UNIVERSITE DE NICE SOPHIA ANTIPOLIS

U.F.R. DE SCIENCES DU LANGAGE

THESE

Sciences du langage

présentée et soutenue publiquement par

Olivier KRAIF

le vendredi 29 juin 2001

Constitution et exploitation de bi-textes pour l'Aide à la traduction

Directeur de thèse : Henri Zinglé

JURY

Pierre Bernhard
Laurence Danlos (rapportrice)
Arnaldo Moroldo
Jean Véronis (rapporteur)
Henri Zinglé

Constitution et exploitation de bi-textes pour l'aide à la traduction

La pierre de Rosette

Basalte noir, Hauteur : 106 cm ; largeur : 76 cm ; épaisseur : 28 cm, découverte en juillet 1799, au Fort Julien près de Rashid « Rosette », port situé à l'est d'Alexandrie, conservée au British Museum, EA 47. Datée de l'an 9 de Ptolémée V-Épiphane : 196 av. J.-C

Remerciements

Je remercie Henri Zinglé pour son ouverture d'esprit et son soutien, qui n'a pas faibli tout au long de ce travail. Merci aux membres du jury pour l'intérêt qu'ils ont manifesté envers ce sujet, situé au croisement de différentes disciplines. Je remercie spécialement Jean Véronis, grâce à qui j'ai pu participer au projet Arcade et bénéficier des corpus adéquats, qui m'ont permis de mener à bien mes expérimentations.

Merci à Hélène, Luc, Kim et Valentina pour leur compétence et leur aide linguistique, en anglais et en italien. Merci à Karine, Mondher, Peggy, Sam et tous les membres du LILLA dont les remarques, les suggestions, les questions, les conseils bibliographiques, et la bonne humeur ont beaucoup apporté à l'ensemble de ce travail.

Un grand merci enfin à Elisabeth, qui m'a accompagné et soutenu pendant toutes ces années de recherche.

Table des matières

Abréviations :

MT :	Mémoire de traduction
STT :	Station de travail pour traducteur
TA :	Traduction automatique
TABE :	Traduction automatique basée sur l'exemple
TABR :	Traduction automatique basée sur les règles
TAL :	Traitement automatique du langage
TAO :	Traduction assistée par ordinateur
UT :	Unité de traduction

NB : les citations dont l'original figure en bas de page sont traduites par nous.

INTRODUCTION

Ningún problema tan consustancial con las letras y con su modesto misterio como el que propone una traducción.

Jorge Luis Borges, Las versiones homéricas, *Discusión*, *1957*

« *Perhaps the single most remarkable observation about machine translation is that it has attracted the attention of a vanishingly small number of researcher with some knowledge of traditional translation*. *And one of the most remarkable facts about translation as a field of inquiry is that it has very rarely been treated as an empirical enterprise*. *As a result*, *the literature on translation theory is replete with simplified versions of linguistic theories about morphology*, *syntax and semantics in the apparent belief that they have something to say about translation*. *But what translators actually do and how they do it remains largely mysterious*.»

Martin Kay, Préface à *Parallel Text Processing*, Jean Véronis, 2000

Introduction

Si l'on considère les rapports de la pratique de la traduction avec le traitement informatique, l'expression *Aide à la traduction* sonne comme un euphémisme. Chronologiquement, on a vu d'abord apparaître le terme de *Traduction automatique*, qui traduisait les expressions anglaises *Machine translation* ou *Mechanical Translation*[1]. Puis est apparue l'expression *Traduction assistée par ordinateur*[2] (TAO), désignant non plus un processus entièrement automatisé, mais le développement d'outils informatiques destinés à aider le traducteur humain. Le terme *Aide à la traduction*, qui se situe dans le prolongement de cette dernière, est d'usage plus récent : il s'adresse toujours à des outils informatisés, bien que la référence à l'automatisation ou à l'informatique n'y soit plus explicite. Pris dans un sens littéral, ce terme pourrait aussi bien s'appliquer à des dictionnaires « papier » ou des collections de fiches terminologiques.

On assiste ainsi à une évolution paradoxale : parallèlement à la progression quasi exponentielle des performances des ordinateurs, et à l'envahissement progressif des technologies de l'information dans tous les secteurs de la vie sociale, il semblerait que les rapports entre traduction et informatique s'orientent vers une diminution graduelle de l'automatisation. Est-ce une illusion d'optique, une simple erreur de perspective ?

Pour bien comprendre l'évolution de la Traduction automatique (TA), il est nécessaire de retracer brièvement l'histoire du domaine (elle-même très brève : une cinquantaine d'années !), qui fut marquée dès le début par des visions prophétiques et … aussi quelques erreurs de perspective.

---

[1] cf. K. Delavenay & E. Delavenay (1960), *Bibliographie de la traduction automatique*, Mouton, Paris. (titre original : *Bibliography of mechanical translation*).

[2] cf. C. Greenfield & D. Serain (1977), *La Traduction assistée par ordinateur : des banques de terminologie aux systèmes interactifs de traduction*, AFTERM, Institut de recherche d'informatique et d'automatique, Paris.

Les premières recherches furent marquées par la publication d'un mémorandum de Warren Weaver (1949), de la Fondation Rockefeller[3], qui proposait de s'inspirer de paradigmes hérités de la théorie de l'information (Claude Shannon, 1949) et des modèles issus de la cryptographie. Henri Zinglé (1993) résume ainsi les questions développées par Weaver :

> « a) l'ambiguïté lexicale peut, selon lui, être résolue en tenant davantage compte du contexte ; b) l'étude des fondements logiques du langage devrait permettre de mieux appréhender le passage d'une langue à l'autre ; c) des résultats significatifs pourraient être atteints avec une approche statistique et en tirant notamment parti des méthodes de décryptage utilisées durant la seconde guerre mondiale. »

Un premier prototype, dédié à la traduction du russe vers l'anglais, est développé sous la direction de L. Dostert, à l'université de Georgetown. Malgré la faible couverture de ce système (250 mots, 6 règles de grammaire, testés sur un corpus de 49 phrases), la démonstration publique, le 7 janvier 1954, aura un grand retentissement.

De 1955 à 1966, de très nombreux projets verront le jour, aux Etats-Unis, au Mexique, au Japon, en Chine, en URSS et dans la plupart des pays d'Europe (Grande-Bretagne, France, RFA, Belgique, Italie, RDA, Pologne, Roumanie, Tchécoslovaquie, Hongrie, Yougoslavie, etc.).

Mais, déjà certaines critiques se font jour quant aux directions empruntées par la plupart de ces groupes de recherche. En 1959, Yehoshua Bar-Hillel, un des pionniers du domaine, met en garde contre les faux espoirs suscités par la TA. Pour lui, les systèmes qui travaillent aux niveaux lexical et morphosyntaxique ne peuvent obtenir des traductions de haute qualité de manière totalement automatisée (en anglais *Fully Automated High Quality Translation*, parfois abrégé par FAHQT). Bar-Hillel donne l'exemple suivant, désormais devenu classique : pour traduire la phrase « *the box was in the pen* », l'ordinateur doit choisir entre deux sens de *pen*, correspondant à « *stylo* » ou à « *parc*, *enclos* ». Or, privé d'informations sur le monde, ou tout simplement incapable de sens commun, la machine ne peut motiver son choix. L'intervention humaine devient indispensable dans une phase de post-édition, ce qui amène à reconsidérer le partage entre travail humain et automatisation :

---

[3] publié en 1955 dans : W. N. Locke & A. D. Booth ed., *Machine Translation of Languages*, MIT Press, Cambridge, MU.

> « Ainsi, à court terme, il apparaît que le seul objectif raisonnable pour la recherche en TA consiste à trouver un équilibre, entre automatisation et post-édition, qui soit commercialement rentable par rapport à la traduction humaine – pour ensuite essayer d'augmenter la compétitivité de cette collaboration en améliorant la programmation, afin de laisser à la machine une part de plus en plus importante dans le processus complet de traduction, vis-à-vis des opérations qu'elle peut effectuer plus efficacement qu'un réviseur humain. »[4] Bar-Hillel (1964 : 172)

Publié en 1966, le rapport ALPAC (pour *Automatic Language Processing Advisory Committee*), rédigé par un comité mandaté par l'Académie nationale des sciences des Etats-Unis conclut que la TA, au vu des résultats décevants d'une décennie de recherches, ne sera pas économiquement rentable à court terme. Le rapport insiste sur les limites des systèmes directs, de langue à langue, inspirés de l'expérience de Georgetown (plus tard appelés « systèmes de première génération »). Il en ressort que le domaine d'application demande à être reconsidéré, et que certaines directions de recherche doivent être privilégiées, notamment par la prise en compte du niveau sémantique et par les approches indirectes, basées sur une représentation intermédiaire. La publication de ce rapport provoquera un véritable séisme dans le monde de la TA : les crédits seront brutalement coupés aux Etats-Unis et en Grande-Bretagne, de nombreux groupes de recherche cesseront leurs activités.

Comme le note Zinglé (1993), Bar-Hillel avait dégagé dès 1951, de manière presque prophétique, les grandes lignes de cette critique : importance du niveau sémantique et des descriptions morphosyntaxiques, faible intérêt des méthodes superficielles (et notamment statistiques), nécessité d'une approche réaliste imposant des restrictions du domaine et du vocabulaire, limitation inhérente à l'automatisation complète, etc. Déjà, les questions posées par Bar-Hillel (1951 : 230, cité et traduit par Isabelle *et al*., 1993a) impliquaient un « glissement » de la TA pure vers l'aide à la traduction (quoique dans ce cas il faudrait interpréter l'expression *aide à la traduction* comme une aide humaine à la traduction automatisée) :

---

[4] "The only reasonable aim, then, for short-range research into MT seems to be that of finding some machine post-editor partnership that would be commercially competitive with existing human translation, and then try to improve the commercial competitiveness of this partnership by improving the programming in order to delegate to the machine more and more operations in the total translation process which it can perform more effectively than the human post-editor."

> «Dans le cas des domaines cibles où la précision absolue est une condition *sine qua non*, il faut renoncer à la TA pure en faveur de la TA mixte, c'est-à-dire un processus traductionnel faisant intervenir l'intelligence humaine. Ce qui soulève la question suivante : Quelles étapes du processus devrait-on confier à un partenaire humain? »

L'écho de cette problématique a traversé toute l'histoire de la TA. Une des grandes figures du domaine, Martin Kay, faisait en 1980 la déclaration suivante :

> « Je veux préconiser une approche selon laquelle on permettrait à la machine de prendre en charge graduellement, presque imperceptiblement, certaines fonctions du processus global de traduction. La machine assumerait d'abord des tâches périphériques. Puis, peu à peu elle s'attaquerait au cœur du processus. La démarche serait toujours empreinte de modestie. Jamais n'essaierait-on de faire plus que ce que l'on sait bien faire. » (cité et traduit par Isabelle *et al*., 1993a)

La modestie est le maître mot de cette attitude. Faut-il en conclure que la TA a fait long feu, victime de l'arrogance illusoire de ses premières ambitions ?

Qu'on nous permette de passer sur des décennies de recherche et développement, et la mise au point des systèmes de première et deuxième génération : si l'on fait un bilan, plus de 30 ans après le rapport ALPAC, force est de constater que la réalité est très nuancée, et que la TA n'a pas dit son dernier mot.

En ce qui concerne la traduction entièrement automatisée, on peut citer le succès du système TAUM-Météo, développé pour les services météorologiques du gouvernement canadien. Ce système traduit des informations météorologiques depuis 1977 de l'anglais vers le français, et depuis 1989 du français vers l'anglais. Comme le décrit Anne Loffler-Laurian (1996 : 22), la réussite de TAUM-Météo est due à la correcte exploitation d'un domaine d'application particulièrement adapté au traitement automatique (et de surcroît particulièrement inadapté à la traduction humaine) :

> « Le type de textes extrêmement particulier auquel le système est destiné possède un vocabulaire limité, ce qui permet un dictionnaire exhaustif des termes à traduire, une syntaxe figée autour d'un petit nombre de structures, ce qui permet une correspondance parfaite de structure à structure, et une organisation discursive bien établie, toujours identique. Les traducteurs humains s'ennuyaient en général très rapidement devant la monotonie du travail et quittaient leur poste en moyenne après six mois. »

Même si l'on considère des systèmes plus généralistes, on constate aujourd'hui que le domaine de la TA se porte plutôt bien. John Hutchins (1996) remarque que les produits destinés aux professionnels comme au grand public connaissent un grand succès, tout

spécialement en Amérique du nord et au Japon. Cela concerne des systèmes implémentés sur des serveurs du type *mainframe* (Systran, Metal, Logos, Fujitsu) mais aussi de nombreux logiciels développés pour des micro-ordinateurs (MicroCAT, Globalink, Korya Eiwa de Catena, AppTek, CITAC, EJ Bilingual, LEX, Neocor, PC-Translator, Systran, etc.), sans compter les services de traduction en ligne (Systran, Logos, Globalink, Fujitsu, NEC, etc.). Parallèlement, on voit se développer un autre marché, très prometteur : la traduction de courrier électronique. CompuServe a rencontré un certain succès avec son service de traduction, brute ou post-éditée.

Sur le créneau de l'Internet, la demande de traduction vise essentiellement à faciliter l'accès à la recherche et à l'assimilation de l'information. Ainsi, il semblerait y avoir un marché important pour la traduction automatique brute, non-révisée, qui concerne par exemple 85 % des services fournis par CompuServe.

Comme le note Hutchins (1996), en Europe, la TA s'est concentrée essentiellement sur le segment professionnel, ce qui représente cependant un marché important :

> « En Europe, les systèmes de TA sont principalement utilisés par de grands services de traduction et par des compagnies multinationales, comme la société de développement SAP qui traduit quelques 8 millions de mots par an à l'aide de Metal et Logos ; Ericsson est un grand utilisateur de Logos pour traduire ses manuels ; et la Commission européenne a connu une rapide augmentation de l'utilisation de Systran, à ce jour 200 000 pages par an (…) »[5]

La pérennité et le succès commercial d'un système comme SYSTRAN, développé à partir de 1969 selon une architecture directe, inspirée en droite ligne par les travaux de Georgetown, montre que les conclusions du rapport ALPAC n'étaient pas sans appel. Certains faux espoirs ayant été abandonnés, la mise en œuvre de traductions médiocres peut néanmoins trouver son intérêt, dans certains contextes. Qu'on examine l'exemple suivant, montrant les versions anglaise et française d'un extrait du corpus JOC (cf. annexe A-I), et la traduction française obtenue sur Internet avec Systran :

[5] " In Europe, MT systems are used mainly by large translation services and by multinational companies, e. g. the software group SAP is translating some 8 millions words a year using Metal and Logos; Ericsson is a large user of Logos for translating its manual; and the European Commission has seen a rapid growth in the uses of Systran, now some 200,000 pages a year (…)"

Version anglaise (corpus JOC)

> *The Commission Translation Service is in the process of equipping itself with modern computer tools providing facilities for word processing, document and terminology retrieval, document transmission and office management in the official languages* (*including Greek*) *on an equal footing so that it is better placed to meet all the demands made upon it.*

Version française (corpus JOC)

> Le Service de traduction de la Commission des Communautés européennes s'efforce en outre de mettre en œuvre des outils bureautiques et informatiques dans le domaine du traitement de texte, de la recherche documentaire et terminologique, de la transmission de documents et de la gestion administrative, au niveau des quels toutes les langues (y compris le grec) sont traitées sur un pied d'égalité, et ceci afin de lui permettre de mieux répondre à l'ensemble de ses obligations.

Traduction automatique de la version anglaise avec Systran (version Internet)

> Le service de traduction de la Commission est en cours de s'équiper des outils modernes d'ordinateur fournissant des équipements pour le traitement de texte, la recherche de document et terminologique, la transmission de document et la gestion de bureau dans les langues officielles (Grec y compris) sur une constante de bas de page égale de sorte qu'il mieux soit placé pour satisfaire toutes les demandes faites sur lui.

Les ruptures syntaxiques, les faux-sens et la bizarrerie des tournures de la version « automatique » peuvent prêter à sourire : mais cela n'enlève rien à l'intérêt d'une telle traduction, qui permet d'offrir l'accès, globalement, au contenu du texte anglais. Même si la reconstruction du sens n'est pas totale, le locuteur français a tôt fait de repérer les anomalies syntaxiques ou sémantiques, et de passer outre, afin de se faire une idée d'ensemble : la mention à la « constante de bas de page » n'induira personne en erreur !

Dès lors, il apparaît que la TA gagne à être située dans le champ élargi de l'aide à la traduction. Elle offre des outils, parmi d'autres. Ce rôle d'assistant, évoqué par Hutchins (1996) dans la citation suivante, n'est qu'une autre forme de la « modestie » prônée par Kay :

> « On a abandonné depuis longtemps l'idée de développer des systèmes complètement automatisés capable d'effectuer, pour des textes tout-venant, un travail proche de celui du traducteur humain. Le but de la Traduction automatique et des activités connexes est de produire, pour des traducteurs

> professionnels et non professionnels, des aides et des outils tirant le meilleur parti de l'ordinateur pour seconder l'intelligence et les compétences humaines. »[6]

Ce changement de perspective, de la TA vers l'aide à la traduction, n'implique donc pas une restriction des applications, mais au contraire un élargissement du champ de vision, englobant tous les apports de l'automatisation permettant de faciliter le processus traductionnel.

Ces apports concernent par exemple la création et la gestion de bases de données terminologiques et de dictionnaires multilingues, comme la base EURODICOTAUM[7] développée par la Communauté européenne : « il s'agit d'un dictionnaire essentiellement terminologique, avec des nomenclatures correspondant aux diverses branches scientifiques et techniques dont la communauté est amenée à traiter : physique nucléaire, nouveaux matériaux, technologies de l'information, biologie, médecine, aéronautique, espace, droit, administration. » (Loffler-Laurian, 1996 : 10)

Plus spécialement dédiées aux traducteurs professionnels, les stations de travail pour les traducteurs (STT) visent à l'intégration d'outils variés destinés à faciliter le traitement des textes, la gestion des traductions déjà faites, l'accès aux informations linguistiques et terminologiques utiles, etc. Il en existe déjà des versions commerciales (Trados's Translation Workbench, IBM's TranslationManager, STAR's Transit, Eurolang's Optimizer) qui d'après Hutchins (1996) « offrent une gamme similaire : traitement de texte multilingue avec fractionnement d'écran, reconnaissance, recherche et gestion de la terminologie, mémoire de traduction (pré-traduction basée sur des textes existants), logiciels d'alignement pour les utilisateurs qui veulent créer leur propre base de données de textes bilingues, conservation du formatage du texte original – et elles supportent un large éventail de langues européennes, à la fois comme langue source ou langue cible. »[8]

---

[6] "The idea of developing fully automatic general-purpose systems capable of near-human translation quality has been long abandoned. The aim of MT research and related activities is to product aids and tools for professional and non-professional translators which exploit the potentials of computers to support human skills and intelligence."

[7] On peut accéder à cette base à l'adresse : http://eurodic.ip.lu/cgi-bin/edicbin/EuroDicWWW.pl

[8] "in facilities and functions, each offer similar ranges : multilingual splitscreen word processing, terminology recognition, retrieval and management, translation memory (pre-translation based on existing texts), alignment software for users to create their own bilingual text databases, retention of original text formatting, and support a very wide range of European languages, both as source and target languages."

Parallèlement à un relatif déclin de la recherche dans le domaine de la TA proprement dite (concernant des textes écrits), de nouvelles voies sont explorées, comme la Traduction automatique fondée sur le dialogue (ou TAFD, Eric Werhli, 1992) ou l'aide à la rédaction (Zinglé, 1996). Comme le remarque Hutchins (1996), ce mouvement se traduit notamment par l'émergence de nouvelles méthodes : « réseaux de neurones, architectures parallèles, et surtout des approches basées sur les corpus : analyse statistique de texte (alignement etc.), génération à partir d'exemples utilisant des modèles statistiques, systèmes hybrides combinant des règles linguistiques traditionnelles et des méthodes probabilistes, etc. »[9]

Une direction semble particulièrement prometteuse : la réutilisation des traductions déjà faites. Pierre Isabelle (1992 : 726), évoquant l'important volume annuel de traductions au Canada, part d'un constat d'évidence :

> « Au Canada seulement, bon an mal an, le volume de traductions atteint au moins un demi-milliard de mots. Il faut se rendre à l'évidence. *La masse des traductions produites chaque année contient plus de solutions à plus de problèmes que tous les outils de référence existants et imaginables* ! Si seulement les services de traduction pouvaient considérer que ces masses de textes qu'ils génèrent constituent une richesse qu'ils pourront réexploiter… »

Il est étonnant que cette idée de « recyclage », pourtant très simple, ne soit apparue que tardivement. D'après Jean Véronis (in Véronis, 2000 §1), citant Alan Melby, « l'idée d'enregistrer des exemples de traduction en vue d'une utilisation ultérieure semble avoir émergé indépendamment, à la fin des années 70, dans différents centres de recherches, parmi lesquels Brigham Young et Xerox PARC. »[10] En 1984, M. Nagao propose une méthode de Traduction automatique basée sur l'exemple (TABE, en anglais *Example-Based Machine Translation*, ou EBMT), consistant à décomposer les phrases en petits segments, chercher dans une base les exemples les plus proches de ses segments, rapatrier les traductions de ces segments, puis recomposer les morceaux traduits en phrases cohérentes. Par la suite, vers la fin des années 80, on voit apparaître les techniques

[9] "neural networks, parallel processing, and particularly in corpus-based approaches: statistical text analysis (alignment, etc.), stastitics-based generation from example texts, hybrid systems combining traditional linguistic rules and statistical methods, and so forth."

d'*alignement* visant à apparier des portions correspondantes, sections, paragraphes ou phrases, d'un texte et de sa traduction. La première méthode d'alignement automatique, mise au point par Martin Kay & Martin Röscheisen (1988), ouvrira la voie à de nombreux développements.

Les textes sources et cibles (on dit aussi : textes *parallèles*), à l'issue des opérations de segmentation et d'appariement, deviennent des bases de données informatisées, offrant une puissance d'indexation qui facilite l'accès aux exemples de traduction. Pour ce type de structure, B. Harris (1988) propose le terme de *bi-texte*, qu'on généralise à *multi-texte* lorsque les textes traduits impliquent plus de deux langues. L'application des techniques d'alignement automatique permettant la constitution massive de corpus bi-textuels, il devient désormais possible de rassembler des *Mémoires de traduction*, bases de données assez vastes pour receler ces « solutions déjà trouvées » dont parle P. Isabelle (1992) : « Le bi-texte constitue de ce fait l'amorce d'une approche à base mémorielle : au lieu de recréer à chaque fois une solution à un problème de traduction particulier, on se donne la possibilité de rappeler les solutions déjà trouvées. »

Des projets de *Traduction automatique basée sur l'exemple* à la constitution de *Mémoires de traduction* (MT), les corpus bi-textuels s'intègrent donc tout naturellement dans le paysage de l'aide à la traduction : comme l'ont montré de nombreuses recherches menées au RALI de Montréal, ils en constituent même un des pivots essentiels, qu'il s'agisse de la compilation de bases de données traductionnelles, de l'extraction de glossaires multilingues, de la vérification automatique de traduction ou de l'automatisation du processus traductionnel.

Fait surprenant : après le détour de décennies de recherche en TA, qui ont montré à quel point il était important d'intégrer une description fine des phénomènes linguistiques mis en jeu par la traduction ; après avoir dégagé l'intérêt de recourir à des représentations intermédiaires abstraites indépendantes des langues, permettant ainsi une plus grande modularité ; après avoir insisté sur la prééminence des niveaux sémantiques et conceptuels dans le processus de traduction… on assiste, avec les méthodes d'extraction de corpus bi-

[10] "According to Alan Melby, the idea of storing sample translations in view of later reuse seems to have sprung up independently at various research centres in the late seventies, including Brigham Young and Xerox Park".

textuel, au développement poussé de techniques totalement alinguistiques, qui ne s'intéressent qu'aux phénomènes les plus superficiels concernant le passage d'une langue à une autre. Il y a comme un retour (ou une régression ?) aux idées fondatrices de Weaver, qui comparait la traduction à une opération de décryptage. Comment ne pas s'étonner, 50 ans plus tard, du développement fourmillant des indices statistiques et des modèles mathématiques issus de la théorie de l'information, qui constituent la plus grande part des travaux dans le domaine de la mise en œuvre des corpus bi-textuels ?

C'est avec une certaine curiosité pour le succès inattendu des modèles statistiques (qui ont souvent eu mauvaise presse auprès des linguistes), que nous proposons, dans la présente étude, de faire le point sur les méthodes « bi-textuelles ». Jusqu'à présent, les résultats ont été encourageants, et l'avenir semble prometteur, car les applications de l'alignement sont nombreuses.

On pourrait estimer que l'alignement constitue une tâche bien modeste. Partir de textes déjà traduits, les couper en morceaux, apparier les paragraphes, phrases, syntagmes ou mots qui se correspondent : tout cela semble bien trivial !

Et pourtant… il suffit de considérer l'exemple précédent tiré du corpus JOC, pour prendre conscience de certaines difficultés. Tâchons de segmenter le plus finement possible la version anglaise, et essayons de déterminer les segments correspondants dans la version française (résultat d'une traduction humaine). On pourrait proposer l'« alignement » suivant :

> (*The* ; Le), (*Commission* ; de la Commission des Communautés européennes), (*Translation* ; de traduction), (*Service* ; Service), (*is in the process of equipping itself with* ; s'efforce en outre de mettre en œuvre), (*modern computer* ; bureautiques et informatiques), (*tools* ; des outils), (*providing facilities for* ; dans le domaine), (*word processing* ; du traitement de texte), (*document* ; documentaire), (*and* ; et), (*terminology* ; terminologique), (*retrieval* ; de la recherche), (*document* ; de documents), (*transmission* ; de la transmission), (*and* ; et), (*office management* ; de la gestion administrative), (Ø ; au niveau desquels), (*in* ; Ø), (*the official languages* ; toutes les langues), (*including* ; y compris), (*Greek* ; le grec), (Ø ; sont traitées), (*on an equal footing* ; sur un pied d'égalité), (Ø ; et ceci), (*so that* ; afin de), (*it is better placed to* ; lui permettre de mieux), (*meet* ; répondre), (*all* ; à l'ensemble), (*the demands made upon it* ; de ses obligations)

D'emblée un certain nombre de problèmes surgissent :

– certaines unités n'ont pas de correspondant évident : comme (angl.) *in* ou (fr.) *sont traités* ;

– dès lors, faut-il apparier (angl.) *Commission* avec (fr.) *de la Commission des Communautés européennes*, ou seulement (angl.) *Commission* avec (fr.) *Commission* ?

– que signifie la correspondance entre (angl.) *modern computer* et (fr.) *bureautiques et informatiques* ? est-elle réutilisable dans un autre contexte ?

– de même, faut-il apparier (angl.) *providing facilities for* avec (fr.) *dans le domaine de*, ou ne pas les apparier du tout ?

– est-il légitime de considérer (angl.) *terminology* et *retrieval* comme deux unités traduites indépendamment, en les appariant respectivement avec (fr.) *terminologique* et *recherche* ?

A bien examiner les appariements donnés précédemment il est difficile de se départir d'un sentiment d'arbitraire : on aurait pu aligner de bien d'autres manières. Tout semble reposer sur des intuitions plutôt vagues. Les problèmes qui apparaissent sont nombreux, et de nature différente :

– problème de la *granularité* de l'alignement : jusqu'où segmenter ? quelle est la consistance des unités segmentées ?

– problème de l'équivalence : que signifie l'appariement ? à partir de quel degré de « différence » doit-on considérer qu'une unité n'a pas de correspondance ?

– problème de la biunivocité : que signifient les unités sans correspondances ? faut-il les considérer comme des exceptions, la règle étant que les unités ont en principe une contrepartie dans le texte traduit ?

– à quoi peut servir ce type d'alignement ?

– l'alignement est-il une structure objective sous-jacente à toute traduction ?

Pour les besoins de notre exemple, nous avons recouru à un alignement quasi lexical, à l'intérieur d'un couple de phrases. Mais toutes ces questions peuvent se poser *de droit* pour n'importe quel type d'alignement, à quelque niveau que ce soit : elles relèvent toutes d'une problématique très générale. Avant même qu'il soit question d'automatisation, elles en remettent en cause le principe : car comment automatiser, avec rigueur et cohérence, une tâche à ce point floue et mal définie ?

Pour aborder ces problèmes, il ne faut pas perdre de vue l'origine du matériel bi-textuel : c'est bien de traduction *humaine* qu'il s'agit. On ne peut donc se contenter d'une approche normative ou idéale de la traduction, comme on l'a souvent fait, à juste titre, en TA. L'aliment d'une mémoire de traduction est un matériau concret issu d'une pratique sociale, et ce n'est qu'à partir d'une approche résolument empirique qu'on pourra espérer en étudier la richesse, la composition et le mode de digestion. Dans la préface à *Parallel Text Processing* (Véronis, 2000), ouvrage de synthèse sur l'exploitation des textes parallèles, Martin Kay fait le constat suivant :

> « Un des faits les plus remarquables à propos de la traduction en tant que champ d'investigation, est qu'elle a rarement été traitée en tant qu'entreprise empirique. C'est pourquoi la littérature sur la théorie de la traduction regorge de versions simplifiées de théories linguistiques traitant de morphologie, de syntaxe et de sémantique, suivant l'idée qu'elles ont quelque chose à dire à propos de la traduction. Mais ce que les traducteurs font vraiment, et comment ils le font, reste largement mystérieux. »

Avant de commencer à répondre aux questions soulevées par la mise en œuvre de l'alignement, il nous faudra essayer de soulever, au moins en partie, un coin de ce mystère.

Le mot *traduction* a un double sens, puisqu'il signifie à la fois un procès et son résultat. Nous nous limiterons au point de vue du résultat, car ce qui nous importe, ce n'est pas tant le « comment », les méthodes de travail employées par les traducteurs, mais plutôt le produit et la fonction de ce produit. Pour notre propos, la question centrale, est : quelles sont donc les relations sous-jacentes à un texte et sa traduction ?

Dans la première partie de ce travail, nous chercherons à décrire ces relations, et nous constaterons que tout se résume à un problème d'équivalence. Comme l'avait si bien vu Roman Jakobson (1963 : 80), « l'équivalence dans la différence est le problème cardinal du langage et le principal objet de la linguistique ». C'est aussi, à un degré peut être plus fort, le problème cardinal de la traduction. Cette photographie rapide de la pratique traductionnelle nous permettra ensuite d'étudier de façon réaliste les articulations possibles entre la traduction humaine et l'automatisation. Nous en dégagerons la place des outils bi-textuels dans le champ de l'aide à la traduction.

Dans les parties II et III, nous étudierons le cœur des techniques dites d'alignement : nous tenterons d'en formaliser et d'en conceptualiser les objets. A la lumière des considérations sur la nature de la traduction humaine, nous nous efforcerons de donner une définition rigoureuse et explicite de l'alignement.

La deuxième partie sera consacrée à un type d'alignement que la plupart des chercheurs s'accordent à considérer comme le moins problématique : l'alignement au niveau des phrases. Nous verrons qu'un grand nombre de techniques ont été développées, produisant de bons résultats. Dans la troisième partie, nous examinerons les méthodes destinées à l'appariement des unités lexicales. Les problèmes spécifiques à ce niveau nous amènerons à définir un autre concept d'alignement : la notion de *correspondance lexicale*.

Afin de mieux délimiter les enjeux et les problèmes posés par la constitution et l'exploitation de bi-textes, une partie importante de notre travail s'articulera autour de l'expérimentation des nombreuses techniques étudiées à ce jour. Nous chercherons, à partir d'une étude empirique détaillée, d'évaluer avec précision les méthodes les plus représentatives – et nous tâcherons, le cas échéant, de proposer des améliorations et des mises au point dans la mise en œuvre algorithmique.

Cette expérimentation s'organisera autour d'un parti pris méthodologique : nous nous limiterons, autant que possible, aux méthodes formelles ne requérant pas de connaissances de nature linguistique. Il ne s'agit pas de conclure à l'inutilité de ces informations : bien au contraire, nous pensons que le traitement des bi-textes ne peut que bénéficier d'un couplage entre les descriptions linguistiques *ad hoc* et les méthodes

quantitatives brutes. Cette restriction délibérée du champ d'investigation est motivée par deux raisons :

- d'une part, nous pensons que les régularités statistiques peuvent révéler de nombreux phénomènes, à leur niveau, sur le rapport qui lie un texte et sa traduction ;

- d'autre part, pour articuler les techniques formelles avec les traitements basés sur des connaissances linguistiques, il est intéressant de dégager les possibilités et les limites des techniques formelles employées seules. C'est tout le but de l'évaluation que nous proposons ici.

Cette étude poursuit donc des objectifs à la fois pratiques et théoriques : d'abord, faire un état de l'art des techniques d'alignement, nombreuses, variées et souvent sophistiquées ; mais surtout donner une vision critique de ces techniques par la mise en lumière des problèmes fondamentaux posés par le bi-texte, cet objet linguistique d'un genre « nouveau », si mal connu, et pourtant vieux comme la pierre de Rosette (Véronis in Véronis, 2000 §1).

Rien ne sert en effet de se concentrer sur tel aspect algorithmique ou mathématique, si l'on ne sait ce que signifient les objets qu'on manipule, d'un point de vue linguistique et communicationnel.

Au terme de cette étude nous espérons voir se dessiner une problématique nouvelle. Les bi-textes manifestent-ils des *structures* spécifiques, ayant une pertinence sur le plan linguistique ? Peut-être verrons-nous surgir un niveau particulier de phénomènes, qui n'apparaissait pas au seul plan monolingue. Car comme le note Guy Bourquin (1991 : 115-116) :

> « On voit que le parti d'aborder des langues par le biais conjugué de la traduction et de l'automatisation (ou plutôt de la préoccupation automatisante) conduit à faire surgir inopinément des convergences dont certaines apparaissent comme la trace de phénomènes universaux encore mal repérés »

Partie I

Problèmes de traduction

« *Non verbum de verbo reddere sed sensum* »

Saint Jérôme

« *Ce que la communication verbale ne dit pas*, *le message le révèle* »

Danica Seleskovitch, *Langue*, *Langage et mémoire*, 1975, p. 178

I Traduction humaine et traduction assistée par ordinateurs : problèmes, enjeux et limites

On ne peut aborder les techniques d'alignement, et leurs applications en traduction assistée par ordinateur, sans d'abord mettre en lumière les problèmes posés par la traduction en général, considérée en tant que processus, acte de langage et démarche communicative.

Les errances et les illusions des débuts de la TA doivent beaucoup à la naïveté des premières tentatives, qui s'inspirèrent d'une vision simplifiée de la traduction, plus souvent guidée par les paradigmes de la cryptographie et de la théorie de l'information, que par une connaissance approfondie de la pratique de la traduction en tant que telle.

Avec les systèmes de première génération on a cru pouvoir traduire par simple substitution et remise en ordre des mots, moyennant des règles syntaxiques simples liées au contexte immédiat des unités lexicales. Cette approche directe, *ad hoc*, dépendant du couple de langues, nécessitait le développement de règles complexes sans généralité, puisque tributaires des langues mises en jeu et du sens de traduction. En outre, la prise en compte de divergences structurelles profondes entre systèmes linguistiques différents, devenait d'autant plus délicate que ces systèmes reposaient sur une faible assise linguistique (Zinglé, 1993).

Les systèmes de seconde génération, basés sur le transfert ou les représentations pivot, ont apporté une amélioration décisive au niveau de l'économie des traitements lexico-syntaxiques, en généralisant le principe de modularité : « les processus d'analyse et de génération sont séparés, ce qui permet de réutiliser les modules d'analyse et de génération d'une langue donnée pour tous les sous-systèmes de traduction utilisant cette langue soit en langue source, soit en langue cible. » (Zinglé, 1993). Mais, dans une démarche réductionniste, on a délibérément mis de côté les aspects cognitifs et pragmatiques liés à l'acte de traduire. On s'est axé sur les aspects technologiques et linguistiques, au détriment d'une réflexion élargie sur la traduction comme acte de communication remplissant de multiples fonctions.

Dans les recherches en TA, toutes les approches s'appuyaient sur un parti pris méthodologique positif, isolant provisoirement les textes de tout contexte extralinguistique, comme le remarque Harry Somers (1993 : 234) : «(...) jusqu'à présent, tous les systèmes de Traduction automatique ont été conçus en partant de l'hypothèse que le texte source contient assez d'information pour entreprendre la traduction[11] ».

Plus exactement, on supposait que l'information contenue dans le texte original, assortie de données linguistiques concernant les deux langues (grammaires et dictionnaires), suffisait à obtenir une traduction du texte source. Hypothèse opératoire certes utile et stratégique, dans la mesure où il est naturel de s'atteler à des problèmes « simples » avant de chercher à résoudre des problèmes compliqués. Mais hypothèse intenable quand il s'agit de dessiner un état des lieux des difficultés, des possibilités et des limites inhérentes aux systèmes, ainsi que des directions les plus prometteuses à explorer.

Or, il importe, dès le départ d'une recherche, de prendre la mesure des problèmes soulevés. Et comme le note M. Kay (in Véronis, 2000, préface), traduire automatiquement implique de capter, même superficiellement, une telle somme de connaissances sur le monde, qu'on y perd vite ses illusions :

> « Traduire, c'est réexprimer le sens, et le sens n'est pas une propriété émergente des textes, qu'il s'agisse de textes monolingues ou de textes multilingues mis les uns à côté des autres. Quant à savoir si des corpus bilingues assez grands, des ordinateurs assez rapides et des statistiques assez sophistiquées peuvent saisir une image du monde assez précise pour une traduction automatique de qualité, la question reste ouverte, mais il y a peu de raisons d'être optimiste.[12] »

Et pourtant, le champ à explorer reste immense, tant sur le plan théorique qu'au niveau des applications. Pour s'en convaincre, il faut d'abord se donner une vision claire du travail du traducteur humain : qu'il s'agisse de traduction humaine ou de TA, « les objectifs des deux opérations – faciliter la communication entre langues étrangères – sont, en principe, les mêmes[13] » (Juan C. Sager, 1994 : 20). Une fois éclaircie la nature de cette

---

[11] "(…) all MT systems so far have been designed with the assumption that the source text contains enough information to permit translation".

[12] "Translation is the reexpression of meaning, and meaning is not an emergent property of texts in a single language nor in several laid side by side. The question of just how large bilingual corpora, fast computers and sophisticated statistics can focus the picture of the world needed for high quality translation remains open, but there is little to support great optimism."

[13] "because the objectives of both operations - the facilitation of translingual / interlingual communication- are, in principle, the same."

activité de communication très spéciale qui implique une démultiplication des codes linguistiques, des participants et des situations, nous pourrons définir un peu mieux les besoins auxquels l'automatisation peut répondre et les problèmes auxquels on peut raisonnablement s'atteler.

Ce retour aux sources de la pratique traductionnelle est ici d'autant plus indispensable que la famille de méthodes que nous allons étudier se fonde sur des textes traduits par l'homme. Dans ce dialogue entre l'homme et la machine, nous commencerons donc par donner la parole à l'humain.

I.1 L'acte de traduire

I.1.1 Définition et problématique

Le mot *traduire*, dérivant du latin *traducere*, signifiant « faire passer d'un point à un autre, conduire à travers », a été introduit avec son emploi moderne par Leonardo Bruni au XVe siècle sous sa forme italienne, et repris en français par Robert Estienne en 1539. Cette construction suggère d'emblée une idée de transformation sous un aspect dynamique. Dans leur dictionnaire de sémiotique, A. J. Greimas & J. Courtès (1993 : 397-398) proposent la définition suivante :

> « 1- On entend par traduction l'activité cognitive qui opère le passage d'un énoncé donné en un autre énoncé considéré comme équivalent.
> 2- La traductibilité apparaît comme une des propriétés fondamentales des systèmes sémiotiques et comme le fondement même de la démarche sémantique : entre le jugement existentiel " il y a du sens " et la possibilité d'en dire quelque chose, s'intercale en effet la traduction; " parler du sens " c'est à la fois traduire et produire de la signification. »

Avec la notion de « passage », cette définition apporte un certain nombre d'éléments significatifs :

- On traduit des *énoncés*, à entendre comme partie ou totalité d'un message, au sein d'une situation d'énonciation. Nous expliciterons donc en quoi la traduction se définit avant tout comme un acte de communication, et quelles en sont les composantes communicatives.

- Ces énoncés sont en relation d'*équivalence*. Nous verrons que la notion d'équivalence comporte plusieurs dimensions, en rapport avec les enjeux de la communication.

- Il est question de « traductibilité », i.e. des limites inhérentes à l'activité traduisante.

- La traduction est une amorce d'explicitation du sens. Nous examinerons en quoi la traduction présuppose toujours une *interprétation* du sens global du message.

Notons que dans la précédente définition il n'est pas précisé que le passage est effectué entre deux langues. Jakobson (1963 : 79), de la même manière, conçoit la traduction dans un sens élargi, incluant la notion de paraphrase. Il distingue trois types de traduction : 1) la traduction « intralinguale » ou reformulation, 2) la traduction « interlinguale » 3) la « transmutation », ou transformation dans un système de signes non linguistique.

Nous nous limiterons quant à nous au seul cas de traduction interlinguale, tel qu'on l'entend communément, i.e. d'un système linguistique vers un autre.

Sur le plan de la pratique traductionnelle, Eugene Nida (1969 : 12) définit le type d'équivalence qui doit être visé :

> « Traduire consiste à produire dans la langue d'arrivée le plus proche équivalent naturel du message de la langue de départ, en premier lieu sur le plan du sens et en second lieu sur le plan du style. » [14]

Par « naturel », Nida indique que la traduction doit respecter, outre la correction grammaticale, les normes et les usages de la langue d'arrivée. Elle doit en somme être fidèle à ce qu'on nomme le *génie* propre de celle-ci. Mais il convient de relativiser ce concept de « naturel », qui, souligne Georges Mounin (1963 : 278), ne doit pas être pris comme un critère stable dans le temps et dans l'espace, facilement objectivable. Ce qui est « naturel » pour certains locuteurs peut sembler étrange à leurs proches voisins.

Cette condition première étant réalisée, la recherche d'équivalence doit privilégier, toujours selon Nida, le fond sur la forme, le contenu (« sens ») par rapport à la forme de l'expression (« style »). Cette hiérarchisation rappelle la maxime de saint Jérôme : « *non verbum de verbo reddere sed sensum* » (ne pas rendre mot à mot, mais respecter le sens), qui introduisait ainsi l'opposition : *ad verbum* / *ad sensum*. Mais cette définition de la notion d'équivalence reste encore largement indéterminée : que faut-il entendre par *sens* ? Nida (1969 : 14) précise que sa définition requiert l'évaluation d'un certain nombre de contraintes antagonistes, et l'établissement d'un « système de priorité » :

> « Comme il ressort clairement de la discussion sur la définition de la traduction, le traducteur est constamment confronté à des séries de distinctions qui l'obligent à privilégier le contenu par rapport à la forme, le sens par rapport au style,

[14] "Translating consists in reproducing in the receptor language the closest natural equivalent of the source-language message, first in terms of meaning and secondly in terms of style. "

> l'équivalence par rapport à l'identité, l'équivalence la plus proche par rapport aux autres types d'équivalence, et la tournure naturelle par rapport à la forme littérale. »[15]

Ces oppositions soulèvent de nombreuses questions, et demandent une élucidation précise de chaque terme : contenu, forme, sens, style, équivalence, naturel, littéralité, etc. En outre, on peut se demander si l'équilibre entre ces couples antagonistes ne dépend pas étroitement des types de texte : en poésie, la forme n'est-elle pas parfois plus importante que le fond ?

Si l'on reste en surface, de nombreuses contradictions émergent dès qu'on s'interroge sur ce qui constitue le principe d'une bonne traduction. Avec un certain humour, Savory, cité par Sager (1994 : 126), confronte des partis pris apparemment irréconciliables, couramment formulés par les traducteurs professionnels :

> « Une traduction devrait….
> - restituer les termes de l'original ; - restituer l'idée de l'original ;
> - se lire comme un texte original ; - se lire comme une traduction ;
> - refléter le style du traducteur ; - être écrit dans le style de l'original ;
> - se lire comme un texte contemporain à la traduction ; - se lire comme un texte contemporain à l'original ;
> - ne jamais comporter de suppression ni d'ajouts ; - pouvoir comporter des suppressions ou des ajouts » [16]

Voici une belle liste d'apories. Et pourtant, aucune de ces assertions n'est fausse : les contradictions qui se font jour nous révèlent que la traduction est un lieu de paradoxe. Le théoricien ne doit cependant pas si tôt baisser la garde : nous verrons que dès l'instant où l'activité traduisante est considérée globalement, réinsérée au sein d'une situation de communication particulière où un grand nombre de paramètres entrent en jeu, il devient

---

[15] "As it may be clearly noted from the discussion of the definition of translation, one is constantly faced by a series of polar distinctions which force him to choose content as opposed to form, meaning as opposed to style, equivalence as opposed to identity, the closest equivalence as opposed to any equivalence, and naturalness as opposed to formal correspondence."

[16] "A translation should.....- render the words of the original ; render the ideas of the original ; read like an original piece of text ; read like a translation ; reflect the style of the translator ; be written in the style of the original ; read like contemporary work of the translation ; read like a contemporary work to the original ; must never have deletions or additions ; may have additions and deletions."

possible de répondre, par delà les apparentes contradictions, à la question posée par Christian Boitet (1993 : 110, nous soulignons):

> « Très souvent, on entend des arguments du type : un traducteur ne peut bien traduire que ce qu'il comprend, ou, les interprètes (simultanés) traduisent bien sans avoir le temps de comprendre. *Qu'entend-on donc par bien traduire* ? ».

I.1.2 D'un message à un autre: l'exégèse traductionnelle

La traduction est certes une activité linguistique, et même doublement linguistique puisqu'elle implique en principe deux langues distinctes. Mais (comme toute activité langagière), il est impossible de la réduire au seul plan linguistique. Ce que Catherine Fuchs (1983 : 176) remarque à propos de la paraphrase peut s'appliquer à la traduction en général :

> « La paraphrase est un phénomène langagier (c'est-à-dire une activité de langage menée par des sujets dans des situations de discours données), qui n'est que partiellement linguistique (c'est-à-dire s'appuyant sur des relations complexes en langue, qui contribuent à l'établissement d'un jugement de paraphrase, sans pour autant le déterminer absolument). »

L'activité traduisante, étant indissociable d'un certain contexte situationnel, n'est pas conditionnée par les seules déterminations du (ou plutôt des) code(s). Pour les mêmes raisons que la paraphrase, le phénomène de la traduction, afin d'être compris dans sa complexité, doit être considéré avant tout comme activité de communication.

I.1.2.1 Les coordonnées pragmatiques du message

Maurice Pergnier (1993 : 212) établit une distinction entre deux types de processus intervenant dans la réception d'un énoncé :

> « La saisie d'un énoncé par un récepteur suppose [...] deux types de processus analytiques : l'un portant sur le code (idiome), l'autre portant sur le message. Afin d'éviter la confusion entre ces deux plans, nous réserverons le nom d'analyse au premier, et appellerons exégèse celui qui mène à la compréhension du message. »

On peut généraliser ces deux types d'analyse à l'étude de la traduction : nous nommerons *exégèse* l'étude des conditions pragmatiques relatives à l'émission et à la

réception des messages source et cible ; nous nommerons *analyse* l'étude des conditions linguistiques déterminant la transformation d'un énoncé dans le code source à un énoncé dans le code cible. Par exemple, l'analyse traductionnelle se penchera sur les transformations structurales mises en jeu dans la traduction d'une *phrase*, la *phrase* étant considérée comme « objet théorique abstrait », selon l'expression de François Rastier (1989 : 37) découlant du code linguistique[17]. Vis-à-vis de l'exégèse, cette même phrase sera considérée en tant qu'*énoncé*, manifestation linguistique d'un message (ou d'une partie d'un message), objet transcendant les seules déterminations linguistiques, et dont les transformations dépendent de conditions pragmatiques particulières.

L'exégèse met de côté tout « ostracisme » linguistique : elle reconnaît que la langue intervient comme un code, mais non pas un code isolé, un code parmi un grand nombre de codifications sociales susceptibles de laisser une empreinte dans le passage à la traduction : canons littéraires, codes de politesse, genre textuel, symboles religieux, systèmes symboliques scientifiques, terminologies, etc. :

> « Il n'existe pas de texte (ni même d'énoncé) qui puisse être produit par le seul système fonctionnel de la langue (au sens restreint de mise en linguistique). En d'autres termes, la langue n'est jamais le seul système sémiotique à l'œuvre dans une suite linguistique, car d'autres codifications sociales, le genre notamment, sont à l'œuvre dans toute communication verbale. » (Rastier, 1989 : 37)

Au-delà de ces déterminations multiples, dépendantes de codes linguistiques et extralinguistiques, une caractéristique importante du message est sa *singularité*. Etant étroitement lié à une situation donnée, ponctuelle dans le temps, il ne prend son sens original que dans cette situation. Il est une sorte « d'hapax linguistique, puisqu'il n'est jamais totalement reproductible et que sa composante implicite n'est pas totalement identifiable. » (Bernard Pottier, 1992 : 15). Le message est en quelque sorte le résultat instantané de la *précipitation* de tous les facteurs pragmatiques qui entrent en jeu dans la communication : codes, supports, participants, présupposés, intentions, etc.

[17] pour autant que cet objet ait été clairement défini sur le plan linguistique, ce qui n'est rien moins qu'évident comme nous le verrons par la suite.

A peine réalisée, cette réunion ponctuelle et accidentelle (au sens philosophique du terme) des éléments constitutifs du message commence déjà à rayonner, à se disperser dans le temps et dans l'espace. Le message explose, irradie dans de nombreuses directions. Certaines de ses manifestations sont labiles et insaisissables : le message produit des *effets* sur les participants ; ceux-ci en gardent une trace, une *mémoire* qui subira des modifications dans le temps. D'autres manifestations du message prennent une forme figée, du moins en apparence : enregistré, le message se transforme en *document*. Sous une forme scripturale abstraite, nous parlerons de *texte*. Inévitablement, cette transcription stabilisée gagne en autonomie[18], et entame un parcours *sui generis* : elle peut rencontrer de nouveaux récepteurs éloignés de la situation initiale, dans l'espace et dans le temps. Imperceptiblement, les codes linguistiques et sociaux changent, mutent, se déplacent : des interprétations nouvelles éclosent, le texte devient source d'innombrables échos, de résonances nouvelles, quand même les récepteurs effectueraient chaque fois un effort d'exégèse en cherchant à restituer le texte par rapport à ses coordonnées de naissance.

Ainsi, tout message écrit est paradoxal, car il comporte deux faces antagonistes : la première, volatile et éphémère, mais dont l'interprétation est donnée une fois pour toute, constitue le *message original* s'inscrivant dans le temps de l'énonciation ; la deuxième, figée dans sa forme, constitue le *texte*, qui sera livré à une série indéfinie d'interprétations successives, mouvantes même dans leur effort de restituer l'interprétation originale[19]. Danica Seleskovitch (citée par Colette Laplace, 1994 : 225) remarque avec justesse que « Le texte n'est statique que sur les rayons d'une bibliothèque ; au moment de sa lecture, il retrouve le dynamisme qui a présidé à sa naissance par l'écriture. Le texte a un déroulement semblable à celui de la parole et sa traduction se situe dans le dynamisme de son appréhension. »

[18] Comme le note Sager (1994 : 57) : "In all written or recorded communications, the speech act results in a 'document' which is the physical manifestation of the intended message of the writer; as soon as it exists it gains a relative independence."

Il apparaît que l'*exégèse*, ainsi entendue, précède et dépasse l'*analyse traductionnelle*, dans la mesure où elle s'intéresse à la situation non médiatisée, ponctuelle, en tant que lieu de production et d'interprétation du *sens*. En effet, c'est dès ce niveau (englobant par rapport au linguistique), que nous situons le *sens*, à l'instar de Rastier (1989 : 16) :

> « En somme, le sens n'est pas immanent au texte comme message, mais à une *situation de communication* comprenant en outre un émetteur et un récepteur, comme aussi un ensemble de conditions (des normes, dont le genre textuel, et une pratique sociale déterminée). Ces conditions peuvent être dites pragmatiques, mais au sens d'une pragmatique *englobante*. »

Nous nous intéresserons donc d'abord au schéma traductionnel correspondant à l'exégèse, impliquant des messages avec leur système de *coordonnées* pragmatiques (Pottier emploie le terme de « cadrage », 1992 : 16) :

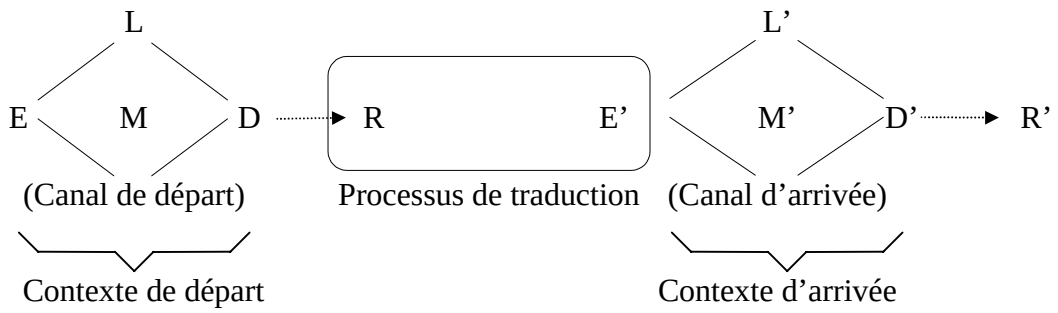
Traduction : Message Source → Message Cible

Comme nous le verrons, la traduction soulève des problèmes particuliers sur les deux plans impliqués par les coordonnées pragmatiques : la différence des codes extralinguistiques, et la singularité des messages.

Si l'on considère le « circuit » de la traduction, on assiste à une situation de communication dédoublée impliquant deux messages, le message source et le message cible. Le traducteur y occupe une position très particulière de médiateur, alternativement récepteur et émetteur. Nous empruntons à Pergnier (1993 : 48) sa représentation du circuit communicationnel, schématisé figure 1 (elle-même inspirée de la représentation classique de Jakobson) :

---

[19] Borges, dans *Fictions*, s'est joué de ce paradoxe. Il imagine un mystérieux personnage, Pierre Ménard, réécrivant le *Don Quichotte*, mot pour mot, au XXe siècle. Même si l'œuvre est matériellement identique, virgule pour virgule, Borges montre que c'est une *autre* œuvre qui est écrite. Car la lecture, et le sens, de l'œuvre de Cervantes, pour un moderne, ne sera jamais le même que pour le contemporain de Cervantes. Quatre siècles de littérature nous en séparent. Entre temps, ce n'est pas seulement la langue qui a changé : c'est le sens de l'acte littéraire, et le sens du monde dans son ensemble. Si le *Don Quichotte* est éternel, il ne cesse d'évoluer, insensiblement, en donnant lieu à des lectures nouvelles.



E : Emetteur, D: Destinataire, R: Récepteur, M : Message, L/L' : Langue de départ/d'Arrivée

figure 1 : *Schéma communicationnel de la traduction*

Ce dédoublement implique une complexité accrue dans la relation émetteur source – récepteur cible, par le jeu d'un certain nombre de déplacements problématiques :

- Le traducteur occupe la double fonction de récepteur et d'émetteur, alors qu'il n'est ni l'émetteur source ni le destinataire du message source. Il occupe une position d'intermédiaire « angélique » au cours de laquelle il doit parfois rester invisible, et parfois se montrer. La situation est en fait plus complexe si l'on considère que le traducteur est rarement un (ré-)émetteur spontané, et qu'il agit la plupart du temps sous l'impulsion d'un commanditaire dont les objectifs et les intentions vont entrer en ligne de compte.

- Le destinataire du message traduit peut être différent du destinataire du message original.

- La situation d'arrivée est la plupart du temps éloignée, dans le temps et / ou dans l'espace, de la situation de départ. Il n'y a que dans l'interprétation (au sens de traduction simultanée) que les situations coïncident.

Ces déplacements entraînent la transformation de nombreux facteurs déterminants dans l'acte de communication. Seleskovitch (citée par Laplace, 1994 : 213) évoque ces facteurs en terme de « compléments cognitifs » :

> « Nos études de la traduction nous ont permis de déceler la provenance et le jeu de ces compléments cognitifs, et d'en relever huit types différents qui interviennent dans la constitution du sens : (1) l'auteur […] (2) le contexte verbal […] (3) le contexte cognitif (4) le bagage cognitif […] (5), (6), (7) et (8) le lieu et le moment où se situe la réalité évoquée par le texte, le destinataire auquel s'adresse un auteur et la situation dans laquelle se déroule la communication jouent également un rôle important dans la constitution du sens […] qui est le point de départ de la compréhension des textes. »

Ces facteurs sont nombreux et hétérogènes, et nous préférons distinguer deux types d'espace pragmatique, la *situation* et le *contexte*. Nous nommerons *situation* d'énonciation le contexte spatio-temporel immédiat lié à l'émission et à la réception d'un message. Nous utiliserons alors le terme de *contexte* dans un sens plus abstrait et plus large, incluant toutes les données sociales et culturelles englobant la situation. L'opposition situation / contexte est ainsi établie le long d'un axe qui va de l'individuel au social, de l'occurrence particulière à la codification conventionnelle[20].

La *situation* d'énonciation se caractérise par un certain nombre de paramètres, que nous nommerons *coordonnées* situationnelles, permettant de repérer le message au sein de son espace énonciatif. Sager (1954 : 57) énumère les dimensions relatives à ces coordonnées : « La situation d'un acte de langage est amorcée par un stimulus et est constituée d'un sujet, d'un temps, d'un lieu et de participants » [21], le stimulus étant ce qui occasionne la production du message. Plutôt que le *sujet* du message, nous préférons considérer à ce niveau l'*intention* de l'émetteur, dans la mesure où les coordonnées désignent les dimensions extralinguistiques déterminant l'interprétation (alors que le *sujet* peut être considéré comme une composante immanente à l'énoncé).

Bien entendu, ces dimensions pragmatiques sont étroitement imbriquées dans toute communication : le stimulus est inséparable des visées intentionnelles des participants, qui sont eux-mêmes plongés dans un certain lieu et un certain temps.

---

[20] Nous ne nous situons pas dans les catégorisations pragmatiques classiques permettant d'identifier différents types de contexte : contexte référentiel relatif aux référents extralinguistiques ; contexte situationnel relatif aux situations sociales conventionnelles ; contexte interactionnel relatif aux actes de langages engagés entre les participants.

[21] "The situation of the speech act is initiated by a stimulus and composed of a topic, time, place, and the participants."

Or, au-delà du simple décalage spatio-temporel, toutes les coordonnées situationnelles sont affectées par la traduction.

En premier lieu, le changement de destinataires peut induire des modifications profondes, comme dans les versions contractées du *Reader's digest*, les traductions de vulgarisation de textes spécialisés, la localisation de CD-ROM[22], etc. Comme le note Fortunato Israël (in M. Lederer & F. Israël, 1991 : 23), en littérature, la traduction d'une œuvre ancienne laisse le choix entre plusieurs options : on peut établir une traduction savante ou « philologique » (Ladmiral, 1986 : 34) pour spécialistes, proche de la littéralité, ou bien une œuvre autonome visant un large public, privilégiant le bonheur de la lecture.

Par ailleurs, le traducteur, en tant qu'émetteur d'un nouveau message, choisit parfois de laisser son empreinte, voire de s'approprier le texte traduit : comme Perrot d'Ablancourt (1606-1664) qui, avec ses « belles infidèles », recherchait l'élégance et le raffinement de la langue, plus que la fidélité à l'original. La composante intentionnelle, déterminant la fonction générale du message, peut ainsi subir des modifications importantes. Par exemple, on traduit des romans, dans certaines éditions bilingues, afin de servir de support didactique, et non dans le seul but de fabriquer une œuvre littéraire, comme l'auteur du texte original. De même, la traduction d'un passage de *Mein Kampf* pourra servir les besoins de la recherche historique, sans intention de propagation idéologique. Le « *vouloir-dire* » du texte et de sa traduction sont, dans ce cas, certainement différents :

> « Le sens qui, dans la communication directe est déjà médiatisé à deux niveaux, une fois comme énoncé, une autre fois comme message - se trouve, dans la traduction, médiatisé une fois supplémentaire à chacun de ces deux niveaux. La traduction, en effet, est le processus dans lequel se trouvent confrontés deux idiomes et deux vouloir-dire. » (Pergnier, 1993 : 50)

[22] Notons que dans ce type de traduction, il est toujours possible de considérer séparément l'opération de traduire, *stricto sensu*, des autres transformations (résumer, adapter, recréer, développer, etc.). Mais cette séparation nous paraît cependant artificielle, car le processus traductionnel englobe fréquemment toutes ces transformations dans un seul et même mouvement. De toute manière il n'existe pas de traduction « pure », mais seulement des traductions plus ou moins littérales, suivant les contraintes des codes linguistiques et les choix de traduction.

Vassili Koutsivis (in Lederer & Israël, 1991 : 143) donne un exemple de divergences sur le plan du vouloir-dire[23], dans le cas particulier des traités internationaux, où certains accords gardent dans leur forme la trace de points de vue irréductibles, superficiellement réconciliés par le jeu des ambiguïtés linguistiques :

> « M. Tabory mentionne plusieurs exemples de 'Terminology as a Tool of Diplomacy' et, parmi d'autres, la divergence délibérée des versions françaises des accords d'armistice de 1949 qu'Israël avait conclus avec le Liban (où le terme anglais " boundary " est traduit par " frontière ") et avec la Syrie (où " boundary " est traduit par " limite ") et qui était due apparemment à la sensibilité de la Syrie à l'égard du mot " frontière ". »

Les problèmes soulevés par la notion de *vouloir-dire* sont nombreux, car ils sont liés aux structures complexes de l'intersubjectivité. En élargissant le concept au message en tant qu'*acte de langage* (en y adjoignant donc une composante de « *vouloir-faire* »), nous dirons que tout message à une *intention*. On peut distinguer, d'une part, l'intention initiale de l'émetteur et, d'autre part, l'intention perçue par le récepteur. Pour Sager (1994 : 67) l'objet du message (« *purpose* ») est le produit de l'intention de l'émetteur et de l'attente (« *expectation* ») du récepteur. La forme du message émis et l'interprétation reçue dépendent alors des représentations que chaque participant a de son interlocuteur.

Le jeu de miroir peut nous mener assez loin : chaque participant à une représentation de sa propre représentation chez son interlocuteur. Dans un dialogue, on a une certaine idée de son interlocuteur, une certaine idée de l'image qu'on offre à cet interlocuteur, et une certaine idée de la façon dont les messages sont décodés. Ce jeu complexe est fondamental dans la communication, puisqu'il décide de l'ajustement des codes et des références communes, afin d'assurer une bonne intercompréhension. La fonction métalinguistique intervient à ce niveau, dans la mesure où les participants s'y accordent sur le sens de leurs énoncés, par des explicitations complémentaires et des corrections additionnelles (les expressions du type « Je voulais dire que… » sont nombreuses et récurrentes). Cette complexité s'applique aussi à l'implicite, à tout ce qui n'est pas formulé et peut éventuellement s'exprimer par le truchement de codes paraverbaux : attitude, intonation, etc., pour l'oral, typographie, graphisme, type de papier, médium audiovisuel, etc. pour

[23] On peut prendre le terme de « vouloir-dire » comme la forme pré-verbale de l'intention communicative, consciente ou inconsciente. Comme le note Laplace 1994 : 198) commentant Seleskovitch : « Le vouloir-dire est la cause du discours, le sens en est la finalité »

l'écrit. Ces systèmes sémiotiques parallèles, conjugués à l'implicite, à la présupposition, sont des vecteurs expressifs puissants permettant de situer la valeur d'un énoncé dans la dimension interactionnelle : politesse, dérision, ironie, raillerie, respect, connivence, humour, etc. La traduction doit bien entendu rendre compte de ces *valeurs*, qui font partie intégrante du message.[24]

Au-delà de ces échos intersubjectifs, ces notions de vouloir-dire et d'intention d'un message soulèvent d'autres questions : où faut-il les situer ? au niveau de la motivation consciente de l'émetteur ? au niveau de son inconscient ? au niveau de la forme de l'énoncé, c'est-à-dire de façon immanente à la manifestation linguistique ? au niveau de l'énoncé pris globalement, en tenant compte des coordonnées situationnelles et du contexte culturel ? Nous verrons qu'il n'existe pas de réponse définitive à ces questions : pour reprendre les termes d'Umberto Eco (in Siri Nergaard, 1995 : 138), leur résolution temporaire dépend du « pari » interprétatif qui détermine les choix du traducteur.

Ainsi, même lorsqu'on cherche à restituer l'intention originale, les « déplacements » intentionnels sont inévitables, et ne sont pas toujours conscients. On ne traduit pas la *Divine Comédie* aujourd'hui comme au XIVe siècle, et de toutes façons, on ne la lit pas de la même manière que les contemporains du *Trecento*.[25] D'une part, la fonction de l'acte

[24] Des erreurs d'interprétation à ce niveau donnent parfois lieu à des dérapages cocasses. Jung Wha Choi (in Lederer & Israël, 1992 : 205) cite un exemple de nature linguistique, mais qui aurait pu concerner n'importe quel type de protocole interactionnel : lors d'un sommet franco-coréen en 1989, le « président de la République coréenne, en rencontrant M. Fabius et avant d'entrer dans le vif du sujet, a dit à son interlocuteur : " Vous êtes très beau ". Ce que l'interprète a traduit tel quel. » Il s'agissait bien entendu d'une formule de politesse…

[25] Il n'est pas nécessaire de comparer des styles littéraires très éloignés dans le temps. Les deux versions suivantes d'un passage de l'*Idiot* de Dostoïevski montrent à quel point les habitudes littéraires, à 30 ans de distance, peuvent influencer la forme et la fonction même de la traduction :
« Je vendis les titres, empochai l'argent, mais, au lieu d'aller chez Andreïev, je filai tout droit au Magasin Anglais où je choisis une paire de boucles d'oreilles avec deux brillants, chacun à peu près de la grosseur d'une noisette. Il me manquait quatre cents roubles, mais je dis qui j'étais et l'on me fit crédit. » *Traduction d'Albert Mousset, 1953, La Pléiade*
« Les billets, je les vends, je prends l'argent, mais j'oublie le comptoir d'Andreev, j'ai filé, j'avais plus que ça en tête, chez les Anglais, et là, je claque le tout pour une paire de pendants d'oreilles, un petit diamant dans chaque, comme une paire de noisettes, comme ça, un peu, il manquait quatre cents roubles, je dis le nom, ils me croient. » *Traduction d'André Markowicz, 1993, Actes Sud*
Dans la première version, les concordances de temps sont respectées, suivant les règles habituelles du français écrit. De même la syntaxe des phrases est classique. La traduction de Markowitcz ne respecte aucune de ces conventions : d'après lui toutes les traductions précédentes de Dostoïevski ont « embelli » artificiellement le texte original, au risque de l'affadir : « Dostoïevski détestait l'élégance, en particulier celle des Français. Il écrivait avec véhémence, sans se soucier de la syntaxe ni des répétitions. Les premières traductions ont tout fait pour policer ce style.»

littéraire change à chaque époque, d'autre part la fonction même de l'acte de traduire se transforme, quand même le traducteur cherche toujours à *servir* l'original. Jacqueline Risset (1985 : 19), dans la préface à sa traduction de la *Divine Comédie*, formule son parti pris en critiquant l'aspect solennel de certaines traductions :

> « Le choix originaire de cette traduction-ci peut se formuler comme la décision de faire émerger - ou de tenter de faire émerger - un aspect de la *Comédie* généralement voilé par l'opération de traduire : la vitesse du texte de Dante ; l'opération de traduire instituant le plus souvent une solennité paralysante, et une sorte de sens supplémentaire, équivalant à ceci : " Le poème que vous lisez, lecteurs ignorants, est un chef d'œuvre de l'humanité ; par conséquent, découvrez-vous, et ne le regardez pas en face... " D'où l'inexorable ennui... »

Pour Risset (1985 : 19), une certaine *littéralité* permet de redécouvrir la vigueur originelle du texte. A propos des traductions d'Antoine Berman, elle note :

> « Traduction littérale (...) grâce à une prosodie moderne, débarrassée de ses symétries obligatoires, et sensible à l'éclat, au tranchant d'un grand texte oublié (assoupi, recouvert par ses propres gloses), et disposant - grâce à la distance accumulée et à la réflexivité de l'époque tardive - une écoute capable d'ouvrir l'accès (à partir du présent) à ces textes lointains devenus proches sans cesser de faire briller leur distance. »

La préface, l'introduction critique, les annotations et le corps de la traduction assument ici une fonction particulière, explicitement distanciée par rapport à l'original : « ouvrir l'accès » au texte de Dante. D'ailleurs cette traduction figure dans une édition bilingue où elle accompagne le texte original. La fonction ancillaire de la traduction est prédominante dans ce cas de figure : son *intention* est d'accompagner, d'éclairer, de faire émerger certains aspects, au même titre que les notes et le reste de l'appareil critique.

On voit donc que le traducteur n'est en rien esclave de l'intention de l'auteur du texte original, même s'il peut chercher à l'intercepter, afin de bâtir sa propre interprétation. Comme l'écrit Seleskovitch, l'intention de l'auteur n'est en somme qu'une « hypothèse » de travail (1984 : 132) :

> « Jamais le traducteur ne traduira *il y a un courant d'air* par *ferme la fenêtre*, car s'il faisait cela, il se situerait dans l'intention de l'auteur et non dans le sens qu'exprime son dire ; l'intention n'est pas explicite et reste donc à l'état

---

(Markowicz cité par Hubert Nyssen, cf. à l'adresse : http://www.liberation.com/quotidien/portrait/port990115.html).

> d'hypothèse, le sens est clairement désigné. En affirmant que l'objet de la traduction est le sens, c'est-à-dire un sémantisme appliqué au discours, je n'entends nullement dire que traduire consiste à expliciter des intentions hypothétiques. »

Ce qui est déterminant, c'est donc la *fonction* du message traduit. A l'intérieur des grandes catégories énumérées par Jakobson (phatique, conative, référentielle, émotive, métalinguistique, poétique), chaque message constitue un mixte fonctionnel original qui découle de l'intention communicative. Les variétés de mixtes fonctionnels sont infinies, et nombreux sont inscrits dans le vocabulaire courant : informer, convaincre, séduire, vendre, effrayer, impressionner, démontrer, expliquer, ordonner, déclarer, émouvoir, décourager, etc. Ces mixtes fonctionnels sont liés de manière préférentielle à des genres textuels et à des situations normalisées : discours à l'assemblée, pièce de théâtre, article de presse, publicité, récit, lettre commerciale, etc. Or, *a priori*, aucune règle n'impose à une traduction de respecter le mixte fonctionnel du texte original, dans la mesure où la traduction peut elle-même assumer des fonctions diverses : rendre accessible un texte éloigné, aider à l'apprentissage d'une langue, divertir, donner lieu à une création littéraire nouvelle, vulgariser une théorie scientifique, etc.

Si la communication directe monolingue est passablement compliquée, on peut dire que la traduction est alors doublement compliquée. La dimension intersubjective où se situe l'*intention* devient, dans l'acte traduire, un vertigineux palais des glaces où se mire toute une galerie de participants virtuels : le traducteur est confronté à l'auteur du message source avec ses intentions conscientes et / ou inconscientes, à ses destinataires supposés, aux virtualités interprétatives du message, aux intentions du commanditaire de la traduction, aux destinataires de la traduction, et finalement … à lui-même, puisqu'il a ses propres motivations esthétiques, idéologiques, didactiques ou autres… Les questions les plus simples : qui veut dire quoi ? à qui ? pourquoi ? peuvent donc admettre des réponses extrêmement variées pour un même message original.

Enfin, au-delà de la *situation* de communication impliquant tous les participants, il nous faudra considérer le *contexte* élargi, où s'opèrent parallèlement des déplacements importants, qui imposent de déraciner le message source de son contexte de départ pour le

transplanter dans le contexte d'arrivée. C'est un lieu commun de signaler que tout texte porte la marque de sa culture d'origine. Comme le note malicieusement Borges :

> « Le scrupuleux roman historique de Salammbô, dont les héros sont les mercenaires des guerres puniques, est un typique roman français du XIXe siècle. »[26]

C'est pourquoi F. Israël (in Lederer & Israël, 1991 : 24) compare la traduction à un « processus d'acculturation » : « Ainsi tout transfert suppose une décontextualisation et engendre inévitablement un processus d'acculturation pour que soit assurée, dans un réseau de relations autre que le contexte de production, la lisibilité de l'ouvrage. » Cette acculturation implique-t-elle nécessairement une perte ? est-elle toujours possible ? Les problèmes soulevés à ce niveau concernent la compatibilité des codes, au sens large, de départ et d'arrivée.

Entre le message source et le message cible, ce n'est pas seulement la langue, c'est la totalité des codes et des repères culturels qui peuvent changer. Aux coordonnées situationnelles sont liées des dimensions pragmatiques plus larges, d'essence sociale, qui sont aussi en relation directe avec l'interprétation du texte :

- la langue, avec toutes ses variations selon divers axes sociolinguistiques aux dénominations plus ou moins barbares : diatopiques (dialectes), diastratiques (sociolectes, technolectes, idiolectes), diachroniques ;
- les données culturelles au sens large : mode de vie, coutumes, valeurs, références communes, symboles, croyances, etc. ;
- les contenus encyclopédiques, dépendants peu ou prou, malgré leur vocation universelle, d'une culture spécifique ;
- les typologies textuelles : discours, récit, roman, manuel, notice, fiche technique, poème, article scientifique, etc.

Tous ces aspects intéressent la traduction : mais le passage d'un système de coordonnées à un autre implique inévitablement des distorsions. Entre le message original et sa traduction s'instaure une dialectique du même et de l'autre, de l'identité et de la différence. Tous les choix de traductions reposent dans cet équilibre : certains éléments demeurent et d'autres changent, en fonction de la définition donnée à la *relation d'équivalence* que le traducteur cherche à établir entre l'original et sa reproduction.

I.1.2.2 L'ancrage phénoménologique de l'interprétation

Dans une perspective très générale, fidèle à la maxime de saint Jérôme, on peut affirmer que l'équivalence traductionnelle revient à une certaine *conservation du sens*.

Reste que si tout locuteur perçoit de manière immédiate le sens du mot « sens », le linguiste ne peut se contenter de recourir à un terme aussi vague et ambigu sans en préciser le contenu. Or, le *sens* est une notion multiforme, difficile à circonscrire sur le seul plan linguistique dans la mesure où elle intègre tout à la fois des composantes sémantiques, cognitives, culturelles et pragmatiques. Le sens est l'objet central d'une discipline : l'herméneutique.

Un bon début d'élucidation consiste à mettre à jour ce que le sens n'est pas. D'un point de vue structural, Rastier (1989 : 16) oppose le *sens* à la *signification*, la seconde étant la projection linguistique du premier, résultant d'une réduction du contexte de la communication. Par rapport au sens, la signification d'une unité lexicale est un produit refroidi, institutionnalisé, résultat de la sédimentation des usages de la langue, et de la structuration des *valeurs*, au sens saussurien, par le système. Suivant l'axe d'une distinction parallèle opérée par Rastier (1989 : 96), la signification est relative au contenu du *type* et non de l'*occurrence* : « La relation signifiant/signifié n'est donc fixée que pour les signes-types ; pour les signes-occurrences, elle varie indéfiniment en fonction des contextes. Il faudra en tenir compte pour caractériser la linéarité des signifiés. »

Bien sûr, la langue change, la signification évolue, mais elle possède une lenteur, une inertie, qui l'autorise à reposer sagement dans les grammaires et les dictionnaires, même provisoirement. Elle a la pesanteur du social, elle est lourde de ses milliers de locuteurs :

---

[26] J. L. Borges, Lune d'En Face, in *Oeuvres Complètes*, Bibliothèque de la Pléiade, Gallimard, Paris, I.

« La langue n'est pas l'actuel du langage, elle est mémoire, elle est un *avant* de la communication, non la communication elle-même ;(...) » (Pergnier, 1993 : 253). Tandis que le sens est le produit d'une création instantanée, fluide, instable, synthèse précaire et labile des codes et des éléments situationnels qui convergent vers son bref miroitement.

Une architecture textuelle peut cependant, par sa construction progressive du sens, lui conférer des fondations, une charpente et une certaine assise. Dans un texte, le contexte linguistique des unités est figé. Cela ne suffit pas cependant à en figer le sens, définitivement. L'univocité évoquée par Rastier (1989 : 16), est toute relative lorsque le texte est abstrait d'une situation de communication :

> « Retenons que la signification *immanente à la phrase* est un artefact des linguistes, et qu'elle demeure inévitablement équivoque. Alors qu'à l'inverse, son sens, réputé oblique, difficile à cerner, reste généralement univoque dans un contexte et une situation donnés ».

Car le sens textuel n'est ni immanent ni transcendant au texte : il est le résultat d'une conjecture, d'un parcours interprétatif, d'une *lecture*. Le texte entretient un double rapport de « *différance* » (du verbe différer, nous empruntons le terme à Derrida) avec le scripteur d'une part, et le lecteur d'autre part. En tant que texte, il hérite *de jure* des composantes du message *d'origine*, mais gagne *de facto* une certaine autonomie. La lecture est en quelque sorte un nouveau point de départ, une nouvelle forme d'écriture, susceptible de prêter aux signes écrits un sens inédit. Tout le roman d'Eco, *le Pendule de Foucault*, est basé sur cette *différance* : une simple note de blanchisserie devient potentiellement le message codé d'un complot universel. A l'instar de Pascal, dans sa tentative de donner *du sens* à l'univers, Eco (in Nergaard, 1995 : 138) pose le sens comme l'enjeu d'un *pari* : « Le sens que le traducteur doit trouver, et traduire, n'est déposé dans aucune langue pure. C'est seulement le résultat d'une conjecture interprétative. Le sens ne se trouve pas dans un *no language's land* : il est le résultat d'un pari. »

Le sens est donc indissociable de l'horizon phénoménologique des *sujets* impliqués dans la communication. Il est un produit de l'intersubjectivité. Il est indissociable de l'*intentionnalité* des participants de la communication. Ainsi, même si la matérialité de la manifestation linguistique lui confère une incarnation objective, descriptible scientifiquement selon un point de vue structural (une *explication*), le sens du texte reste fondamentalement lié à l'opération subjective de *compréhension*. Paul Ricoeur (1994) décrit ainsi cette ambivalence : « Devenu un objet autonome, le texte se situe à la charnière

de la compréhension et de l'explication et non sur leur ligne de rupture. » Mais cette ambivalence n'est pas symétrique. Elle se fonde sur la compréhension qui « ne cesse de précéder, d'accompagner et d'achever les procédures explicatives. » La compréhension constitue un « *noyau irréductible* » qui prend sa source dans l'acte d'un sujet : « (...) l'explication ne saurait se substituer au noyau de compréhension qui reste le cœur de l'interprétation des textes. Par noyau irréductible, j'entends les phénomènes suivants : d'abord, la provenance des significations les plus autonomisées d'une intention de signifier qui est l'acte d'un sujet. » Cette composante intersubjective, évoque ce que Alan Melby (1995) appelle « *agency* », notion qu'on pourrait traduire en terme phénoménologique par *intentionnalité*, notion qui subsume à la fois la capacité de faire des choix librement décidés (*intentio volitive*) et la réceptivité au sens (« sensitivity to meaning », *intentio cognitive*), et qui distingue selon lui de manière fondamentale le traducteur humain du traducteur mécanique.

En tant que le sens est toujours le produit d'une interprétation, on peut donner le nom d'*exégèse* à l'opération du traducteur qui précède l'écriture du texte de destination. L'exégèse détermine les orientations synthétiques du traducteur quant aux divers niveaux d'équivalence retenus pertinents pour l'opération de traduire. Elle désigne le travail de recherche et d'analyse qui, en fonction des coordonnées pragmatiques du texte, tend à reconstruire l'interprétation du texte par rapport à son référentiel de départ. La notion de « fidélité », dans cette recherche des origines, implique donc l'intervention active d'un *sujet*, créateur d'une nouvelle interprétation :

> « (...) le concept de fidélité est lié à la conviction que la traduction est un type d'interprétation (comme le résumé, la paraphrase, l'évaluation critique, la lecture à haute voix d'un texte écrit). Et l'interprétation doit toujours viser, même si elle part de la sensibilité et de la culture du lecteur, à retrouver non pas l'intention de l'auteur, mais l'intention du texte, ce que le texte dit ou suggère par rapport à la langue dans lequel il est écrit et relativement au contexte culturel où il est né »[27] (Eco, in Nergaard, 1995 : 123)

[27] Ma il concetto di fedeltà ha a che fare con la persuasione che la traduzione sia una delle forme dell'interpretazione (come il riassunto, la parafrasi, la valutazione critica, la lettura ad alta voce di un testo scritto) e che l'interpretazione debba sempre mirare, sia pure partendo dalla sensibilità e dalla cultura del lettore, a ritrovare non dico l'intenzione dell'autore, ma l'intenzione del testo, quello che il testo dico o suggerisce in rapporto alla lingua in cui è espresso e al contesto culturale in cui è nato."

La recherche de fidélité est une tâche contradictoire. Entre fonction ancillaire et rôle créateur, entre contrainte et liberté, le traducteur est un intermédiaire invisible et omniprésent, à la fois maître et serviteur, pivot de cette communication à double détente qu'on nomme traduction : « Toute traduction est appropriation, bonne ou mauvaise, et cette appropriation est autant le résultat d'une contrainte que l'affirmation d'une liberté. » (Israël in Lederer & Israël, 1991 : 18).

I.1.2.3 Les niveaux d'équivalence

Même si le *sens* reste une notion indissociable d'un certain ancrage phénoménologique, dépassant le cadre du présent travail, sa projection peut être précisée dans une perspective pragmatique élargie, suivant différentes dimensions sémiotiques. En effet, il est souhaitable d'articuler le jugement d'*identification* entre deux messages, jugement global porté par un sujet plongé dans une situation de communication, avec les *niveaux d'équivalence* caractérisant des relations d'identité sur le plan des codes et des référentiels socialement partagés : langues, cultures, contenus encyclopédiques, etc. Par rapport au code linguistique, Fuchs (1981 : 57) établit cette distinction, entre « équivalence en langue » et « identification dans l'usage de la langue » :

> « En effet, (…), s'il est indéniable que les séquences linguistiques ne peuvent jamais être totalement identiques (mais seulement équivalentes), en revanche il est également incontestable que , dans leur activité paraphrastique en situation, les sujets les traitent comme si elles étaient identiques : dans une telle perspective, il n'y aurait donc pas lieu d'opposer identité et équivalence, mais il faudrait au contraire distinguer, pour mieux les articuler, équivalence en langue et identification dans l'usage de la langue. »

Ainsi entendue, l'équivalence peut être décrite le long d'un continuum de niveaux d'interprétation qui « s'enchaînent les uns aux autres de façon continue ; ils vont du plus linguistique (c'est-à-dire lié aux formes) au moins littéral (c'est-à-dire à l'interprétation la plus libre). » (Fuchs, 1981 : 128). Fuchs distingue quatre plans, « qui constituent autant de paliers dans ce continuum interprétatif ; nous les appellerons respectivement : plan " locutif ", plan " référentiel ", plan " pragmatique " et plan " symbolique " ».

Ces considérations concernent la paraphrase au même titre que la traduction. Le niveau « locutif » est celui des équivalences linguistiques, où se situent par exemple les valeurs sémantiques des unités lexicales : le rôle d'un dictionnaire bilingue est de répertorier ce type d'équivalence. Fuchs rattache de même au plan locutif les « opérations énonciatives », « par lesquelles le sujet assigne à l'énoncé un certain nombre de valeurs référentielles (de temps, d'aspect, de modalité, de détermination, etc.), c'est-à-dire ancre l'énoncé par rapport à sa situation énonciative, à son " moi-ici-maintenant " » (Fuchs, 1981 : 129). Les références qui ne sont pas inscrites dans le système ressortissent au plan « référentiel », comme dans la paraphrase de « il est venu ici le mois dernier » : « il est venu à Paris en janvier ». Au niveau « pragmatique », l'auteur situe les valeurs illocutoires et perlocutoires du message. Quant au niveau « symbolique », « l'interprétation se fonde sur tout ce à quoi la séquence peut renvoyer symboliquement. » (Fuchs, 1981 : 130), et Fuchs y place les figures de style (métaphores, métonymies), les formes littéraires (parabole, allégorie) ou encore les interprétations psychanalytiques.

Bien sûr ces niveaux coexistent, et même si la paraphrase – ou la traduction – privilégie, dans la relation d'identité, un plan particulier, l'interprétation ne peut les considérer séparément : « Pour pouvoir paraphraser, il faut nécessairement opter pour un niveau d'interprétation ; ceci étant, on sait bien qu'en fait, un énoncé n'est pas toujours à entendre à un seul niveau (…) » (Fuchs, 1981 : 134).

Cependant, la prise en compte de différents niveaux ne nous met pas à l'abri de l'écueil d'une vision trop étroitement mono-fonctionnelle de la langue. Par exemple, pour Quine, la traduction est possible dans la mesure où il y a « identité entre les systèmes linguistiques » (« *identity across linguistic systems* »[28]). Il énumère trois types d'identité inhérents au fonctionnement des systèmes linguistiques :

[28] Quine, W. (1960), *Word and Object*, Cambridge, Mass.: MIT Press, p. 130

1. Identité conceptuelle
2. Identité référentielle (connotative et dénotative)
3. Identité basée sur les universaux du langage (qui requiert l'établissement de grammaires contrastives complètes afin de déterminer les identités structurelles entre les langues mises en jeu).

Cette stratification est quelque peu restrictive, car limitée à des aspects cognitifs, logiques et linguistiques. Or il est rare que la fonction référentielle soit exclusive de toutes les autres, même dans les textes scientifiques ou techniques. La notion d'équivalence traductionnelle s'étend de droit à toutes les dimensions constitutives de la communication. Sager (1994 : 200), dans cette perspective, énumère trois axes nécessaires à la compréhension en profondeur d'un message :

> « Pour notre modèle de traduction, nous postulons trois dimensions de la compréhension, les dimensions cognitives, pragmatiques et linguistiques, chacune comportant différents niveaux. Les traducteurs doivent être en mesure de lire et de comprendre à certains niveaux cognitifs et à tous les niveaux pragmatiques et linguistiques. » [29]

Prenons garde de ne pas y voir des plans phénoménologiquement séparés, coexistant côte à côte : ces niveaux portent des éclairages différents sur les mêmes phénomènes. La maîtrise de ces différents niveaux suppose, de la part du traducteur, d'avoir une connaissance de codes linguistiques déterminés (objets de la linguistique) et de certaines constructions conceptuelles (objets de la psychologique cognitive et de la logique). Quant à la dimension pragmatique, il faut admettre qu'elle occupe une position particulière, en ce qu'elle s'intéresse à l'*usage* de ces objets. Dans ce que Bar-Hillel appelait ironiquement la « poubelle pragmatique », on met commodément tous ce qui ne rentre pas strictement dans le psychologique et le linguistique. C'est que la dimension pragmatique est largement indéterminée : elle constitue en quelque sorte l'*horizon* phénoménologique où des individus sociaux interagissent au moyen de codes appréhendés par leurs structures cognitives.

---

[29] "For our model of translation we postulate three dimensions of understanding, the cognitive, the pragmatic and the linguistic, with several levels each. Translators must be able to read and understand at particular cognitive levels and at all pragmatic and linguistic levels."

Nous laisserons au pragmatique cette indétermination, en lui conférant un statut de synthèse englobante, embrassant les deux autres niveaux, comme le montre la figure 2 :

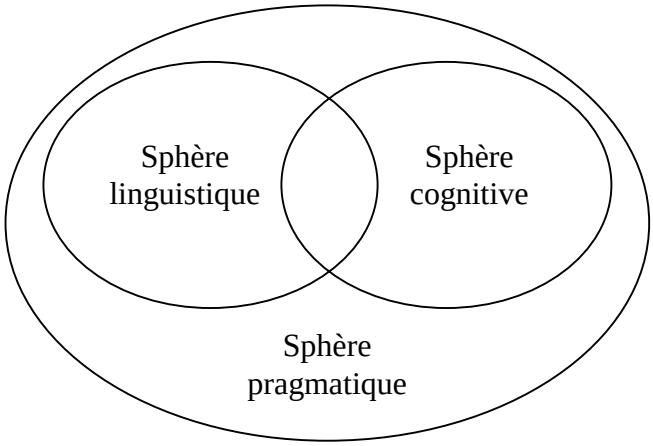


figure 2 : *les trois niveaux phénoménaux impliqués dans la compréhension d'un message*

Dans une communication spontanée entre des participants présents, cette synthèse est instantanée, immédiate. Mais dans l'opération de traduire, le repérage des composants pragmatiques nécessite un effort réflexif de construction : reconstruction du *sens* du texte original (ce que nous avons appelé *exégèse*), construction d'un *sens* nouveau avec le texte cible, après repérage des différentes instances de la communication : destinataires, exigence du commanditaire, etc. Même l'interprétariat n'est pas étranger à ce type de médiation : une interprétation se prépare à l'avance, et exige un effort de documentation, une connaissance minimale des participants, etc.

Pour le traducteur, le respect des différents niveaux d'équivalence, pragmatique, cognitif ou linguistique n'est donc pas strictement le produit d'un « jugement », comme dans « le jugement de paraphrase » évoqué par Fuchs. Dans le processus traductionnel, la définition du (ou des) niveau(x) d'équivalence est le résultat d'un choix. Dans l'espace des coordonnées du message, le traducteur détermine à l'avance le plan sur lequel l'original et sa projection vont se superposer. La relation d'équivalence n'est pas l'objet d'une reconnaissance, comme dans la paraphrase, mais le fruit d'une élaboration.

Dans cette construction, circonscrire les éléments pragmatiques est donc la première étape à accomplir. D'après Christine Durieux (in Lederer & Israël, 1991 : 171) : « (...) la

mission finale d'un texte est un paramètre qu'il importe d'identifier et dont il faut tenir compte. Le traducteur doit impérativement déterminer la ou les missions du texte à traduire avant de se lancer dans l'exécution de la traduction. »

En fonction de ces missions, nous distinguerons entre plusieurs niveaux d'équivalence, non exclusifs les uns des autres : équivalence dynamique, équivalence conceptuelle / référentielle, équivalence stylistique.

I.1.2.3.1 Equivalence dynamique

Nous empruntons à Nida (1969 : 142) l'expression d'*équivalence dynamique* :

> « C'est pourquoi l'équivalence dynamique doit être définie en fonction du degré de similitude entre la réponse des récepteurs du message cible et la réponse des récepteurs du message source. Cette réponse ne peut jamais être identique, dans la mesure où les repères historiques et culturels sont parfois trop éloignés ; il faut cependant chercher à atteindre un haut degré d'équivalence dans les réponses, sans quoi la traduction n'a pas remplit sa mission. »[30]

Ce qui est visé par ce type d'équivalence, c'est la *réponse* des récepteurs de la traduction. L'équivalence recherchée se situe donc au niveau global de l'*effet* produit sur le récepteur. De façon similaire, Seleskovitch (1992 : 157) oppose les aspects statiques et dynamiques du texte, isomorphes à l'opposition classique *ad verbum / ad sensum* : « il faut voir si l'on parle du texte tel qu'il existe dans sa forme figée – le stylistique et le cognitif – ou si l'on parle de l'effet produit par le texte qui est le notionnel et l'émotionnel. »

Par *effet* on ne désigne rien d'autre que le produit de la fonction du message, considéré du côté des récepteurs. Il concerne toute la dimension des actes de langage. La liste de ces actes n'est pas fermée : tout au plus peut-on en citer les principales catégories, à l'instar d'Austin (1962), quand il fait référence aux classes d'énoncé *verdictifs*, *exercitifs*, *promissifs*, *comportatifs*, *expositifs*, etc. dans sa douzième conférence.

La notion d'effet incluant toutes les composantes pragmatiques, elle n'a de pertinence qu'au niveau de la *globalité* du texte. Ainsi, pour certaines traductions, la recherche d'équivalence dynamique peut impliquer un véritable décrochement de la lettre

de l'original. Les unités locales – unités lexicales, syntagmes, phrases – n'auront pas nécessairement d'équivalent dans le texte cible.

Dans cette recherche de réponse chez les récepteurs, les contraintes linguistiques peuvent s'effacer totalement derrière les enjeux extralinguistiques. Dans le domaine de la traduction juridique, Koutsivis (in Lederer & Israël, 1991 : 142) note qu'« (...) il faut traduire de manière telle à obtenir dans l'espace juridique réglementé par le texte traduit les mêmes résultats que ceux fournis par le texte original dans son propre espace. » Le traducteur assume alors la fonction de l'émetteur initial, et s'attache à respecter son intention : « La ligne directrice dans notre travail est de se mettre à la place du législateur. » (in Lederer & Israël, 1991 : 152).

Lorsque le texte cible doit subir des transformations profondes pour être recevable dans la langue d'arrivée, et produire le « même » effet, on parle d'*adaptation*. Cette liberté par rapport aux structurations linguistiques locales est donc l'aboutissement, du point de vue pragmatique, d'une recherche de proximité. Pour G. Bastin (in Lederer & Israël, 1991 : 165) « l'adaptation est une fidélité, un choix de fidélité non plus à un vouloir-dire limité à des énoncés linguistiques, à des expressions linguistiques mais bien à une visée globale, ce que M. Koutsivis appelait la volonté de l'auteur. » On constate que la définition d'un certain niveau d'équivalence, comme ici le vouloir-dire global de l'auteur, se fait au détriment du niveau linguistique. Dans le choix d'un niveau d'équivalence, il n'est pas rare que la conservation d'un niveau implique le sacrifice des autres.

[30] "Dynamic equivalence is therefore to be defined in terms of the degree to which receptors of the message in the target language respond to it in substantially the same manner as the receptors in the source language. This response can never be identical, for the cultural and historical settings are too different, but there should be a high degree of equivalence of response, or the translation will have failed to accomplish its purpose."

I.1.2.3.2 Equivalence dénotative : conceptuelle et / ou référentielle

Les énoncés linguistiques entretiennent, avec les réalités extralinguistiques, des rapports extrinsèques de *désignation*. Lorsqu'une traduction recherche l'équivalence au niveau du *designatum*, nous parlerons d'équivalence *dénotative*[31].

Le niveau dénotatif est fondamental dans la traduction, dans la mesure où il est directement lié, d'un point de vue psycholinguistique, aux facultés de compréhension et de rétention mémorielle. Pottier (1992 : 68) note ainsi que « toute compréhension d'un texte en [langue naturelle] est de nature conceptuelle (Co), et son siège est la *mémoire*. Or la mémoire enregistre le sémantisme dans un code délié des langues naturelles. On oublie très vite dans quelle langue une information a été reçue. Les noèmes tentent de caractériser la rétention mémorielle. » Par sa « noémique », Pottier désigne un substrat cognitif, indépendant de telle ou telle langue, qu'on peu identifier de manière universelle indépendamment des cultures. Une expérience citée par Rastier (1994 : 73) montre de quelle manière, lors de la compréhension, le plan de l'expression s'efface par rapport au plan du contenu : « Psycholinguistiquement, la compréhension est dominée par l'oubli : une phrase n'est pas finie que son début est déjà oublié. Par exemple, si dès la fin de la phrase *La neige dévalait furieusement la pente* on demande aux sujets qui viennent de la lire si le mot avalanche s'y trouve, un sur cinq environ acquiesce. »

Le plan dénotatif, dans la mesure où il est détaché d'une formulation déterminée, est le lieu privilégié de l'équivalence traductionnelle. Pottier (1992b : 15) le présente comme un stade intermédiaire impliqué dans le processus psychologique de la traduction :

> « Quand un sujet parlant traduit d'une langue naturelle 1 dans une langue naturelle 2 (ou qu'il manipule une langue pour faire des « résumés »), il fait le parcours sémasiologique, passe par un univers conceptuel non verbal (niveau dit noémique), pour devenir émetteur suivant le parcours onomasiologique, vers un nouveau message en langue 2 »

Pottier (1992b : 112) propose une intéressante décomposition des différents stades, allant du conceptuel au linguistique, intervenant dans la génération d'un énoncé. Il prend

[31] L'expression « désignationnelle » nous paraissait un néologisme un peu lourd. Mais c'est bien l'identité du *designatum* qui est visée.

l'exemple de l'arrestation de trois voleurs par des gendarmes, et décompose l'élaboration de l'énoncé final suivant divers degrés d'abstraction. Les niveaux qu'il dégage sont :

1. *L'événement conceptualisé*. Ce que Pottier appelle le « schème analytique », représentable dans un langage formalisé, par exemple en recourant à des représentations graphiques abstraites. On se situe alors sur un plan conceptuel, interlinguistique.
2. *Le passage en langue naturelle* : détermination du « schème d'entendement ». Dès ce niveau, les structurations linguistiques entrent en jeu. On y opère le choix des lexèmes, par exemple : « *gendarme* », « *voleur* », « *arrêter* ». Les structures actancielles (au niveau sémantique) en découlent.
3. *La prédication* : choix d'une « base de vision ». On choisit un élément comme base et on lui affecte un prédicat. On effectue une sélection entre toutes les paraphrases diathétiques : « *Les gendarmes ont arrêté trois voleurs* » ou « *Trois voleurs ont été arrêtés par les gendarmes* »

Les autres étapes de transformation linguistiques sont facultatives :

4. *La topicalisation* : « *Les gendarmes, ils ont arrêté trois voleurs*. »
5. *La focalisation* : « *Ce sont les gendarmes qui ont arrêté trois voleurs*. »
6. *L'impersonnalisation* : « *Il y a les gendarmes qui ont arrêté trois voleurs*. »
7. *La réduction d'actance* : « *Trois voleurs on été arrêtés*. »

Une traduction pourra, suivant les possibilités qu'offre la langue cible, conserver ou supprimer l'un de ces niveaux optionnels. L'important, pour Pottier, est que le « schème analytique » soit conservé. Traduire reviendrait alors à refaire le chemin, à partir d'un même schème analytique, dans une autre langue. Nous pensons, avec Jakobson (1963 : 82), que ce chemin existe toujours, car « toute expérience cognitive peut être rendue et classée dans n'importe quelle langue existante ». Nous verrons cependant (cf. p. 69) que des différences culturelles peuvent induire des problèmes de traductibilité.

Notons que ces conceptions ont inspiré une famille de systèmes de TA, basés sur la représentation abstraite des textes à l'aide d'un langage « pivot » (cf. p. 194).

La nature des objets (au sens large : chose, état de chose, événement, …) dénotés soulève un certain nombre de problèmes théoriques qui sont au cœur des enjeux de la traduction. Par *référent*, on désigne des réalités extralinguistiques, qui échappent par essence à l'emprise d'une langue particulière et ne posent théoriquement pas de problèmes vis-à-vis de la traduction. Par exemple, un nom propre désignant un référent particulier possède, en principe, toujours un équivalent dans une autre langue : lui-même (via quelques aménagements « morpho-phono-graphémiques »). Dans ce cas, l'identité référentielle et l'extériorité du rapport entre signe et référent garantit un fonctionnement purement dénotatif indépendant du système de la langue, rapport arbitraire qu'on reproduit sans mal dans n'importe quelle autre langue. Mais il s'agit d'un cas limite : le rapport de désignation est rarement pur, et transite presque toujours par des signifiés, des concepts et des représentations[32].

Or nous pensons que pour appréhender correctement la notion d'équivalence dénotative, il est nécessaire de distinguer entre les *signifiés* linguistiques d'une part, les *concepts* et les *référents* d'autre part. Dans la mesure où nous plaçons ces deux derniers pôles hors langue, ils peuvent assumer des rôles équivalents par rapport aux unités linguistiques : à la différence des signifiés, ils sont à l'extrémité de la relation de désignation.

Certes, concepts et référents entretiennent entre eux des rapports complexes (les concepts représentent, subsument, classent, abstraient, etc.), dont traitent les disciplines connexes aux sciences du langage : psychologie cognitive, logique ou philosophie. Du fait de l'assimilation entre concepts et signifiés linguistiques, on a longtemps attribué des

[32] Même dans le cas des noms propres, la relation de désignation peut impliquer toutes sortes de filtres, car ils s'insèrent toujours à l'intérieur de systèmes de classification. Comme le note Claude Lévi-Strauss (1962 : 258), les noms propres représentent des « *quanta de signification* » : « En tant qu'ils relèvent d'un ensemble paradigmatique, les noms propres forment la frange d'un système général de classification : ils en sont à la fois le prolongement et la limite. ». « A cet égard, ils ne diffèrent pas foncièrement des noms d'espèces, comme l'atteste la tendance du langage populaire à attribuer, selon leur espèce respective, des noms humains aux oiseaux. En français, le moineau est Pierrot, le perroquet Jacquot, la pie Margot, le pinson Guillaume, (…) » (1962 : 241)

statuts dissymétriques au concept et au référent, par rapport à l'usage langagier : dans l'histoire de la pensée occidentale, on assiste une remarquable pérennité de la conception triadique (Rastier, 1990) qui fait du concept un intermédiaire indispensable entre les mots et les choses[33].

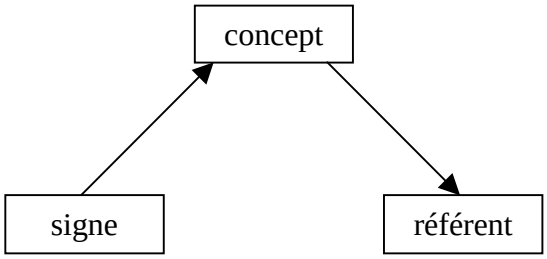


figure 3 : *les trois pôles de la triade sémiotique, d'Aristote à Lyons*

Dans le modèle triadique, le *sens* d'un signe est assimilé au *concept* (ou « représentation », ou « idée ») qui lui est lié. Par exemple, au signifiant « bleu » on attache un concept (une plage de longueur d'onde dans le spectre lumineux) ou une représentation mentale, qui réfère lui-même à une classe de phénomènes (les objets réfléchissant cette tranche de longueur d'onde).

Avec la philosophie analytique, cette relation intermédiaire est escamotée, au détriment du concept : chez le premier Wittgenstein (*Tractatus Logico-philosophicus*, 1961 : 29-107), le nom est en *référence directe* avec l'objet : « 3.203 Le nom signifie l'objet. L'objet est la signification du nom ». La signification est alors issue du rapport d'isomorphisme entre monde et langage. Le monde est décrit par l'ensemble des

[33] Aristote, décrit ainsi la relation triadique : « La parole est un ensemble d'éléments symbolisant les états de l'âme, et l'écriture un ensemble d'éléments symbolisant la parole. Et, de même que les hommes n'ont pas tous le même système d'écriture, ils ne parlent pas tous de la même façon. Toutefois, ce que la parole signifie immédiatement, ce sont les états de l'âme qui, eux, sont identiques pour tous les hommes ; et ce que ces états de l'âme représentent, ce sont des choses, non moins identiques pour tout le monde. » Aristote, *Peri Hermenias* I, 16a, 3-8.
On retrouve la même conception dans la formule scolastique : « *vox significat mediantibus conceptibus* ». Chez Arnaud et Nicole (1683), le triangle sémiotique se traduit dans des termes similaires : « Les mots sont des sons distincts et articulés dont les hommes ont fait des signes pour marquer ce qui se passe dans leur esprit », et les idées « représentent des choses ou des manières des choses » (II,1). Ogden et Richard (1923 : 11) puis Lyons (1978 : 83) reproduisent la triade et assignent un rôle spécifique à la sémantique linguistique, qui s'intéresse à la relation entre les mots « *Symbol* » et les choses « *Referent* », indépendamment de la médiation des concepts « *Thought of Reference* ».

propositions : « 4.001 La totalité des propositions est le langage ». Et ce rapport est de nature binaire : une proposition bien formée peut avoir deux sens : vrai ou faux. « 4.024 comprendre une proposition c'est savoir ce qui arrive quand elle est vraie ». Selon cette conception, le langage vient en quelque sorte redoubler la réalité. Le sens consiste en une « relation projective » de la proposition vers le monde. Il y a identité formelle entre la « forme logique » et la « forme de la réalité » car « 3.032 représenter par le langage 'quelque chose de contraire à la logique', on ne le saurait pas plus que représenter en géométrie par ses coordonnées une figure contraire aux lois de l'espace (…) ». La référence directe shunte le niveau conceptuel : pour que « bleu » désigne une certaine classe de phénomènes, il suffit d'appeler ces phénomènes « bleus », sans qu'il soit nécessaire de recourir à une quelconque construction mentale.

Que les concepts occupent une fonction d'interface entre la langue et le monde, ou que la relation soit directe, ces conceptions ont cependant un point commun : la langue est considérée de manière générale, dans son universalité, sans tenir compte des spécificités propres à tel ou tel système de signifiés linguistiques. Reprenons l'exemple du mot « bleu ». C'est une unité polysémique : suivant les contextes, les classes de référents désignés sont variables : un « bleu de travail » est un vêtement, « avoir un bleu » c'est avoir un hématome, on dit « c'est un bleu » pour un soldat débutant, « les bleus » désignent l'Equipe de France, etc. Un certain nombre d'expressions imagées tournent autour du « bleu »: « une peur bleue » est une très forte peur, « je suis bleu » signifie, en Gironde, « je n'en reviens pas » (on dirait, dans d'autres régions « je suis vert »). Il est clair que tous ces emplois ne renvoient pas aux mêmes concepts ni aux mêmes référents. Et pourtant, ils semblent tous partager une signification commune, un dénominateur commun. S'agit-il d'un concept ? qu'en est-il de ce noyau de signification dans d'autres langues ?

On peut rétorquer que les différents emplois de « bleu » n'impliquent pas nécessairement une identité de contenu, même partielle. Mettons provisoirement de côté le problème de la polysémie, sur lequel nous reviendrons au chapitre I.1.3.2. Considérons une seule acception de *bleu* : la couleur. Dans les emplois suivants : *le ciel bleu, des yeux bleus, un reflet bleu, une mer bleue, une chemise bleue, une encre bleue, une peau bleue, un steak bleu* est-on sûr de toujours désigner la même tranche de fréquence lumineuse pour chaque classe d'objet ? en d'autres termes, s'agit-il toujours du même concept ou de la même

représentation ? Il semble que la valeur du mot *bleu*, considéré hors contexte, ne soit pas si facile à caractériser : on peut lui rattacher une nébuleuse de concepts ou d'idées. Par concept, on désigne cependant une entité relativement stable, comme le concept de triangle. La représentation triadique semble négliger cet aspect.

Or, ces variations ont des conséquences directes dans la traduction : *bleu* peut-être traduit, en italien, tantôt par *azzurro,* tantôt par *blu* ou *celeste* (mais la liste n'est pas fermée). Les particularités de l'unité *bleu*, ses possibilités d'emploi, sa polysémie n'ont pas de correspondance dans les autres langues : elles appartiennent au français. Bourquin (1993 : 29) propose d'appeler *notion* les contenus immanents à une langue donnée, à la différence des *concepts*. *Bleu* renverrait donc à une *notion*, à la différence du concept physique de couleur. Pour nous situer sur un plan linguistique, nous préférons le terme saussurien de *signifié*, qui renvoie à toutes les virtualités liées à la *valeur* de l'unité au sein du système (peut être pourrait-on réserver le terme de *notion* à des concepts culturellement marqués).

Ainsi, pour appréhender correctement le problème de la désignation et ses prolongements dans l'activité traductionnelle, il faut restituer aux signifiés linguistiques leur juste position dans le schéma global du fonctionnement sémiotique de la langue.

Afin de déterminer ce qui appartient au linguistique et de situer la part relative des objets connexes, Rastier (1990 :21) propose une configuration à quatre pôles, reproduite figure 4.

Cette autonomisation du sémantique par rapport au référent et au concept permet d'appréhender les structurations sémantiques particulières à chaque langue : oppositions sémiques, classes, structures hiérarchiques, polysémie, etc. Réciproquement, le plan conceptuel gagne une certaine autonomie, dans la mesure où il vise des représentations et constructions mentales indépendantes d'une langue spécifique. Dès lors, le concept, comme le référent, est à considérer comme une entité « objective », puisqu'il est un point de convergence intersubjectif. En tant qu'objet, il est le résultat de diverses opérations de fabrication, et ses contours sont plus ou moins nets en fonction de ces opérations : abstraction à partir de l'expérience (induction, synthèse), ou bien définition théorique de ses propriétés.

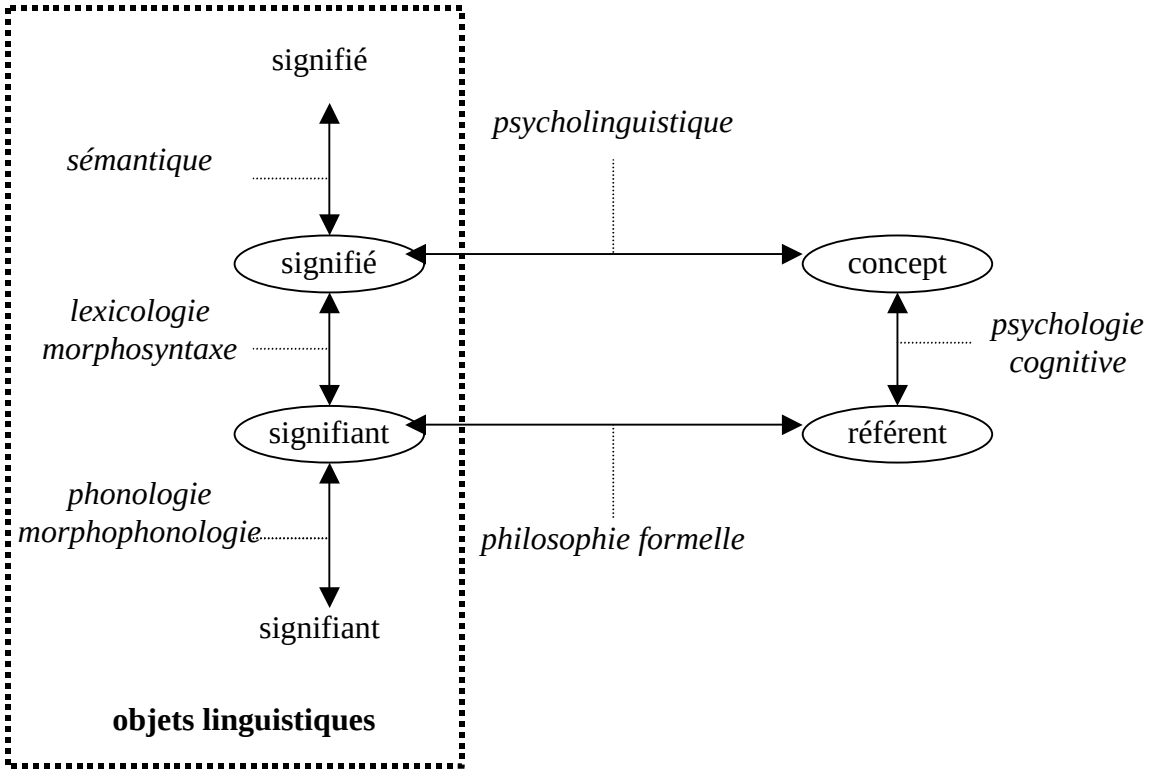


figure 4 : *relations entre objets linguistiques et extralinguistique*

Ainsi, les concepts peuvent être considérés sous plusieurs angles, comme représentations spontanées ou comme des objets intellectuels rattachés à des disciplines scientifiques, des techniques, des pratiques sociales, juridiques, etc. Par exemple, le savoir scientifique *construit* et *définit* des concepts à partir d'un arrière plan théorique et de protocoles expérimentaux. Dès lors, on ne peut plus confondre le signifié de « eau » avec le concept physico-chimique H_2O. De même, on pourra donner une définition précise du concept « bleu » à l'intérieur d'une discipline précise : pour un graphiste, le bleu fondamental correspond à des fréquences lumineuses précisément étalonnées.

Il apparaît que l'équivalence dénotative ne vise pas à la conservation des signifiés linguistiques, inséparables de leur système d'origine, mais de leurs *designata*. Les divergences entre organisations linguistiques différentes ne sont donc pas un obstacle à la traduction. Le fait que le gaélique désigne le « vert » et le « bleu » avec une même unité « glas », n'empêche pas le locuteur irlandais de référer aux mêmes phénomènes que les Français.

Ce que le schéma de la figure 4 ne montre pas, c'est que concepts et référents occupent des positions interchangeables en tant que *designata*. En effet, les relations entre les signes linguistiques et les objets extralinguistiques peuvent fonctionner suivant la tripartition de Pierce (indice, icône et symbole) quel que soit le type d'objet : référent, ou concept. Ainsi, le rapport symbolique de *désignation* n'est pas réservé à la relation entre un signifiant et un référent, mais existe aussi entre signifiant et concept. C'est par exemple, ce qui lie un *terme* à un concept : un rapport extrinsèque, conventionnel, qui peut être partiellement motivé mais qui reste arbitraire dans son principe.

L'équivalence dénotative concerne donc aussi bien des désignations de concepts que de choses. Du point de vue de langue, concepts et choses sont équivalents. En effet par *choses,* on entend en général des classes d'objets similaires. Or le rapport entre concepts et classes d'objets est réciproque : de manière analytique, la définition d'un concept (« compréhension ») détermine aussi une classe d'objets (« extension ») – à l'inverse, de manière synthétique, la donnée d'une classe d'objets similaires constitue un concept, par l'abstraction des propriétés communes de ces objets. Le sens *concepts* → *objets* correspond à une sémantique des conditions nécessaires et suffisantes ; le sens *objets* → *concepts* est lié aux structurations prototypiques du langage (où une classe est définie par la ressemblance avec le représentant le plus typique, cf. Kleiber, 1990).

Distinguer les signifiés des concepts, ou des référents, pose cependant quelques problèmes. Entre les organisations de signifiés, et les organisations conceptuelles on observe un *continuum* : il n'existe pas de ligne de démarcation nette entre les deux types de contenu. Il n'existe pas de concepts qui ne soient pas entachés du « péché originel » d'avoir été formulé à l'intérieur d'une langue et d'une culture donnée. Malgré les tentatives de Leibniz, de Frege, de Russel, il n'existe pas non plus de langue parfaite susceptible de forger des concepts avec une absolue universalité. Walter Benjamin[34] (1923 & in Nergaard, 1993 : 227) émet l'hypothèse d'une « langue pure » (« *Reine Sprache* ») constituant l'horizon vers lequel toutes les langues convergent :

[34] Dans son Introduction à sa traduction des *Tableaux parisiens* de Baudelaire. Titre original :« Die Aufgabe des Überstzers »

> « En chacune d'elles, prise comme un tout, est entendue une seule et même chose, qui toutefois n'est accessible à aucune d'elles prise dans sa singularité, mais seulement à la totalité de leurs intentions réciproquement complémentaires : la langue pure. »[35]

Mais comme le souligne Eco (in Nergaard, 1995 : 134) : « cette Reine Sprache n'est pas une langue. »

D'après Guy Bourquin (in P. Bouillon & A. Clas, 1993 : 29), les discursivités fortement conceptualisées peuvent être considérées comme de simples « notations d'un invariant interlingue externe, transcendant ». Il s'agit là d'une idéalisation, caractérisant certains usages restreints de la langue. C'est pourquoi, ce que Eco (in Nergaard, 1995 : 136) dénomme le *contenu propositionnel* ne s'applique qu'à une catégorie limitée d'énoncés :

> « La notion de contenu propositionnel vaut seulement pour les énoncés très simples exprimant des états du monde : énoncés qui d'une part ne sont pas ambigus (comme dans le cas des figures de rhétorique), et d'autre part ne sont pas auto-réflexifs, i.e. produits pour attirer l'attention non seulement sur leur signifié mais aussi sur leur signifiant. (...) la notion de contenu propositionnel s'applique à des processus de dénotation et non de connotation ».[36]

Que les unités dénotées soient des concepts ou des référents, il ne faut pas perdre de vue notre perspective pragmatique englobante : ce qui fait le contenu dénotatif d'un texte, ce ne sont pas les signes qui le composent, mais le rapport établi par l'interprétation, dans une situation d'énonciation précise, entre ces signes et les réalités extralinguistiques désignées. Car le parcours à accomplir entre un signe et sa désignation n'est jamais établi à l'avance : l'aboutissement de la désignation peut-être très proche du signifié (ce que Pottier, 1992a : 48, appelle un *orthonyme*), ou bien transiter par des représentations

---

[35] Nous ne disposons que de la traduction italienne de cette citation :"Piuttosto, ogni affinità metastorica delle lingue si basa sul fatto che in ciascuna di esse, presa come un tutto, è intesa una sola e medesima cosa, che tuttavia non è accessibile a nessuna di esse presa singolarmente, ma solo alla totalità delle loro intenzioni reciprocamente complementari : la pura lingua."

[36] "La nozione di contenuto proposizionale vale dunque solo per enunciati molto semplici che esprimono stati del mondo e che, da un lato, non siano ambigui (come accade con le figure retoriche) e, dall'altro, non siano autoriflessivi, tali cioè da essere prodotti ai fini di attrarre l'attenzione non solo sul loro significato ma anche sul loro significante (come i valori fonici o prosodici). (..) la nozione di contenuto proposizionale si applica ai processi di denotazione ma non a processi di connotazione."

conceptuelles, un contexte référentiel spécifique, et s'éloigner de l'orthonymie par l'application de filtres successifs, métaphoriques ou métonymiques.

Ainsi, des chemins différents peuvent conduire au même *designatum*. Pour reprendre un exemple célèbre de Frege : *étoile du soir* et *étoile du matin*, bien que dénotant des concepts différents peuvent désigner un même référent, Vénus.

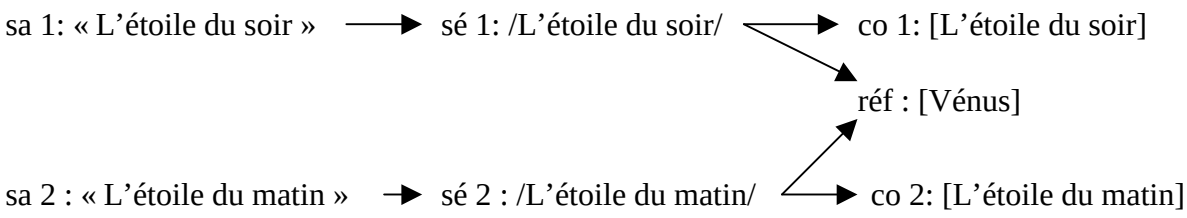


figure 5 : *variété des parcours dénotatifs* :
deux concepts pour un même référent

C'est l'interprétation qui doit déterminer le point d'aboutissement de la désignation : ce peut être le concept [étoile du soir], en référence à la première étoile apparaissant dans le ciel vespéral ; mais dans un autre contexte, l'*étoile du soir* pourrait aussi bien désigner une chanteuse de cabaret.

On peut imaginer une configuration inverse à la précédente, où le même concept correspond à des référents différents. C'est le cas des expressions françaises et italiennes, rigoureusement symétriques : *les bleus* et *gli azzurri*, le *drapeau tricolore* et *la bandiera tricolore*, *transalpin* et *transalpino*, *le tour* et *il giro*.

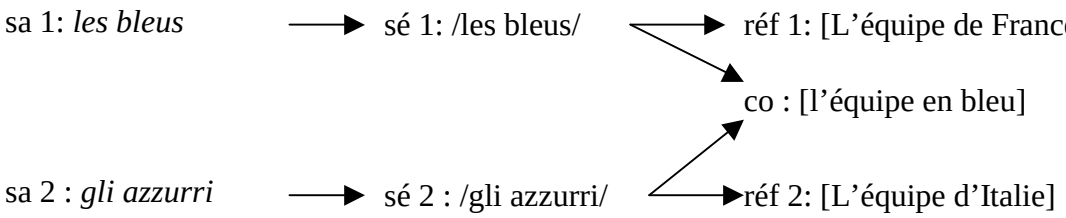


figure 6 : *variété des parcours dénotatifs* :
deux référents pour un même concept

Cette identité conceptuelle force les traducteurs à emprunter l'expression étrangère : on emploie fréquemment *squadra azzurra* pour désigner les bleus transalpins.

Dans le mécanisme de désignation, le traducteur doit non seulement tenir compte du *designatum* visé *in fine*, mais aussi du chemin parcouru. Lorsque le même chemin peut-être emprunté dans une autre langue, il peut être préférable de s'y conformer. Par exemple, on pourra préférer traduire *étoile du matin* par *Morgenstell* en allemand ou *stella mattutina* en italien, même si c'est la planète Vénus qui est visée. Car le parcours dénotatif n'est jamais innocent, et il imprègne toujours, d'une certaine manière, la chose désignée : certes, les expressions *le vainqueur d'Austerlitz* et *le vaincu de Waterloo* font toutes deux référence à Napoléon : mais s'agit-il vraiment de la *même* personne ? Ne peut-on considérer deux personnages, Napoléon conquérant et Napoléon humilié, en des moments et des circonstances différentes de l'histoire – et ces deux Napoléon sont-ils tout à fait identiques au Bonaparte des campagnes d'Italie ?

Le Wittgenstein du *Tractatus* et le Wittgenstein des *Cahiers* sont à ce point différents qu'on parle des « deux Wittgenstein ». En fait, Il y a toujours un risque à conférer une extériorité absolue au *designatum* par rapport à sa dénomination. Entre sens et référence, comme l'affirme Fuchs (1981 : 65), il n'y a pas de solution de continuité :

> « l'existence en langue de termes comme les déictiques ou les anaphoriques manifeste une propriété capitale du langage, à savoir le fait qu'il n'y a pas d'un côté le sens (exclusif de toute référence) et de l'autre la référence (directe et extra-linguistique), mais un continuum entre les deux. »

Le *designatum* est avant tout l'extrémité d'un cheminement qui va de la langue au monde, d'une trajectoire sans cesse reparcourue par l'interprétation du récepteur, qui reste toujours libre d'en infléchir la courbe et d'en déterminer le point d'arrêt.

Ainsi, même si la réalité objective fournit la base référentielle et conceptuelle, le *primat* sur lequel le traducteur s'appuie pour restituer l'équivalence dénotative, cette « objectivité » est toujours construite, c'est un produit stabilisé à la limite de l'intersubjectivité. Comme le note Sager (1994 : 153), il y a toujours un « *transfert conceptuel* », un filtrage inévitable de la subjectivité du traducteur : « le contenu peut être filtré par la perception propre du traducteur vis-à-vis du sujet, et comporter l'addition de ses interprétations personnelles ou de ses choix là où il les sent adaptés à la situation de

médiation. »[37]. Cette forme de transformation conceptuelle n'est pas particulière à la traduction, car elle est engagée dans toute situation de communication au niveau du processus de *compréhension*.

En conclusion, la traduction force le traducteur à la résolution d'un antagonisme : l'équivalence dénotative se situe au-delà de la langue et de ses signifiés particuliers, sur le plan des référents et des contenus conceptuels – mais dans la mesure où l'interprétation du texte est conditionnée par les formulations linguistiques, le traducteur doit aussi tenir compte du cheminement parcouru, et tenter de le restituer, afin de conserver les virtualités interprétatives du texte original.

I.1.2.3.3 Equivalence connotative et / ou rhétorique et / ou expressive

Nous avons vu que la *désignation* instaure une relation complexe susceptible d'être interprétée à différents niveaux (c'est pourquoi nous avons cherché à éviter l'usage du terme *dénotation*, qu'on emploie parfois dans un autre sens, en logique comme en linguistique[38]).

A ces niveaux s'ajoute un autre plan, parfois jugé « secondaire », qui peut pourtant être déterminant dans la définition de l'équivalence traductionnelle : c'est le plan des connotations.

On donne souvent une définition négative de la connotation : « on considère que la connotation est un concept servant à nommer tout ce qui, dans la signification, ne relève pas de la dénotation (ces deux notions se partageant exclusivement la totalité du champ de

---

[37] "The content may even be filtered through the translators' own perception of the subject matter, with the addition of personal interpretations or selections where felt appropriate for the situation of the mediation"

[38] En logique, chez J.S. Mill, le couple connotation / dénotation recouvre généralement l'opposition compréhension / extension des concepts– comme le remarque Kerbrat-Orecchioni (1977 : 13), cette acception conduit, dans certains emplois « à une antonymie parfaite entre la connotation logique et la connotation linguistique ». En sémantique structurale on oppose les sèmes dénotatifs, considérés comme *distinctifs*, aux sèmes non-distinctifs (*virtuels* chez Pottier, *afférents* chez Rastier). Ces sèmes connotatifs forment ce que l'on appelle parfois la *connotation référentielle*. Mais dans la mesure où il n'y a pas de différence de nature entre ces deux types de sèmes, qui forment un *continuum* (du plus au moins pertinent, plus ou moins distinctif, plus ou moins saillant, etc.), nous ne retiendrons pas cette acception de l'opposition dénotation / connotation.

la production du sens) »[39]. Communément, on rattache à la connotation des significations secondes, des valeurs axiologiques supplémentaires qui s'ajoutent à la signification première : valeurs affectives, sociolinguistiques, poétiques, etc. Pour Nida (1969 : 91), la connotation concerne l'impact émotionnel des formes linguistiques : « Cet aspect du sens qui a trait à nos réactions émotionnelles est appelé sens connotatif. »[40]

De manière plus générale, on peut considérer la *connotation* comme le produit d'un système second, dont la face signifiante implique des structures du système de la langue. C'est le sens donné par Hjelmslev, pour qui une langue de connotation forme un méta-système : il définit le signe connotatif comme un signe dont le signifiant est un signe de la langue (englobant le couple Sa / Sé) :

$$\frac{(\frac{Sa}{Sé})}{Sé_{connoté}}$$

Mais ce schéma est restrictif si l'on considère que le signifiant de connotation implique toujours un signe complet, avec son signifiant et son signifié. Kerbrat-Orecchioni insiste sur le fait que le signifiant de connotation est « plus diversifié que celui dont relève la dénotation », et se construit tantôt sur le Sa, tantôt sur le Sé de dénotation, tantôt sur les deux, tantôt sur des signifiants de nature spécifique. Le schéma de Hjelmslev doit donc être interprété dans une vision plus large : « L'idée Hjelmslevienne, pour être correcte, doit être reformulée de la façon suivante : " les codes connotatifs présupposent les codes dénotatifs " » (Kerbrat-Orecchioni , 1977 : 85)

Ainsi, les signifiants de connotation sont hétérogènes et extrêmement diversifiés car ils peuvent s'attacher à toutes les facettes de la langue et de son usage : variations phonétiques, prosodiques, syntaxiques, lexicales, variables situationnelles, etc. Par ailleurs, les signifiés de connotations sont spécifiques (ironie, humour, variables sociolinguistiques, valeurs poétiques, etc.) et largement autonomes par rapport au contenu dénoté. Kerbrat-Orecchioni (1977 : 229) résume ainsi le rapport connotation / dénotation :

> « - Une unité de dénotation a pour support un signifiant lexical ou syntaxique, et elle apporte des informations explicites sur l'objet dénoté par le message verbal.

[39] P. Dubois, article « connotation », in *Encyclopaedia Universalis* 2000

[40] "This aspect of the meaning which deals with our emotional reactions to words is called connotative meaning."

> - Une unité de connotation réutilise à son profit n'importe quel élément du matériel de la dénotation, et ses informations sont de nature et / ou de statut différent. Tantôt elles concernent autre chose que le référent immédiat du discours (sujet énonciateur, situation d'énonciation, type particulier d'énoncé), tantôt elles viennent enrichir sous forme de valeurs additionnelles et suggérées, le contenu dénotatif. »

En ce qui concerne les contenus connotés, Dubois[41] retient quatre classes principales :

- les « *connotations stylistiques* », relatives aux variantes diatopiques (origine géographique), diastratiques (origine sociale, langue de spécialité, genre textuel), diachroniques.

- les « *connotations énonciatives* », qui renseignent sur l'affectivité, les valeurs, l'idéologie du locuteur.

- les « *connotations associatives* », incluant les jeux de langage et les figures de rhétorique (calembour, allitération, rime, métaphore, métonymie, synecdoque, oxymore, allusion, ironie, etc.), opérant des rapprochements sur le plan de l'expression (homonymie, paronymie) ou sur le plan du contenu (synonymie, antonymie, hyponymie, hyperonymie, etc.).

- « *connotations implicites* », « qui désignent les présupposés, les inférences et les non-dits des énoncés comme porteurs de significations particulières. »

Les deux premières catégories peuvent être ramenées à ce qu'on appelle parfois la « connotation d'usage », le signe héritant d'un sens lié à la situation où il est habituellement employé. Les connotations associatives et implicites véhiculent aussi des significations secondaires, mais à partir d'une élaboration plus complexe qui met en jeu l'interprétation globale du message.

En ce qui concerne la connotation d'usage Greimas (1970 :101) propose une classification plus complète, où il distingue deux vastes champs de signification :

[41] article « connotation » in *Encyclopaedia Universalis*, 2000

- « l'homme dans la société », concernant « paraître de la société et paraître de l'homme ». Ces connotations découlent de l'articulation des communautés linguistiques (stratification sociale, fonctionnelle, découpage géographique, etc.) et d'une typologie sociale des individus (taxinomie de caractères et de types psychologiques, dépendant de facteurs culturels).

- les espaces sémiotiques extérieurs. Certains espaces sociaux, comme le droit ou la religion, connotent les unités linguistiques de valeurs spécifiques : autorité, pouvoir, prestige, magie, transcendance. La poésie ou le mythe produisent des effets de sens et de vérité, de nature à véhiculer l'émotion, le mystère ou le sentiment du sacré… Enfin certains « objets » culturels sont connotés comme lourds de sens ou dotés de puissance : noms célèbres, proverbes, événements historiques, …

On constate que la connotation récupère tout ce qui ne rentre pas strictement dans le rapport de désignation (de même que la pragmatique hérite de tout ce que la linguistique rejette hors de son champ d'étude). Par souci de précision, nous préférons employer le terme de connotation dans une acception plus restreinte, limitée à la *connotation d'usage*.

D'après Greimas, les contenus connotés font apparaître une « nouvelle dimension » de la sémiosis : « tout objet sémiotique existe simultanément sur *le mode de l'être* et sur *le mode du paraître*. » (1970 : 99) En tant qu'il désigne, c'est-à-dire en temps qu'il *est* un signe, le signe porte une certaine signification. Mais en tant qu'il *apparaît* dans un certain contexte, le signe se charge des valeurs liées à ce contexte et au code auquel il appartient, et reçoit un sens supplémentaire, sans rapport immédiat avec sa signification « normale ».

Ce qui pose problème, vis-à-vis de la traduction, c'est la capacité de tout construit sémiotique à signifier *simultanément* suivant les deux modalités évoquées par Greimas :

- sur le mode de « l'être » du signe, à travers le rapport de désignation,

- sur le mode du « paraître » du signe, à travers les situations d'usage, les sonorités, les rapprochements possibles avec d'autres signes, etc. A ce dernier niveau, nous

situons les connotations d'usage, les effets rhétoriques et les phénomènes liés au plan de l'expression (rythmes, sonorités, etc.).

Pour la traduction, il existe donc deux plans porteurs de signification, qui induisent des niveaux d'équivalence concurrents et parfois antinomiques :

– l'équivalence dénotative ;

– l'équivalence connotative et / ou rhétorique et / ou expressive.

Dans les textes littéraires, il n'est pas rare que le plan connotatif devienne prioritaire par rapport au strict jeu des dénotations. Eco (1995 : 126-127) donne l'exemple du passage suivant, tiré du *Pendule de Foucault* :

« *Ma tra picco e picco si aprivano orizzonti interminati - al di là della siepe, come osservava Diotallevi...* »[42]

Dans ce passage, l'auteur indique qu'il n'est pas nécessaire de traduire « *siepe* » (fr. « haie ») littéralement. En fait « *Al di là della siepe* » fait référence à un vers de *l'Infinito* de Leopardi. Eco veut signifier que son personnage, Diotallevi, rapporte sa perception du paysage à son expérience de la poésie. Pour la traduction française, Eco préconise donc que cette connotation poétique soit transposée dans un vers de Baudelaire : « Mais entre un pic et l'autre s'ouvraient des horizons infinis - *au-dessus des étangs, au-dessus des vallées* [43], comme observait Diotallevi. » (trad. Jean-Noël Schifano). Ce qui importe, pour l'auteur, c'est la référence intertextuelle, de nature connotative : « Au-delà d'autres choix stylistiques plus évidents, chaque traducteur a inséré un renvoi à un passage de sa propre littérature, identifiable par le lecteur visé par la traduction. »[44] (1995 : 127)

[42] U. Eco (1988 & 1996), *Il Pendolo di Foucault*, Milano, Bompiani, p. 355
[43] La référence implicite concerne le poème de Baudelaire, « *Elévation* », tiré des *Fleurs du mal*
[44] " Al di là di altre evidenti scelte stilistiche, ciascun traduttore ha inserito un richiamo a un passo, della propria letteratura, riconoscibile dal lettore a cui la traduzione mirava. "

La conservation des valeurs attachées au plan de l'expression est sans doute plus problématique, car il est clair que les transformations impliquées par la traduction altèrent d'abord la forme des signifiants. La maxime de saint Jérôme condense un des enjeux les plus universels de la traduction : il ne faut pas rendre la *parole* mais le *sens*, il faut sacrifier le plan de l'expression, intrinsèquement lié à la langue d'origine, pour sauvegarder le plan du contenu, d'un statut ontologique jugé plus élevé[45]. Cette précellence du contenu est cependant parfois malmenée, notamment dans les textes littéraires, et plus spécialement la poésie. Comme l'a noté Jakobson, l'expression poétique est auto-référentielle, dans la mesure où la forme perd son caractère arbitraire et devient en elle-même porteuse de sens. Les plans du contenu et de l'expression n'y sont plus parallèles, disjoints, mais s'interpénètrent et se mêlent dans un jeu de courts-circuits qui travaillent simultanément sur toutes les facettes de la langue. Lorsque Ponge écrit, dans les *Mots et les choses* :

« Les ombelles ne font pas d'ombre, mais de l'ombe, c'est plus doux »

la poésie jaillit d'une soudaine collision : le calice des ombelles, finement ajouré, projette tout à la fois une certaine qualité d'ombre et un nouveau mot de la langue française. Signifiant et signifié s'unissent tout à coup, enfin réconciliés, à travers la douceur d'un oubli. « L'ombe » est enfantée dans le frisson fragile de l'ombre et du soleil, par la disparition fugace d'une lettre qui tombe, et roule par terre.

La poésie, en pétrissant la matière sémiotique, réunit les jumeaux, Sa et Sé, en frères siamois désormais inséparables. Comment réaliser cette même fusion dans une autre langue ? Il est certes possible de transposer des procédés stylistiques (rythmes, sonorités, rimes) et rhétoriques (tropes, figures) dans des formules plus ou moins équivalentes. Mais dès lors, comme le signale Israël (in Lederer & Israël, 1991 : 22), « la traduction littéraire ne peut être que la mise au point d'une autre œuvre, c'est-à-dire d'un texte autonome de même statut ». Le choix de mettre en avant certains aspects devient alors une délicate question d'interprétation, car toutes les composantes du message ont voix au chapitre. Faut-il respecter la littéralité, faut-il privilégier le rythme, faut-il conserver les sonorités,

[45] On peut y déceler un écho de la hiérarchie platonicienne, puis chrétienne, subordonnant la chair à l'esprit.

les images, les références culturelles, faut-il restituer l'idéologie de l'auteur ? Aucune de ces questions n'admet de réponse *a priori*.

Notons que les textes scientifiques et techniques ne sont pas exempts d'aspects rhétoriques. Des contraintes de simplicité syntaxique, d'univocité lexicale, de registre sociolinguistique, de phraséologie préétablie imposent à ces textes des critères rigoureux quant à la forme. A la différence des contraintes de l'expression poétique, ces spécificités ne sont pas signifiantes en elles-mêmes, mais il est important de les respecter si la destination du texte traduit le requiert. Dans ce cas il ne s'agit pas tant de rechercher une équivalence formelle stricte, que d'appliquer des systèmes de contraintes formelles propres au message de destination.

I.1.2.3.4 Niveaux d'équivalences et signifiés locaux

Nous avons vu que la notion d'équivalence peut être considérée suivant différents niveaux. A chacun de ces niveaux, les choix de traduction s'inscrivent dans un jeu d'oppositions binaires combinant plusieurs dimensions :

- L'équivalence pragmatique implique une prédominance du tout sur les parties, une prévalence de l'intention et des fonctions du message par rapport aux moyens linguistiques mis en œuvre localement.

- L'équivalence dénotative, plus souvent décomposable au niveau d'unités textuelles inférieures (paragraphes, phrases, syntagmes, unités lexicales) entraîne une subordination des signifiés linguistiques (lexicaux, morphosyntaxiques) par rapport aux designata (référents, concepts) visés. Dans ce cas, la forme de l'expression est secondaire.

- Les équivalences connotatives, rhétoriques et expressives aboutissent au contraire à un recul du plan dénotatif par rapport aux formes de l'expression : les aspects incidents de la construction du sens (figures, registres, rythmes, sonorités, etc.) sont mis au premier plan. Cela ne signifie pas que le plan du contenu est escamoté : ces formes expressives sont signifiantes, mais pas de façon directe

comme dans un simple rapport dénotatif. Ce type d'équivalence implique toujours un « jeu de langage »[46], un mode de signification indirecte, médiatisé par le tissu des relations unissant signifié et signifiant dans le système de la langue (cf. le schéma de la connotation, p. 56).

Dans tous les cas, on constate que les signifiés locaux (i.e. les signifiés des unités lexicales et des structures morphosyntaxiques) n'apparaissent qu'en tant qu'élément de construction, tandis que l'équivalence traductionnelle ne s'intéresse qu'au résultat final : et lorsqu'il s'agit de réaliser la copie d'une œuvre architecturale, briques, moellons et ciment s'effacent bien vite derrière l'aspect général et la fonction de l'édifice. Comme le note Seleskovitch (1984 : 132-133), l'identité du sens est dépouillée de toute forme linguistique : « Lorsque je parle d'identité, je ne parle pas d'identité de moyens, je parle d'identité de résultat ; le sens qui nous reste est un souvenir cognitif, dépourvu de toute forme mais identique en sa teneur informe. »

On assiste à ce que certains considéreront comme un paradoxe : l'équivalence traductionnelle n'implique en rien la conservation du contenu sémantique des unités mises en jeu. De manière très schématique, on peut expliciter la hiérarchisation des différents niveaux par le jeu des oppositions binaires pertinentes pour chaque type d'équivalence (cf. tableau 1). Ainsi, la recherche d'équivalence pragmatique surdétermine les autres niveaux d'équivalence, qui lui sont toujours subordonnés.

[46] Pour Wittgenstein, la relation de désignation occupe une place particulière mais non essentielle face à la diversité et l'hétérogénéité des jeux de langages possibles. Il s'agit seulement d'une phase préparatoire : « dénommer est analogue au fait d'attacher une étiquette à une chose. On peut dire que c'est là la préparation à l'usage des mots » (1961 : 126). L'usage littéraire ou poétique n'est lui-même qu'une forme particulière de jeu, basé sur un ensemble de règles, parmi lesquelles ont trouve le rapport de désignation : « il est d'innombrables et diverses sortes d'utilisation de tout ce que nous nommons ''signes'', ''mots'', ''phrases''. Et cette diversité, cette multiplicité n'est rien de stable, ni de donné une fois pour toute ; mais de nouveaux types de langages, de nouveaux jeux de langage naissent, pourrions-nous dire, tandis que d'autres vieillissent et tombent en oubli. » (1961 : 125).

Oppositions binaires	*Type d'équivalence*		
	dynamique	**dénotative**	**connotative rhétorique expressive**
Globalité du message	+		
Signifié locaux	-		
Designata		+	
Signifié locaux		-	
Formes de l'expression			+
Signifié locaux			-

tableau 1 : *aspect secondaire des signifiés locaux par rapport aux niveaux d'équivalence*

I.1.2.4 Le palier du texte

Notre perspective communicationnelle impose une vision « holiste » du message. Pour paraphraser Israël (in Lederer & Israël, 1991 : 22), dans l'approche globale qui caractérise le plan pragmatique, « l'unité de traduction n'est plus le mot, le syntagme ou la phrase mais le texte tout entier. » Il faut donc s'attacher à la *textualité*, en tant que principe d'unité. A l'instar de Rastier (1987 : 147), par textualité on entendra « ce qui rend un texte irréductible à une suite d'énoncés, voire de mots ».

Pour Jean-Michel Adam (1992 : 21), la complexité des phénomènes mis en jeu interdit d'embrasser l'objet textuel dans une perspective unifiante, et il est plus économique d'adopter une approche « modulaire », consistant en plans complémentaires relativement autonomes. En se situant au seul niveau de ce qu'il appelle « texte » (par opposition aux « discours » régis par des normes sociales) Adam (1992 : 22-28) distingue cinq modules descriptifs : la « visée illocutoire » (qui définit la cohérence extralinguistique du propos), les « repères énonciatifs » (énonciation actuelle / non-actuelle, énonciation universelle du discours logique, théorique - scientifique etc.), la « cohésion sémantique » (rapport fiction/réalité, unité thématique), la « connexité textuelle » (articulation des chaînes de propositions, anaphores, parenthésages, titres, etc.), l'« organisation

séquentielle » (liée à cinq familles de séquences prototypiques : narrative, descriptive, argumentative, explicative et dialogale).

En généralisant les notions de *cohérence* et de *cohésion*, nous distinguerons deux niveaux de description :

– de manière immanente au texte, la textualité se manifeste par sa *cohésion* à laquelle participent un certain nombre d'éléments :

  - la cohésion thématique : par la récurrence d'éléments thématiques prédominants ;
  - la cohésion énonciative : par la donnée d'un ancrage énonciatif particulier, actuel (je, tu, ici, maintenant, …) ou non actuel (il, on, …) ;
  - la cohésion stylistique : par le choix d'un certain niveau de langue, des connotations d'usages et des phraséologies, des rythmes, et d'autres aspects formels ;
  - le réseau des liens anaphoriques et cataphoriques traversant les énoncés ;
  - la connexité des chaînes de proposition : articulations logiques, hiérarchisation, énumérations (si… alors …, d'une part… d'autre part, premièrement …) ;
  - la macrostructure formelle : enchaînement des titres et des parties, mise en page, typographie, ponctuation.

  Ces facteurs de cohésion se situent à la fois sur le plan du contenu et sur le plan de l'expression. Le terme de « grammaire de texte » est parfois employé pour des organisations locales (ce qu'Adam, 1992 : 28, appelle « connexité »), mais la cohésion globale se laisse difficilement formaliser sous la forme d'une grammaire comparable aux structures syntaxiques du niveau phrastique. Car « les structures textuelles sont essentiellement sémantiques. Elles relèvent plutôt de normes et de régularités que de règles (...). » (Rastier, 1990 :10). Pour Rastier, la cohésion sémantique se manifeste par des *isotopies* facultatives, c'est-à-dire des récurrences sémiques non-grammaticalisées ou prescrites par le système (1987 : 157). Il note

ainsi que la cohésion textuelle est avant tout le résultat d'un parcours interprétatif qui permet d'anticiper les isotopies potentielles :

> « En général, on considère l'isotopie comme une forme remarquable de la combinatoire sémique, un *effet* de la récurrence des sèmes. Ici, bien au contraire, ce n'est pas la récurrence des sèmes déjà donnés qui constituent l'isotopie, mais à l'inverse, la présomption d'isotopie qui permet d'*actualiser* des sèmes (…). D'un principe régulateur, l'isotopie devient un des principes constitutifs du sémantisme textuel. » (1987 : 153)

Si la cohésion sémantique est immanente au texte, ce n'est donc qu'au niveau de ses virtualités interprétatives. La cohésion sémantique n'est une propriété intrinsèque qu'en tant que réservoir de significations possibles, de même que le contenu sémantique d'une unité lexicale isolée n'est qu'un faisceau de significations potentielles. Le *sens*, comme produit de l'interprétation, nécessite l'interprétation d'un sujet transcendant.

– A la cohésion, on peut opposer la *cohérence* « définie comme le rapport entre contenu linguistique et référent. » (Rastier, 1987 : 160). Elle est d'ordre extralinguistique, et concerne, entre autres, la cohérence terminologique, l'univocité des désignations, la validité des désignations, la consistance des constructions conceptuelles, la cohérence logique des raisonnements et des argumentations, les valeurs de vérité des assertions référentielles, etc. Comme la cohésion, la cohérence est une résultante du parcours interprétatif, conditionnée par ce qu'Adam appelle la « visée illocutoire » du message, le but de la communication : « La cohérence n'est pas une propriété linguistique des énoncés, mais le produit d'une activité interprétative. L'interprétant prête a priori sens et signification aux énoncés et ne formule généralement un jugement d'incohérence qu'en dernier ressort. » (1992 : 22).

Au-delà des prototypes textuels, identifiés par Adam au niveau des séquences propositionnelles (*narratif*, *descriptif*, *argumentatif*, *explicatif*, *dialogal*), il existe une autre forme de typologie découlant de normes sociales. Ainsi tout texte peut se situer dans un rapport de rupture ou de conformité avec un *genre* conventionnel, auquel il se rattache en fonction de la situation de communication. Ces genres sont extrêmement variés, plus ou

moins strictement codifiés, et concernent tous les secteurs de la vie sociale : manuel scolaire, dissertation, roman, essai, rapport parlementaire, ordonnance médicale, discours de réception à l'Académie Française, histoire drôle, dictionnaire, courriel, charade, scénario de film, bulletin météo, flash d'information, comptine, thèse de doctorat, etc.

Lors de la traduction, le choix du genre textuel surdétermine et conditionne toutes les options suivantes. Comme le précise Sager (1994 : 186), la détermination du genre constitue la première étape du repérage pragmatique :

> « Tandis que les équivalences cognitives et linguistiques sont établies pour une grande part au niveau de la phrase ou d'unités plus petites, les équivalences pragmatiques doivent être déterminées d'abord dans une phase préparatoire, et celles-ci se manifestent au niveau du type textuel avant d'être réalisées pour des unités plus petites en des points appropriés à l'intérieur des documents »[47]

Pour reprendre un terme de Sager (1994 : 83), déjà employé par Van Djik, la « *macrostructure* » du message découle du choix du genre. Cette macrostructure est en quelque sorte un cadre préétabli, socialement normé, régissant l'ensemble des paramètres de la communication écrite. C'est la projection, sur le plan textuel, d'une situation de communication conventionnelle. Pour Sager (1994 : 83), son repérage correspond à la première étape du codage et du décodage d'une communication écrite :

> « Le choix du type textuel est, dans la production d'un texte autonome ou contraint, c'est-à-dire une traduction, l'étape qui détermine la macrostructure du codage lui-même. Le sujet et les objectifs communicatifs étant déterminés, le scripteur doit décider de l'ordre, de la structure et du modèle de communication avant de choisir les techniques de communication et de commencer à écrire. La reconnaissance du type textuel est aussi la première opération réalisée par le lecteur ou le traducteur en face d'un document. »[48]

---

[47] "While the cognitive and linguistic equivalents are mainly established at the level of the sentence or in smaller units during the translation phase, the pragmatic equivalents have to be selected first in the preparation phase and at the level of the text type before being also realised in smaller units at appropriate points in the documents."

[48] "The choice of text type is that step in the production of an autonomous or dependent text, i.e. a translation, which determines the macrostructure of the coding itself. With a topic and a purpose chosen, the writer must decide on the order, structure and pattern of communication before choosing the techniques of communication and actually starting to write. Recognition of text type is also the reader's and translator's fist reaction to a document."

Car le genre textuel cristallise, de manière synthétique, un véritable faisceau de paramètres pragmatiques relatifs à une situation de communication socialement normée. A la typologie des textes, on peut rattacher, entre autres :

- Une typologie des émetteurs : hommes politiques, professeurs, écrivains, journalistes, lecteurs de magazines, etc.

- Une typologie des récepteurs : grand public, catégories socioprofessionnelles, classes sociales, classes d'âge, public spécialisé, etc.

- Une typologie des situations, indissociable d'une typologie des medias et des supports de la communication écrite. A chacun de ces types textuels correspond un médium précis et une situation spécifique de réception (i.e. de lecture) : discours lu, dépêche d'agence de presse, publicité radiophonique, affichage publicitaire, article de presse, bande dessinée, télégramme, page web, rapport administratif, manuel d'utilisation, note de travail, éditorial …

- Une typologie des variations diastratiques autorisées : le discours publicitaire en France, par exemple, sans prétendre affecter un registre surveillé, s'autorise rarement le recours à des expressions connotées triviales, argotiques ou vulgaires.

- Une typologie des mixtes fonctionnels : nombreux sont les types de texte qui assument explicitement leur profil fonctionnel. Par exemple, la plupart des publications de presse situent précisément leurs productions, tant au niveau du contenu que de la fonction revendiquée : « magazine de divertissement », « magazine d'information », « quotidien régional » « journal satyrique », « magazine de bricolage », etc. Les genres littéraires, tout au moins dans leur forme canonique, sont avant tout définis par un certain mixte fonctionnel plus ou moins compliqué : un roman d'aventure sert à s'évader, un roman à l'eau de rose sert à émouvoir et à faire pleurer, un roman policier sert à faire réfléchir sur une énigme, à tarauder la curiosité du lecteur, à le plonger dans l'angoisse, etc. Dans le seul champ des textes techniques, qu'on pourrait croire limités à la seule fonction informative, on relève également des fonctions variées: « On pourrait ainsi classer les textes techniques en catégories d'après leur nature : par exemple,

un rapport d'étude sert à informer; un article de vulgarisation sert à expliquer; une notice d'utilisation sert à faire marcher; un encart publicitaire sert à faire vendre, etc. » (Durieux, in Lederer & Israël, 1991 : 170).

– Une typologie des structures formelles : par exemple, un article de journal est le plus souvent constitué d'un titre, un chapeau et le corps du texte ; les horoscopes sont toujours classés par signes, dans le même ordre ; une interview doit prendre la forme d'une alternance entre questions assez courtes et réponses ; un article scientifique comporte presque toujours les sections suivantes : titre, auteur, affiliation, résumé, mots clés, introduction, parties, conclusion, remerciements, bibliographie, annexe.

Tous les types textuels n'imposent pas les mêmes contraintes. Certains sont très ouverts, et ne définissent pas de critères rigoureux le long des axes énumérés : c'est souvent le cas des situations de communication interindividuelle (lettre, courriel, billets, etc.). A l'opposé, il existe des types éminemment contraints, fréquents dans la sphère publique (p. ex. les vœux du Président de la république, le discours de réception à l'Académie française, etc.) où toutes les instances sont fixées à l'avance : l'émetteur, les récepteurs, les variations diastratiques, le mixte fonctionnel, etc.

Ces conventions textuelles demandent donc une grande maîtrise de la part du traducteur, qui doit pouvoir repérer, dans la culture source, l'allégeance à un type, ou au contraire le décalage voire la rupture avec les schémas conventionnels. Une maîtrise identique est requise dans la culture d'arrivée, où les choix typologiques vont orienter le travail de production du texte cible. Par exemple la traduction d'une publicité du français vers l'italien nécessite quelques adaptations : alors qu'en France le vouvoiement est de rigueur (en dehors des « cibles » jeunes), le tutoiement publicitaire est fréquent sur la péninsule.

Notons qu'il n'existe pas toujours de stricte correspondance entre types textuels, d'une culture à l'autre. Dans le domaine de la traduction littéraire, le cas est fréquent, dans la mesure où chaque langue possède des traditions et des canons littéraires qui lui sont

propres. Israël (in Lederer & Israël, 1991 : 25) donne l'exemple de la traduction de Racine en anglais :

> « (...) le traducteur anglo-saxon qui aborde l'œuvre de Racine ne dispose pas, au sein de sa propre culture, d'un véritable modèle de référence. C'est pourquoi le poète américain Robert Lowell a cherché, dans sa version de *Phèdre*, à se rapprocher de Dryden et de Pope, atténué le tour abstrait du langage, supprimé la rime et remplacé l'alexandrin par le décasyllabe, le tout afin de donner au texte racinien une possibilité de fonctionnement dans le système littéraire anglais. »

I.1.2.5 Traduction et langues de spécialité

Dans les domaines scientifiques et techniques, l'exercice de la traduction requiert la maîtrise de ce qu'on nomme des *langues de spécialité* (qu'on appelle aussi parfois *langues spécialisées* ou *sous-langages*). Cette compétence spécifique soulève le problème du statut de ces langues de spécialité : s'agit-il de langues à part entière nécessitant l'apprentissage d'un système spécifique, comme lorsqu'on aborde un dialecte différent ?

Comme le suggère Sager (1994 : 47), la définition des langues de spécialité doit reposer sur un soubassement pragmatique : « La limite entre langue générale et langue de spécialité ne peut (…) être déterminée que par des critères pragmatiques découlant de l'usage. »[49]

Ainsi, Sager (1994 : 28) fait reposer la notion de « *sous-langages* », sur le socle d'une *communauté linguistique* spécifique: « L'identification et la description des sous-langages ne sont rien d'autre que l'abstraction des éléments communs aux actes de paroles effectués par des locuteurs appartenant à un groupe social cohérent, défini par des critères professionnels, situationnels, géographiques, historiques ou autres. »[50] Cependant, en assimilant les langues de spécialité à des *sociolectes* particuliers, il est préférable de les définir sur la base de ce que Rastier (1989 : 49) appelle des *pratiques* : « Un sociolecte relève plutôt d'une pratique sociale que d'un groupe social déterminé : nous possédons

[49] "The threshold between general and special subject language can, however, be delineated only by pragmatic criteria derived from usage."
[50] "The identification and description of sublanguages is in effect nothing more than the abstraction of common elements of real speech acts by speakers belonging to a coherent social group defined by professional, situational, geographical, historical or other criteria."

tous plusieurs compétences sociolectales liées à ces pratiques (sport, politique, enseignement, etc.). »

Cette approche pragmatique soulève un certain nombre de problèmes :

- Comment définir rigoureusement les limites d'une pratique sociale déterminée, dont dépend l'« unité » et la consistance de ce qu'on entend par langue de spécialité ? Par exemple, si l'on s'intéresse au domaine de la recherche en linguistique, il y a-t-il un sens à chercher un commun dénominateur entre un manuel, une communication orale, un échange d'information sur une liste de diffusion, un rapport de soutenance, une entrée de dictionnaire, un manuel d'utilisation d'étiqueteur morphologique, un rapport d'activité ?

- La caractérisation pragmatique d'une langue de spécialité est-elle pertinente d'un point de vue linguistique ? Autrement dit, est-il possible d'extraire des traits représentatifs, caractérisant les propriétés linguistiques des discours spécialisés ?

La première question pose le problème de la gradation du phénomène des langues de spécialité. Dans la mesure où on a affaire à un *continuum*, à l'intérieur duquel tous les degrés sont envisageables, les démarcations peuvent apparaître artificielles voire arbitraires. En outre la classification des langues spécialisées recoupe la stratification des typologies textuelles : il est peut-être plus économique et plus pertinent de rétrécir le champ à des types de textes précis au sein d'une pratique sociale déterminée.

Les travaux de Douglas Biber (1988) ont montré qu'il était possible de caractériser les types de textes de manière inductive sur la base de traits linguistiques tels que le passif, les formes interrogatives, les pronoms, les adverbes de temps et de lieu, les déictiques, les

marqueurs de temps et d'aspect, les modalités, la coordination, etc. [51] Dès lors, on peut se demander ce qu'il reste des traits caractérisant la spécialité, si l'on fait abstraction de toutes les propriétés relatives à la typologie textuelle. Habert *et al*. (1997 : 152) posent la question à propos d'un corpus de manuels médicaux montrant « les régularités propres à tout discours didactique (pluriels génériques, présent de vérité générale, etc.) qui 'parasitent' la perception du sous-langage proprement dit. »

Pour certains auteurs, comme Zelig Harris *et al*. (1988), les langues des disciplines scientifiques ou techniques peuvent être décrites en terme de « sous-langage » impliquant des restrictions à la fois sémantiques et syntaxiques : ces sous-langages seraient caractérisés par le recours à un vocabulaire limité et par la récurrence de schémas prédicats - arguments spécifiques. Mais comme le remarquent Habert *et al*. (1997 : 149), ces restrictions n'équivalent pas à une inclusion : « La dénomination *sous-langage* tient du faux ami. Ces sous-langages ne sont pas forcément en effet des sous-ensembles de la langue générale. Certains traits de la langue générale s'y retrouvent, d'autres leur sont propres ». Pourtant, même si, comme le notent ces auteurs, certaines structures peuvent être décrites en terme de « grammaires locales », la syntaxe reste gouvernée par les règles du système fonctionnel de la langue. Pierre Lerat (1995 : 75) remarque qu'« il n'existe pas de règles propres à la syntaxe de quelque " langue de spécialité " que ce soit, à proprement parler. »

Il nous paraît plus éclairant de situer, comme Sager (1994 : 25), les langues de spécialité le long d'un axe bipolaire opposant « langage naturel » et « langage artificiel ». Dans cette opposition, l'« artifice » constitue à édicter une norme, à énoncer un certain nombre de règles destinées à contrôler l'usage, afin de maîtriser le fonctionnement du

[51] 67 traits au total, répartis en 16 catégories. Une analyse factorielle permet de regrouper ces traits en 5 dimensions cohérentes, et d'en déduire, par des méthodes de classification automatique, une typologie induite comportant huit catégories : *intimate interpersonal interaction*, *informational interaction*, « *scientific* » *exposition*, *learned exposition*, *imaginative fiction*, *general narrative exposition*, *situated reportage*, *involved persuasion*. Benoît Habert *et al.* (1997 : 29) décrivent cette méthode de la manière suivante : « La statistique multidimensionnelle est mise à contribution pour repérer les oppositions majeures entre associations de traits linguistiques. Elle rassemble les traits qui ont tendance à apparaître ensemble. Elle constitue dans le même temps les configurations de traits qui sont systématiquement évités par les mêmes rassemblements. Cette démarche permet d'obtenir des pôles multiples, positifs et négatifs, correspondant à ces constellations. Ces pôles deux à deux constituent des dimensions. Chaque texte, par son emploi des traits linguistiques étudiés, se situe en un point déterminé de l'espace à *n* dimensions déterminé par cette analyse. »

langage en tant qu'*outil* développé par et pour la discipline spécialisée. Sager donne les exemples de la *Nomina Anatomica* ou de la nomenclature de la chimie, qui sont des sous-langages régissant les constructions lexicales avec des règles spécifiques de composition et de dérivation. Les langages d'interrogation de bases documentaires sont des langages artificiels, comportant un vocabulaire fixé et structuré hiérarchiquement au sein d'un thesaurus, et des connecteurs permettant de combiner les unités pour former des requêtes. Ce qui nous intéresse, dans cette notion de langage artificiel, c'est la définition d'un *continuum* couvrant un très large éventail de phénomènes :

- Du côté « naturel », on trouve bien sûr toutes les langues humaines, qui incluent cependant une part d'artifice dès le moment où l'on énonce certaines normes susceptibles de fixer des limites *a priori* à l'usage. Les langues écrites pour lesquelles on explicite de manière détaillée les règles de la grammaire et du « bon usage » font un pas de plus vers le contrôle artificiel de l'usager sur le code.

- A l'opposé, on trouve par exemple les langages informatiques totalement explicites et figés dans leur syntaxe comme dans leur interprétation.

- Entre ces deux pôles se situent de nombreux usages intermédiaires. Par exemple, certaines conditions de communication imposent redondance et absence d'équivocité afin d'assurer les meilleures chances de transmission : les communiqués de météo marine, ou de navigation aérienne, répondent à ces exigences, en limitant les constructions syntaxiques et en multipliant les formules stéréotypées à partir d'un lexique fermé. Le traitement informatique, de plus en plus indispensable dans la gestion documentaire d'importants volumes de données, oblige parfois à imposer aux rédacteurs des restrictions artificielles au niveau syntaxique. C'est ce type de rédaction *contrainte* ou *contrôlée* qui a été mise en œuvre dans la perspective du projet Kant, système de traduction automatique développé à l'Université Carnegie-Mellon de Pittsburg, dédié à la traduction en 35 langues des 100 000 pages de documentations techniques que la société Caterpillar produit chaque année.

Or les usages des langues de spécialité se répartissent tout le long de ce continuum. Les nomenclatures comme la *Nomina Anatomica*, les terminologies, les abréviations employées dans les dictionnaires, les systèmes de notations formels de type mathématique, etc., constituent des sous-systèmes plus ou moins fermés, explicités, contrôlés, visant à structurer artificiellement certains usages linguistiques. Mais les langues de spécialité, si on les caractérise sur la base des pratiques spécialisées, ne peuvent se réduire à ces sous-systèmes artificiels. De par leur fermeture, ceux-ci présentent des limites intrinsèques. Sager (1994 : 45) remarque qu'« un langage artificiel ne peut servir de métalangage pour discourir de lui-même ou de tout autre langage, cette fonction ne pouvant être assumée que par un langage moins contraint »[52]. A ce titre, Sager (1994 : 48) établit une distinction fondamentale entre deux types d'actes de langages, *innovatifs* et *rétrospectifs* :

> « Les actes de langage rétrospectifs servent à confirmer et consolider un savoir, et donc à accumuler un réservoir de connaissances (comme dans une base de données) qui pourrait, sous une forme épurée, se réduire à la définition et à la description formelle des concepts mis en jeu et de leurs relations au sein d'une représentation des connaissances.
> Les actes de langage innovatifs permettent de développer de nouvelles théories, d'exprimer des jugements spéculatifs et des hypothèses, ce qui nécessite le recours à toute la gamme des phénomènes et des potentialités linguistiques. »[53]

Du fait des usages innovatifs, les langues naturelles s'opposent aux langages artificiels par une propriété fondamentale : elles sont *ouvertes*. Elles s'autorisent, à tout moment, le recours à des unités et à des structures qui ne sont pas encore inscrites dans leurs codes, sauf en germe. Il est évident que les usages scientifiques et techniques font un usage intensif de cette possibilité de création, au moins à chaque fois qu'ils ont affaire à des contenus nouveaux, car la fonction métalinguistique est prééminente dans le discours scientifique. Même si cette ouverture concerne essentiellement le vocabulaire, elle exige une certaine souplesse syntaxique, comme la littérature scientifique en témoigne. Les

---

[52] "An artificial language cannot function as the metalanguage in which to discourse about this language or any other; for this function, a less restricted language is required."

[53] "Retrospective speech acts serve the purpose of confirmation and consolidation of knowledge, and thus constitute a reservoir of knowledge (as in a knowledge base) and could, in their purest form consist of formal descriptions or definitions of the concepts and relations between them in a knowledge structure.
Innovative speech acts are then those which develop new theories, express speculative judgements and hypotheses, many of which require the full range of natural language phenomena to deal with the new ideas as they are being developed."

besoins communicatifs imposent des règles mais n'en dépendent pas, car « les langues de spécialité n'adoptent de règles spéciales que dans la mesure où cela n'inhibe pas la communication. »[54] (Sager, 1994 : 47).

C'est pourquoi il faut aborder les langues de spécialité, comme Lerat (1995 : 28), en terme de « pluri-systèmes ». D'après ce dernier, les usages spécialisés partagent en général les caractéristiques suivantes :

- recours à des systèmes de signes non-linguistiques ;
- priorité donnée au palier de l'écrit ;
- morphologie spécifique incluant du lexical général et du lexical spécifique (cf. en chimie, le couple de suffixes *-ique* / *-ate*) ;
- syntaxe générale, avec des prédilections stylistiques et des phraséologies professionnelles ;
- sémantique non ethnocentrique, universalité potentielle des notions scientifiques et techniques.

Bien entendu, la traduction en domaine spécialisé requiert une connaissance de ces pluri-systèmes et une maîtrise de différents niveaux de contraintes. Nous retiendrons essentiellement trois de ces niveaux :

- *Niveau phraséologique*

Certaines pratiques sociales adoptent des normes phraséologiques à la fois originales et stéréotypées, comme dans la langue juridique, la liturgie religieuse, le discours journalistique, le discours politique, etc.

[54] "Special subject languages adopt special rules only to the extent that do no inhibit communication."

Par exemple, la phraséologie des textes législatifs est centrée sur la nécessité d'énumérer de façon explicite des circonstances, des motifs, puis des décisions, des recommandations, des directives, etc. Ainsi, une même « phrase » étant susceptible de s'étaler sur un texte tout entier, l'articulation logique de ses alinéas doit être transparente, ce qui impose une manière spéciale d'« aligner » les expressions adverbiales et les verbes, comme dans ce fragment de résolution européenne[55] :

> *PROPOSITION DE RÉSOLUTION*
> *sur les droits des personnes handicapées*
> *le Parlement européen ,*
> *- vu les pétitions n o 63 / 93* (…) ;
> *- vu la communication de la commission du 30 juillet 1996*
> [*suivent 8 tirets « vu… »*]
> *considérant que, selon des estimations officielles, 37 millions de personnes handicapées en tant que citoyens de l'Union européenne ne bénéficient pas de leurs pleins droits civils et humains* ; [*suivent 2 « considérant… »*]
> *invite les États membres à inclure une clause de non-discrimination* (…) ; *invite* (…); *demande* (…) ; *estime* (…) ; *estime* (…) ; *insiste* (…); *demande aux institutions communautaires et aux États membres de revoir leurs politiques d'accès et d'emploi et à la commission de publier un Code de bonne pratique en matière d'emploi des personnes handicapées et de faire en sorte que la directive relative à la passation des marchés publics de fournitures, lorsqu'elle sera révisée, garantisse une prise en compte appropriée des critères sociaux dans les contrats publics* ; *se félicite de la communication de la commission* (…); *etc.*

Cet exemple montre que certains critères stylistiques, comme la simplicité des phrases et l'évitement des répétitions, n'ont pas de pertinence dans ce domaine-ci. Notons que la traduction de ces formules nécessite d'en connaître les équivalents techniques :

> fr. : *vu…*
> angl. : *having regard to …*
>
> fr. : *considérant que…*
> angl. : *whereas …*

Lorsqu'on observe la traduction anglaise de la précédente résolution, on relève néanmoins des variations d'ordre stylistique : le français cherche malgré tout, dans la mesure du possible, à éviter certaines répétitions, tandis que la version anglaise paraît plus

[55] source Internet : http:\\www.europarl.eu.int : rapport A4-0391.

« systématique » sur ce plan. Par exemple, en ce qui concerne les recommandations, la version française présente de nombreuses variations :

fr. : *invite*…
angl. : *calls* …

fr. : *demande*…
angl. : *calls* …

fr. : *estime*…
angl. : *calls* …

fr. : *insiste*…
angl. : *calls* …

fr. : *estime certes*…
angl. : *acknowledges that* …

Ces phraséologies se manifestent donc sous la forme de régularités, observables sur un plan statistique. Lerat (1995 : 75) remarque que l'analyse syntaxique des langues spécialisées doit porter sur des habitudes langagières, et non sur des règles strictes : « Ce qui est attendu, en revanche, d'une analyse syntaxique des langues spécialisées, c'est l'aptitude à rendre compte linguistiquement d'habitudes d'expression statistiquement dominantes dans tel type de texte, c'est-à-dire d'un style. »

– *Niveau du domaine*

Toute pratique se caractérise par un discours, situé dans ce que Rastier (1989 : 39) nomme un « *domaine sémantique* » :

> « A chaque type de pratique sociale est associé un type d'usage linguistique que l'on peut appeler discours : ainsi des discours juridiques, politiques, médicaux, etc. Les discours ainsi entendus correspondent à ces formations paradigmatiques que sont les domaines sémantiques. Au sein d'un domaine sémantique, il n'existe pas, en règle générale, de polysémie. »

Si la définition d'un domaine sémantique permet d'interpréter correctement un terme, la réciproque n'est pas vraie. Les textes spécialisés chevauchent souvent plusieurs domaines et les disciplines empruntent souvent leurs termes à la langue générale ou à des

disciplines voisines, de sorte qu'il est rarement possible d'affecter un domaine *a priori* à un terme : le terme *loi*, qui appartient à la langue générale, s'interprète différemment en droit, en philosophie, en mathématique, en statistique, en physique, etc. Ainsi, de très nombreux usages métaphoriques transcendent les domaines (p. ex. les métaphores physiques courantes en linguistique : *molécule sémique*, *atome lexical*, *satellite du verbe*, *structure nucléaire*, etc.). Cette perméabilité des frontières entre spécialités explique selon Sager (1994 : 29-30) les difficultés de classification rencontrées en lexicographie :

> « Au seul niveau sémantique de la langue, l'établissement de frontières n'est pas évident dans la mesure où tous les types de sous-langages partagent des unités lexicales, ce qui rend difficile le rattachement de ces unités à des domaines spécialisés. L'hésitation des dictionnaires à utiliser des étiquettes de domaine est particulièrement révélatrice de cette question. » [56]

En outre, les langues de spécialité emploient de très nombreuses unités polysémiques dans un sens général, comme les verbes précédemment cités : *inviter*, *estimer*, *demander*, *insister*, etc. Comme le montre R. Krovetz (1998), les unités employées dans des acceptions différentes au sein d'un même discours spécialisé sont très fréquentes. La citation de Rastier ne doit pas laisser entendre que la polysémie est absente des langues de spécialité, même si elle l'est, dans une certaine mesure, au niveau des sous-systèmes spécialisés que sont les terminologies.

Enfin, notons que l'interprétation du contenu requiert parfois une connaissance approfondie du domaine. Il arrive qu'une part d'implicite demande à être explicitée au cours du processus de traduction, soit que la langue d'arrivée exige certaines précisions sémantiques, soit que les récepteurs de la traduction ne possèdent pas exactement les mêmes référentiels que ceux du texte original. On peut reprendre l'exemple donné par Boitet (in Bouillon & Clas, 1993 : 112): « A partir de *la* grammaire linéaire droite G1, on construit *le* système d'équation associé, et on en déduit *une* expression régulière pour L(G1). » Avec une autre traduction des articles *la, le une*, la phrase aurait perdu son sens, car dans ce domaine précis on sait que pour G1 il n'existe *qu'un* système d'équation associé et *plusieurs* expressions régulières possibles. Pour faire le bon choix, le traducteur

[56] "At the semantic level of language alone, the borderline for a subdivision is prima facie less obvious because all types of sublanguages share common lexical items and find it difficult to attribute lexical unequivocally to special subject fields. We note this particularly in the hesitation of dictionaries to use subject labels."

devait connaître au moins les rudiments de la théorie. Toutefois, l'habitude et l'expérience dans un domaine permettent de pallier en partie l'absence de compréhension. Comme le suggère Boitet, un traducteur humain est toujours capable de mettre en œuvre une compréhension superficielle qui s'appuie à la fois sur des indices de surface et un certain bon sens :

> « En général, un traducteur humain comprend facilement le contexte pragmatique et communicatif d'un texte. Par contre, sa connaissance du domaine spécialisé concerné est souvent très superficielle, voire nulle. Pourtant, s'il est expérimenté, il arrive dans une certaine mesure à faire illusion, c'est-à-dire à traduire comme s'il avait compris, en se fondant uniquement sur son habitude du type de texte. Il s'agit donc de compréhension apparente (humaine). » (ibid.)

– *Niveau terminologique*

Fonctionnellement, toute terminologie s'applique, dans un domaine d'activité précis, à la désignation de concepts (contenus extralinguistiques) par des lexies (les termes), de manière univoque et non ambiguë. En d'autres termes, il s'agit de l'aménagement artificiel d'un vocabulaire, visant à clarifier les rapports de désignation, et parfois à maîtriser la création lexicale en assurant une cohérence dans les mécanismes de dérivation et de composition. En théorie, la relation biunivoque entre le terme et le concept se traduit par deux contraintes : le même terme ne doit pas désigner deux concepts différents (dans le même domaine), et le même concept ne doit pas être désigné par deux termes différents. Deux aspects attirent notre attention :

- il existe des « pré-terminologies », non encore normalisées ni explicitées, qui sont employées et reçues à des degrés divers au sein des communautés concernées. Le caractère provisoire et inachevé de ces terminologies, les conflits qu'elles suscitent lorsqu'elles recouvrent des controverses entre spécialistes, font que la relation biunivoque citée plus haut n'est plus respectée. Les controverses peuvent porter sur les termes, mais aussi sur les concepts, lorsque ceux-ci ne sont pas acceptés de manière unanime. Dans un travail de traduction, le traducteur ne

disposera pas toujours de terme équivalent dans la langue d'arrivée ; parfois, il en aura plusieurs[57].

– les terminologies sont toujours liées à un domaine précis. Elles n'ont de pertinence qu'indexées à ce domaine de définition. Des domaines différents peuvent assigner à de même lexies des concepts différents, et inversement. Pour reprendre un exemple cité par Daniel Gouadec (1993 : 41), les brasseurs au Royaume-Uni nomment *liquor* l'eau du robinet qui entre dans la composition du brassin, et non simplement *water*.

De même que la traduction de contenus conceptuels exige parfois une collaboration étroite entre traducteur et spécialiste, la traduction des termes demande des informations spécifiques sur les terminologies, incluant les équivalences entre les langues. Lorsque celles-ci sont explicitées, le travail de traduction se borne à un simple transfert lexical, une fois les termes identifiés.

En revanche, lorsque le traducteur a affaire à des concepts pour lesquels il n'existe aucun terme connu dans la langue d'arrivée, son rôle se complique, car il devient potentiellement créateur de nouveaux termes. Ce rôle créateur n'est pas à négliger : Pierre Coste, le premier traducteur de l'*Essay on Human Understanding* de Locke (1690), a introduit le terme de *conscience* pour traduire le néologisme anglais *consciousness* (dans le sens cognitif et non moral). L'usage est resté en français, et le terme philosophique de *conscience* a conservé les acceptions que lui a prêtées pour la première fois le traducteur de Locke (Leibniz, dans ses *Nouveaux Essais*, essaiera d'introduire le terme « *conscienciosité* », mais sans succès...).

Le développement des nouvelles technologies de l'information et de la communication a donné naissance à un gigantesque chantier terminologique en français, la plupart des termes étant d'origine anglo-saxonne. Parfois, certains paramètres intentionnels peuvent entrer en ligne de compte et influencer la traduction, certains termes étant liés à des marques commerciales (p. ex. faut-il traduire *browser* par *navigateur* ou *explorateur*, ou un autre terme ?). Comme le note Christine Durieux (in Lederer & Israël, 1991 : 173) à

[57] La linguistique fournit de beaux spécimens de fluctuations terminologiques : l'anglais *meaning* peut se traduire par *sens*, *contenu*, *signification*, *signifié*, *sémantisme*, etc..

propos de la traduction technique, « traduire de la documentation technique est aussi un acte de marketing. »

I.1.2.6 Les problèmes culturels

Avec les contraintes imposées par les conventions de typologie textuelle ou la définition d'un domaine, on passe graduellement de la *situation* de communication au *contexte* élargi, englobant l'ensemble des déterminations culturelles partagées par les membres de la société. Lorsqu'un message est traduit dans une langue étrangère, et donc dans une culture étrangère, le traducteur réalise alors un processus d'*acculturation* : le message d'origine, déraciné de son contexte d'émission, doit devenir *assimilable* au sein d'une culture différente. Lorsque deux cultures sont très éloignées, dans le temps ou dans l'espace, on peut légitimement se demander si cette acculturation n'aboutit pas nécessairement à une perte irréversible de sens.

Une partie du problème découle du fait que langue et culture sont difficilement détachables l'une de l'autre. D'après Mounin (1963 : 234) « le contenu de la sémantique d'une langue, c'est l'ethnographie de la communauté qui parle cette langue ». De manière plus radicale, certains auteurs considèrent la langue comme le réceptacle d'une *vision du monde* particulière. C'est ce que sous-entend l'hypothèse de Sapir-Whorf (cité et traduit par Mounin, 1963 : 46) : « tous les observateurs ne sont pas conduits à tirer, d'une même évidence physique, la même image de l'univers, à moins que l'arrière-plan linguistique de leur pensée ne soit similaire, ou ne puisse être rendu similaire d'une manière ou d'une autre ».

Ces différences de *visions du monde* vont jusqu'à affecter des catégories considérées comme universelles. Pour reprendre un exemple donné par Mounin (1963 : 46) c'est le cas « des noms de nombre, toujours en hopi, qui contraignent à distinguer grammaticalement l'addition de quantités dans l'espace (dix hommes) de l'addition de quantités dans le temps (dix jours) ». Au niveau des conceptions spatio-temporelles, Nida (1959, trad. in Nergaard, 1995 : 151) [58] cite l'exemple des Quechuas de Bolivie, pour qui le futur est derrière, et le

[58] Texte original : Nida, E. (1959) « Principles of Translation exemplified by Bible Translating » in R. Brower (ed.), On Translation, Cambridge, Harvard University Press, pp. 11-31

passé devant : cette représentation serait motivée par le fait qu'on peut voir en esprit ce qu'on a déjà vécu, mais pas ce qu'on vivra. Cette vision n'est certes pas plus illogique que la nôtre. Tout simplement, elle procède d'une *autre* logique.

Ainsi, la langue a une telle prégnance sur la pensée qu'elle va jusqu'à conditionner l'organisation des abstractions conceptuelles. Dans un article intitulé « Catégories de pensée et catégories de langue », E. Benveniste (1966 : 69-76) montre que les catégories logiques énoncées par Aristote ne sont rien d'autre que la transposition des catégories de langue propres au grec ancien :

> « on discerne que les " catégories mentales " et les " lois de la pensée " ne font dans une large mesure que refléter l'organisation et la distribution des catégories linguistiques. Nous pensons un univers que notre langue a d'abord modelé. » (1966 : 6)

L'irréductibilité des cultures les unes aux autres ne doit cependant pas nous conduire à une vision autistique des langues : une hétérogénéité absolue de visions du monde, de repères culturels, voire de modes de pensée, interdirait la traduction et condamnerait même toute tentative de communication.

Si la langue est bel et bien la manifestation la plus évidente des spécificités culturelles, et d'une manière originale de découper le monde[59], elle n'impose pas un cadre préformé à la structuration de nos pensées. La langue reste un instrument qu'on peut plier à ses exigences. Nous pensons avec Seleskovitch (citée par Laplace, 1994 : 195) qu'il n'y a pas identité entre langue et conceptualisation : « Les idées doivent se couler dans les

[59] Lévi-Strauss (1962 : 117) décrit ainsi le travail de structuration culturelle, dont on ne peut jamais présumer : « D'autre part, et même promus à ce niveau humain qui peut seul leur conférer l'intelligibilité, les rapports de l'homme avec le milieu naturel jouent le rôle d'objets de pensée : l'homme ne les perçoit pas passivement, il les triture après les avoir réduits en concepts, pour en dégager un système qui n'est jamais prédéterminé : à supposer que la situation soit la même, elle se prête toujours à plusieurs systématisations possibles. ». Par exemple, dans l'observation d'une même espèce d'oiseau, des cultures différentes adoptent des points de vues très variés : « Si le grimpereau éveille l'intérêt des Australiens c'est, comme l'a montré Radcliffe-Brown, parce qu'il hante le creux des arbres ; mais les Indiens des prairies de l'Amérique du Nord prêtent attention à un tout autre détail : le pic à tête rouge est censé être protégé des oiseaux de proie parce qu'on ne trouve jamais ses vestiges. Un peu plus au sud, les Pawnee du haut Missouri établissent une relation (comme les anciens Romains semble-t-il) entre le pic et la tempête et l'orage, tandis que les Osage associent cet oiseau au soleil et aux étoiles. Mais pour les Iban de Bornéo (…) une variété de pic (…) reçoit un rôle symbolique en raison de son chant triomphal. » (Lévi-Strauss, 1962 : 74)

catégories que leur impose la langue, mais elles ne se confondent pas plus avec ces catégories qu'elles ne se confondent avec la langue ».

S'il y a « intraduisibilité », cela ne concerne que des unités linguistiques prises isolément. Comme le remarque Seleskovitch (citée par Laplace, 1994 : 234), il ne faut pas confondre cette « intraduisibilité » avec l'impossibilité de traduire : « […] nous ferons la distinction entre deux notions souvent confondues lorsqu'on parle de mots " intraduisibles " : celle de traduire dans le sens de passer d'une langue à l'autre en mettant un mot à la place de l'autre, et celle d'exprimer la même chose dans les deux langues. » Quand on traduit, la *situation* est la seule unité de mesure valable entre les deux énoncés. « Si l'on peut prouver que deux langues différentes analysent l'expérience non-linguistique de manière différente, ce n'est pas en se fiant à l'analyse linguistique, puisque des structures totalement différentes peuvent signifier arbitrairement des situations tout à fait semblables. » (Mounin, 1963 : 268). Ce primat de la situation apparaît de façon évidente au niveau d'expressions jouant le rôle de marqueurs pragmatiques, comme ces locutions énumérées par Seleskovitch (citée par Laplace, 1994 : 236):

> « Si, lorsqu'on frappe à sa porte le Serbe dit " slobodno ", le Français " entrez ", l'Allemand " herein " et l'Anglais " come in ", c'est à partir de la situation où chacun se trouve et de ce qu'il veut exprimer qu'il trouve l'expression juste et non à partir de l'expression équivalente dans une autre langue. Traduite, celle-ci donnerait " libre " pour le serbe, " dedans " pour l'allemand et seul l'anglais " come in " traduirait le français " entrez " »

De façon générale, chaque langue induit un point de vue particulier sur des situations communes, ce qui ne pose pas problème pour la traduction. On peut citer un exemple donné par Mounin (1963 : 268) : « quand un japonais dit : *c'est un puits profond*, son analyse sémantique se réfère à l'importance d'un volume creux et vide ; quand un mongol dit : *c'est un puits profond*, c'est la partie creuse, remplie d'eau qu'il nomme » : ces divergences peuvent causer des contresens, mais elles n'empêchent en rien la traduction. Si une formulation est ambiguë dans la langue d'arrivée, il suffit pour le traducteur d'expliciter la situation de manière plus précise, en fonction des données contextuelles. La traduction ne devient problématique qu'en l'absence de tout contexte.

Quand on soulève les problèmes de traduction, on évoque souvent des difficultés de nature lexicale (Mounin 1963 : 61-63), avec des questions du genre : « Comment traduire

montagne pour les Indiens de la péninsule absolument plate du Yucatán dont l'éminence la plus haute atteint 30 m. » ? Comment nommer une montagne dans une langue de la plaine ne possédant pas le mot *montagne* ? Comment exprimer *palmier* dans la langue des esquimaux ?

D'une certaine manière, il s'agit là d'un faux problème, nourri par une tendance à hypostasier les catégorisations du lexique pour les transformer en données objectives. Le problème des couleurs en est une bonne illustration : on entend parfois des assertions du type « il y a sept noms de couleurs en français », « dans telle langue africaine on ne distingue que deux couleurs », etc. Certes, on remarque que chaque langue effectue un découpage original[60] du champ chromatique. Mais une langue ne lexicalisant que deux couleurs est-elle capable d'exprimer des nuances plus précises ? A l'évidence oui. Mounin (1965 : 214) remarque que les couleurs sont souvent définies, « par référence à des technologies de teintures, de peinture, de marquage ou de coloriage, par référence au matériau d'origine, au produit colorant, au procédé, à la nuance définie par comparaison avec un objet de couleur standard. » Grâce à la métonymie on peut pallier les défaillances lexicales les plus drastiques. Il suffit de caractériser une couleur par un objet qui la porte : « ciel », « azur », « mer », « neige », « cerise », « orange », « marron », « canari », « bordeaux », « cachou », « brique » sont des usages métonymiques courants en français, et l'on peut même noter que la grande majorité des qualificatifs de couleur proviennent de tels procédés. L'absence de dénomination pour désigner un objet ou un concept n'est jamais insurmontable, car outre l'emprunt ou le calque, les tropes autorisent une grande créativité. Jakobson (1963 : 82) cite l'exemple des Chukchee, du nord-est de la Sibérie, qui ont su se doter d'expressions imagées assez explicites dès qu'ils ont eu besoin de dénommer des objets nouveaux :

écrou = « clou tournant »
acier = « fer dur »
craie = « savon à écrire »
montre = « cœur martelant »

[60] On peut citer Lévi-Strauss (1962 : 73) : « [le système hanunóo] distingue les couleurs, d'une part en relativement claires et relativement foncées, d'autre part selon qu'elles sont habituelles aux plantes fraîches ou aux plantes séchées ; les indigènes rapprochent ainsi du vert la couleur marron et luisante d'une section de bambou qui vient d'être coupé, alors que nous-mêmes la rapprocherions du rouge si nous devions la classer dans les termes de l'opposition simple entre les couleurs rouge et verte qu'on rencontre en hanunóo. »

Le problème se déplace lorsqu'on élargit le champ du linguistique au culturel : les langues peuvent sans doute exprimer les mêmes expériences par des procédés différents, mais comment rapprocher des expériences irréductibles ? Clare Donovan (1990 :110) évoque cette difficulté :

> « La traduction, pas plus que le dialogue monolingue, ne peut surmonter le décalage d'expérience et de sensibilité qui existe entre deux interlocuteurs. Il est aisé d'imaginer à quel point les Hopis d'Amérique auraient, par exemple, du mal à comprendre le concept de " pointage " dans une usine, non pas parce que leur langue ne possède pas de mot pour ce concept ou parce qu'elle est fondée sur une analyse du temps autre que la nôtre, mais parce que la chose désignée reste étrangère à la culture hopi. La difficulté éventuelle de traduction ne résulte pas de la différence linguistique, mais de l'écart culturel entre les deux communautés linguistiques. On peut toutefois envisager une " traduction ", à condition d'admettre qu'elle devra être très explicite. »

Cet écart culturel n'est donc pas un problème spécifique à la traduction, mais une difficulté inhérente à toute forme de communication. Il n'est insoluble, en ce qui concerne la traduction, que si l'on s'impose la recherche d'une identité absolue. Mais comme nous l'avons vu précédemment, traduire, c'est construire des *équivalences*. Si l'on a besoin de traduire *pointage* dans la culture hopi, c'est qu'un échange s'est déjà engagé préalablement, et il faut se demander pour quelles raisons on veut traduire le terme de *pointage* : en fonction de ces raisons, il existera toujours une manière adaptée de décrire ce concept nouveau. Ainsi, à la question posée par Seleskovitch (citée par Laplace, 1994 : 235), « " Bread " est-il vraiment l'équivalent de " pain " ? » on peut répondre par l'affirmative dans les nombreux contextes où *Bread* joue le même rôle que *pain*, même si « " Bread " pour l'Américain c'est une matière spongieuse, coupée en tranches et enveloppée de Cellophane ; pour le Français, le pain c'est une longue baguette croustillante et dorée… ». Pour que le sens d'un texte puisse être reçu, il n'est pas nécessaire que tous les récepteurs partagent la même expérience. Comme le remarque Jakobson (1963 : 78), prenant le contre-pied de Russel, on peut connaître le sens du mot " fromage " sans avoir goûté au fromage : « Tout représentant d'une culture culinaire ignorant le fromage comprendra le mot français *fromage* s'il sait que dans cette langue ce mot signifie " aliment obtenu par la fermentation du lait caillé ", et s'il a au moins une connaissance linguistique de " fermentation " et " lait caillé " »

Fondamentalement, et Jakobson (1963 : 81) insiste sur ce point, une langue n'est pas limitée parce qu'elle *peut* exprimer : « toute expérience cognitive peut être rendue et classée dans n'importe quelle langue existante. »

Même si, comme nous l'avons montré, l'équivalence traductionnelle se situe hors langue, ce contenu extralinguistique ne se fonde pas nécessairement sur une expérience concrète partagée : il peut être reconstruit, développé, explicité, commenté. Par exemple, quand le chiffre 4 est cité dans un texte coréen, il peut être intéressant d'indiquer que c'est un chiffre porte-malheur semblable à notre 13. De même, en anglais, on pourra traduire *TGV* par *French high speed train*. Dans l'autre sens, on cherchera au contraire à « impliciter », c'est-à-dire à gommer les explications jugées inutiles. Ce rapport implicite / explicite intervient au niveau culturel, au niveau référentiel comme au simple niveau linguistique, dans la mesure où « chaque langue n'explicite qu'une partie du tout qu'elle désigne et ces explicites ne se recouvrent pas » (Seleskovitch, 1984 : 174). L'explicitation peut découler de différences grammaticales aussi bien que lexicales. Dans l'exemple déjà donné : « A partir de *la* grammaire linéaire droite G1, on construit *le* système d'équation associé, et on en déduit *une* expression régulière pour L(G1). », le traducteur du russe vers le français a été contraint d'expliciter, du fait de la distinction article défini / indéfini, en puisant dans des connaissances extralinguistiques concernant la théorie des langages.

Globalement, on constate que le passage d'une culture à l'autre requiert un effort *d'adaptation*. Certaines images, pour garder leur sens, doivent s'insérer dans les cadres de la culture d'arrivée. Florence Herbulot (in Lederer & Israël, 1991 : 106) cite le cas de la traduction suivante (extraite d'un programme japonais sur la planète terre, traduit de l'anglais) :

« On voit que la terre est entourée d'une ceinture d'air chaud en forme de beignet. »

A l'évidence, le mot *beignet*, traduisant *doughnut*, n'est pas la meilleure solution, car si pour un Américain les beignets évoquent une forme d'anneau, en France ils sont plus proches du « patatoïde » que du tore : la métaphore devient donc illisible.

L'adaptation est comparable à un changement de référentiel, au passage d'un système de coordonnées à un autre. Comme l'explique Herbulot (in Lederer & Israël, 1991 : 195), c'est le cas lorsqu'on passe du coréen au français, en ce qui concerne la

manière de compter les années : « Un enfant né en décembre 1989 aura dès le mois de janvier 1990 une durée de vie qui se sera étalée sur deux années de calendrier selon la comptabilité coréenne. On dira donc en coréen qu'il a deux ans alors qu'en français il n'a que 2 mois. »

L'adaptation impose également de considérer les systèmes de connotations, qui n'ont pas de consistance interculturelle. Ainsi, dans les textes bibliques, *âne* et *cheval* n'ont pas les mêmes connotations que la culture française : dans la bible, l'âne est perçu comme un animal noble tandis que le cheval est un animal de guerre fort déplaisant (M. Gravier, in Lederer & Israël, 1991 : 36). Nida (in Nergaard, 1995 : 178) donne quelques exemples de ce type de décalage : sur le plan connotatif, le mot *cœur* doit être traduit par *foie* en kabba-laka (Afrique Equatoriale), *abdomen* en conob (Guatemala), *gorge* dans les îles Marshall.

Pour G. Bourquin (1993 : 28), les différences culturelles nous font atteindre, dans certains domaines, les limites de la traductibilité :

> « Telle autre langue peut avoir, dans le domaine concerné, une tradition sémantico-discursive totalement différente : ainsi en va-t-il lorsqu'il faut passer du discours juridique anglo-américain au discours juridique français ou d'un pays européen. On quitte à ce moment le domaine de la traduction pour celui du commentaire de texte ou de l'exégèse. »

Nous pensons au contraire que ce type de travail ne déborde pas la tâche du traducteur : puisqu'on se situe sur le plan de l'équivalence dynamique, les adaptations requises font partie intégrante du travail de traduction, tout comme l'activité de commentaire et d'exégèse. Les introductions, préfaces, postfaces et annotations critiques sont des outils à la disposition du traducteur. Il peut aussi y avoir un aspect didactique dans le fait de traduire, quand certaines informations doivent être abondamment explicitées.

Pour illustrer les problèmes d'exégèse, le terrain des traductions bibliques est riche de cas exemplaires. Nida (in Nergaard, 1995 : 156) donne l'exemple suivant : alors qu'on trouve, dans l'évangile selon saint Luc, une expression équivalente à « règne de Dieu », l'évangile selon saint Matthieu emploie une formulation qui se traduit littéralement par « royaume des cieux ». Doit-on considérer que ces deux expressions se réfèrent à des « réalités » différentes, ou bien faut-il les identifier, et les traduire par une formule identique ? Certains exégètes ont soutenu qu'il n'y avait pas identité, et que les deux termes font référence à des « récompenses » diverses. Mais Nida nous rappelle que

l'évangile de Matthieu était destiné à des croyants venus du Judaïsme, pour lesquels le nom de Dieu était frappé d'un tabou. C'est donc ce réflexe d'évitement qui aurait conduit Matthieu à employer des substituts comme « ciel », « pouvoir », « majesté ». Tandis que pour le public païen de l'évangile de Luc, non seulement le nom de Dieu n'est frappé d'aucun interdit, mais les tournures équivalentes font défaut. Nida en conclut, par l'analyse exégétique de ces divergences culturelles, qu'on peut raisonnablement assimiler les deux expressions et les traduire par une même formule.

Le problème de l'intraduisibilité, maintes fois débattu[61], trouve peut être son origine dans une vision maximaliste de la traduction, qui exigerait une totale conservation du sens. Il faut pourtant bien admettre que cette recherche est impossible, et s'attacher à la réalité des faits : comme le souligne Nida (in Nergaard, 1995 : 153), « (…) tous les types de traduction comportent (1) une perte d'information, (2) un ajout d'information et / ou (3) une déformation de l'information »[62]. Dans le changement de culture, ces transformations ont une valeur positive, puisque ce sont elles qui permettent l'adaptation.

La question de l'intraduisible demeure un problème pour la poésie, qui reste indissociable d'une forme linguistique particulière. Pour le reste, les niveaux d'équivalence que nous avons dégagés n'imposent pas d'identité absolue et laissent, malgré les écarts culturels, de nombreuses possibilités de réalisation pour la traduction.

I.1.2.7 Stratégies de traduction

Une fois établies les variables pragmatiques et culturelles susceptibles d'orienter le parcours de la traduction, nous pouvons maintenant mettre en évidence les problèmes de *choix* auxquels sont directement confrontés les traducteurs au cours de leur pratique professionnelle. Nous avons regroupé la multiplicité des attitudes et des positions stratégiques sous le terme générique de *choix*, au sens large, ce qui nous permet de subsumer à la fois les décisions délibérées et les options inconscientes ou implicites. Pour

[61] cf. Benedetto Croce (1936) « L'intraducibilità della rievocazione » (reproduit in Nergaard, 1993 : 215-220).

[62] "Vale a dire, tutti tipi di traduzione comportano (1) perdita di informazione, (2) aggiunta di informazione e/o (3) deviazione dell'informazione."

reprendre une formule de Jean-René Ladmiral (1986 : 35) : « "condamné à être libre", le traducteur est un "décideur" ».

Jiří Levý (1967, trad. in Nergaard, 1995 : 63)[63], dans une tentative de formalisation de la traduction, compare le processus traductionnel à un *processus décisionnel* :

> « Du point de vue de la pratique du traducteur, à chaque étape de son travail (c'est-à-dire d'un point de vue pragmatique), l'acte de traduire est un PROCESSUS DECISIONNEL : on a une série de situations consécutives – de coups, comme dans un jeu – situations qui imposent au traducteur la nécessité de choisir entre un certain nombre d'alternatives (qu'on peut la plupart du temps définir de manière exacte). »[64]

Ce processus est en quelque sorte commandé par la gestion économique du travail de traduction, où l'on mesure la valeur du résultat obtenu par rapport à l'effort investi. « Le travail de traduction réel est (...) pragmatique ; le traducteur recherche le maximum d'effet pour un minimum d'effort. C'est-à-dire qu'il applique intuitivement la stratégie dite du MINIMAX. » (ibid. : 79)[65]

Cette vision heuristique s'intéresse aux notions de coût et de valeur d'une traduction, à travers les trois aspects suivants :

- l'évaluation du résultat obtenu.
- l'évaluation du coût du processus.
- le choix d'une stratégie de traduction.

Nous examinerons plus loin (p. 190) la question du coût. Dans un premier temps, nous considérerons la question des *choix* traductionnels indépendamment du coût, par

---

[63] Texte original : Jiří Levý (1967) « Translation as a Decision Process », in *To Honor Roman Jakobson: Essays on the Occasion of his Seventieth Birthday*, The Mouton, La Hague, II, vol. 3, pp. 1171-1182

[64] Dans la traduction italienne : "Dal punto di vista pratico del traduttore, in ogni momento del suo lavoro (cioè dal punto di vista pragmatico), l'attività del tradurre è un PROCESSO DECISIONALE: una serie de un certo numero di situazioni consecutive – di mosse, come in un gioco –, situazioni che impongono al traduttore la necessità di scegliere tra un certo numero di alternative (molto spesso definibile esattamente)."

[65] "il lavoro di traduzione reale, comunque, è pragmatico; il traduttore decide per questo una delle soluzioni possibili che promette il massimo dell'effetto, con il minimo dello sforzo. Vale a dire, egli decide intuitivamente per la cosiddetta STRATEGIA MINIMAX". Cette heuristique, développée en Intelligence artificielle, consiste à évaluer le coût maximal pour chaque possibilité présente, et à choisir, par suite la possibilité minimisant ce coût. Lors de l'exploration d'un arbre de choix, on peut développer cette stratégie à différents niveaux de profondeur, où l'évaluation de chaque niveau est, de manière récursive, le résultat du MINIMAX appliqué aux niveaux inférieurs.

rapport aux *besoins* communicatifs (i.e. les fonctions assumées par le message traduit), et aux différentes *interprétations* possibles d'un même texte original.

En rapport avec les critères de choix qui ont été précédemment dégagés, on peut signaler diverses « postures » ou stratégies de traduction couramment adoptées dans l'exercice de la traduction professionnelle. Deux axes, corrélés entre eux, sont généralement cités :

- traduction libre *vs* littérale
- traduction « cibliste » (en anglais « *target oriented* ») *vs* « sourcière » (en anglais « *source oriented* »)

Le premier axe concerne la distance prise par rapport à la forme du texte original. Notons que cette notion de littéralité est sujette à caution, et dépend étroitement du couple de langues mises en jeu. Dans le passage d'un système à l'autre, le parallélisme entre morphologie, syntaxe et lexique des deux langues peut être trop ténu pour autoriser une quelconque conservation des structures de l'expression originale. Car il faut bien distinguer *littéralité* et *mot à mot* :

- il y a *littéralité* lorsque l'on conserve, si possible, la structuration morphosyntaxique et lexicale de l'expression originale tout en respectant l'idiome d'arrivée (le « génie » de la langue). La littéralité n'est donc pas incompatible avec des modifications profondes dans les structures morphosyntaxiques (passivation, changement d'actance, etc.), lorsqu'elles sont commandées par les usages de l'idiome (ces points seront développés au chapitre I.1.3).

- il y a *mot à mot* lorsque la conservation des unités et de la structure de l'expression originale donne un résultat « étrange », qui peut être ou non compatible avec le système grammatical d'arrivée, mais pas avec l'idiome et ses usages. Le mot à mot est souvent le produit de ce qu'on nomme une *interférence* entre le système de la langue source et celui de la langue cible : l'emploi incorrect d'un faux ami est un exemple typique de ce type d'interférence.

Nous nous éloignons de la conception de J.C. Catford (1965), qui distingue trois degrés, respectivement mot à mot, littéral et libre, pour lesquels il donne les exemples suivant :

angl. : *It 's raining cats and dogs*

1. Mot à mot : *Il est pleuvant des chats et des chiens.*
2. Littéral : *Il pleut des chats et des chiens.*
3. Libre : *Il pleut à verse.*

On ne peut dire que les traductions 1 et 2 soient des bonnes traductions : toutes les deux sont mot à mot, la première comportant une *interférence* de plus, au niveau grammatical, tandis que la seconde ne comporte qu'une interférence au niveau lexical (par la décomposition d'une expression figée). En revanche la traduction 3 est littérale : elle est correcte, et peut être déduite des deux codes linguistiques (sans tenir compte de facteurs pragmatiques particuliers). La littéralité est donc le résultat d'une opération de *transcodage*, pour reprendre le terme de Seleskovitch (cf. infra, p.95).

Vis-à-vis de la polarité libre / littérale, Sager (1994 : 155) donne un éclairage intéressant :

> « La traduction littérale cherche à conserver le plus haut degré d'équivalence formelle au niveau des mots, des locutions, des propositions et des arguments ; elle est habituellement associée avec la définition d'unités de traduction plus petites et un concept d'équivalence plus étroit.
> La traduction libre se concentre sur la transmission du contenu sous une forme aidant le lecteur à saisir, comprendre et capturer le texte plus facilement, au mépris parfois des limites de phrases ou de paragraphes. »[66]

L'opposition *libre vs littéral* recouvre donc en partie les oppositions déjà dégagées (cf. tableau 1) : *globalité du texte vs unités locales*, et *signifiés vs forme de l'expression*. Mais, paradoxalement, ces oppositions peuvent se renverser si l'on considère que la forme

[66] « The close translation is concerned with observing the highest possible degree of formal equivalencies at the word, phrase, clause and argument level; this is usually associated with the admission of smaller units of translation and a narrower concept of equivalence.
« The free translation concentrates on conveying the content in such a form that it becomes easier for the reader to grasp, understand or capture the text, even to the extent of translators disregarding unit boundaries at the sentence or paragraph level »

de l'expression occupe parfois le premier plan, comme en poésie. Israël (in Lederer & Israël, 1991 : 21) remarque avec justesse que « l'approche littérale donne volontiers l'avantage au sens notionnel sur la forme – sonorités, allitérations, rythmes – forme qui, en raison de sa matérialité même, résiste au transfert ».

La littéralité n'est pas un gage de neutralité fonctionnelle, et ses effets dépendent du type de texte traité. Pour un texte littéraire, la littéralité relève du parti pris et marque une forme d'appropriation du texte par le traducteur : « En dépit des apparences, la littéralité est donc, elle aussi, une forme d'appropriation car elle défonctionnalise le texte en lui ôtant sa respiration, en abjurant toute recherche esthétique (...) » (ibid., in Lederer & Israël, 1991 : 22).

Risset (1985 : 20), dans la préface à sa traduction de Dante, explique comment le choix d'être littéral résulte d'une décision interprétative. Ce qui lui semble saillant, dans le système de la tierce rime, c'est le rythme, l'impression de rapidité. Elle motive ainsi son choix : « Comment faire pour traduire la rapidité ? D'abord, être littéral, le plus littéral possible, et dans tous les sens – mais ceci tout en décidant de ne jamais renoncer à être *absolument moderne*. »

Sur le plan culturel, l'opposition sourcier / cibliste formulée par Ladmiral[67] (1986) prolonge la précédente polarité. Le problème posé est le suivant : doit-on gommer, dans l'opération de traduire, toute trace de l'*étrangeté* du texte original ? En reprenant une expression de Mounin (1955), doit-on voir le texte avec des « verres colorés », ou bien employer des « verres transparents » destinés à faire oublier qu'il s'agit d'une traduction ? Est-il toujours légitime d'effectuer un processus d'acculturation pour replacer, ensuite, le texte traduit dans la culture d'arrivée seulement ? Wilhelm von Humboldt (1816, trad. in Nergaard, 1993 : 136)[68] expose un parti pris résolument sourcier, lorsqu'il conçoit la

[67] Ce dernier revendique la paternité de ces deux termes, quoiqu'il leur attribue un contenu très classique puisqu'il les définit par l'accent mis tantôt sur le « signifiant », « la lettre » (sourcier), tantôt sur le « signifié », « l'esprit » (cibliste). De ce fait, Ladmiral prend résolument parti pour la position cibliste. Sans entrer dans ce débat, nous préférons quant à nous l'acception plus originale employée par Eco, où l'opposition *source / target oriented* est entendue relativement à la distance du point de vue – ou point de mire – culturel, plus ou moins exotique par rapport à la culture d'arrivée.

[68] Texte original : W. von Humboldt (1816) « Einleitung zur Agamemnon – Übersetzung » traduit en italien dans, Giovanna Franci & Adriano Marchetti (a cura di) (1991) *« Ripae ulterioris amore » Traduzione e Traduttori*, Gênes, Martietti, pp. 17-32

traduction comme un moyen « d'augmenter l'importance et la capacité expressive de sa propre langue »[69], en y important un « esprit » et des formes littéraires qui lui étaient auparavant étrangers. Il revendique une forme de fidélité : « Mais si avec la traduction on doit acquérir pour la langue et l'esprit de la nation ce qu'ils ne possèdent pas, ou possèdent autrement, il faut exiger avant toute chose, simplement, la fidélité. »[70] Cela ne signifie pas que le traducteur est libre de faire violence à l'idiome d'arrivée : « La traduction a atteint ses ambitieux objectifs si, plutôt que la bizarrerie, elle fait sentir l'étrangeté ; en effet, quand apparaît de la bizarrerie, ce qui obscurcit l'étrangeté, le traducteur montre qu'il n'est pas à la hauteur de l'original »[71].

Ainsi, d'après Eco (in Nergaard, 1995 : 125), une traduction d'Homère ne peut être qu'en grande partie sourcière, et il faudra par exemple conserver « l'aurore aux doigts de rose » chaque fois qu'elle est mentionnée : « Le lecteur doit comprendre qu'à cette époque, l'aurore avait toujours les doigts de rose, à chaque fois qu'elle était nommée. »[72]

D'un point de vue cibliste, on cherchera au contraire à gommer toute forme d'étrangeté. Sur le plan linguistique, on évitera les emprunts, lorsqu'une unité n'a pas d'équivalent évident. De même, on cherchera à substituer les expressions idiomatiques, les proverbes ou les dictons par des expressions équivalentes dans la langue d'arrivée, plutôt que de les calquer. Sur le plan culturel, les références seront adaptées afin d'être plus facilement accessibles pour le lecteur. Eco donne l'exemple de la traduction en russe de son roman, *Il nome della rosa*, dont le texte est émaillé de citations en latin (du type « *De pentagono Salomonis* »), destinées à renforcer l'atmosphère moyenâgeuse de l'intrigue. Pour un locuteur de langue romane, même ne sachant pas le latin, une bonne part de ces citations reste accessible. Mais pour un locuteur slave, les expressions latines translittérées

---

[69] Dans la traduction italienne : "(…) aumentare l'importanza e la capacità espressiva della propria lingua."

[70] "Ma se con la traduzione si deve acquisire per la lingua e lo spirito della nazione ciò ch'essa non possiede o possiede altrimenti, si deve esigere anzitutto semplice fedeltà"

[71] "La traduzione ha raggiunto i suoi alti fini se invece della stranezza fa sentire l'estraneo; infatti dove appare la stranezza in sé e questa addirittura oscura l'estraneo, il traduttore tradisce di non essere all'altezza dell'originale."

[72] "se Omero ripete troppo sovente "l'aurora dalle dita di rosa", non bisogna tentare di variare quell'epiteto, solo perché oggi ci insegnano che non è bene ripetere troppo lo stesso aggettivo. Il lettore deve capire che a quel tempo l'aurora aveva sempre le dita di rosa, ogni volta che veniva nominata"

en alphabet cyrillique ne suggèrent plus rien. Le traducteur a donc opté pour transposer ces citations en slavon ancien tel qu'il était usité par l'église orthodoxe au moyen âge.

Ce type de réinsertion est nécessaire dans la mesure où toute production prend son sens dans un dialogue intertextuel : l'interprétation d'un message est toujours située par rapport aux messages précédents. Bakhtine-Volochinov, cités par J.-M. Adam (1992 : 43), comparent ce lien étroit au maillon d'une chaîne :

> « Toute énonciation, même sous une forme écrite figée, est une réponse à quelque chose et est construite comme telle. Elle n'est qu'un maillon de la chaîne des actes de parole. Toute inscription prolonge celles qui l'ont précédée, engage une polémique avec elles, s'attend à des réactions actives de compréhension, anticipe sur celles-ci, etc. »

L'opposition sourcier / cibliste se brouille lorsque la traduction cherche à restituer la valeur créatrice et innovatrice du texte original. Gilles Deleuze[73] affirme qu'avoir un *style*, c'est parler en une langue étrangère dans sa propre langue : par exemple, des auteurs comme Proust ou Céline se situent en quelque sorte à la limite de la norme du système de leur langue. Ils sont créateurs d'une nouvelle façon d'écrire. Leur traducteur doit (ou peut)-il reproduire ce décalage dans la langue d'arrivée, quitte à violer les normes et usages de celle-ci ? Si l'on considère l'empreinte considérable de certaines traductions sur la langue de leur époque (comme la traduction biblique de Lüther par rapport à l'allemand moderne[74]), ce rôle créateur de la traduction n'est pas à négliger. Le point de vue sourcier peut ainsi aboutir à des choix ciblistes : la recherche de fidélité, par rapport à un texte ancien, peut aboutir à une langue résolument moderne, débarrassée de tout archaïsme. En ce qui concerne la traduction littéraire, il n'y a donc pas d'alternative claire entre sourcier et cibliste, mais une dialectique complexe du même et de l'autre, une tension permanente entre proximité et distance.

La critique de Risset (1985 : 18) illustre l'intrication de cette dualité :

---

[73] cf. *L'abécédaire* de Deleuze, diffusé par ARTE dans l'émission *Métropolis* en 1996.

> « aujourd'hui – avec, disons, Céline, avec Freud – peut-on traduire ce Dante bizarre, ce Dante qui « ne méprise rien » ? André Pézard a courageusement essayé, dans son édition de la Pléiade, en recourant aux archaïsmes, aux néologismes, aux tournures dialectales. Mais, avec des points de surprenante réussite, l'entreprise apparaît malgré tout comme une reconstitution un peu trop volontariste ou plutôt restrictive : d'une part l'archaïsme, dans la langue de la traduction, renvoie à un Moyen Age français, et non italien. D'autre part, l'archaïsme même donne une l'image d'un texte nostalgique, alors que Dante, inventant sa langue, est tout entier tourné vers le futur. »

Il n'existe donc pas de recette *a priori* : c'est toujours l'interprétation personnelle du traducteur qui doit guider ces choix. Comme le note Eco (Nergaard, 1995 : 125), « quant à savoir si une traduction doit être *source* ou *target oriented*, je crois qu'il n'y a pas de règle générale, mais qu'on peut se servir des deux critères alternativement, de manière très souple, suivant les problèmes posés par le texte auquel on a affaire. »[75]

Pour R. H. Bathgate (1980:114), ces choix peuvent être classés en fonction du degré de difficulté impliqué par la recherche d'équivalence. Il propose un modèle normatif où les stratégies les plus simples sont essayées d'abord. Bien sûr, la traduction littérale est le premier degré de cette échelle :

> « Modèle normatif : d'abord on traduit littéralement, puis on se demande si la traduction littérale exprime le sens recherché. Si ce n'est pas le cas, on repart de la forme originale pour retrouver le sens juste. Puis on se demande si le sens juste est également approprié à la situation : est-ce vraiment ce que l'on dit dans ce type de contexte. »[76]

---

[74] Comme le note Neergard (in Neergard, 1993 : 19) « Rosenzweig observe que tandis que la littérature contemporaine de Lüther est difficilement compréhensible pour le lecteur du XXe siècle, sa traduction de la Bible est écrite en une langue qui inaugure l'allemand d'aujourd'hui. Cette remarque montre à quel point la langue utilisée par Lüther dans sa traduction biblique a influencé l'allemand moderne écrit. » ("Per quanto riguarda il tedesco moderno scritto, è prova del forte impatto che ebbe il linguaggio usato da Lutero nella sua traduzione biblica il fatto ricordato da Rosenzweig, il quale osserva che mentre la letteratura tedesca contemporanea a Lutero è difficilmente comprensibile per il lettore di oggi, la sua traduzione della Bibbia è scritta in un tedesco che inaugura quello ancora in uso.")

[75] "Di fronte alla domanda se una traduzione debba essere *source* o *target oriented*, ritengo che non si possa elaborare una regola, ma usare i due criteri alternativamente, in modo molto flessibile, a seconda dei problemi posti dal testo a cui ci si trova di fronte."

[76] "Normative model : First one translate literally; then one asks oneself whether the literal translation conveys the meaning intended. If not, one departs from the original form to convey the right meaning. One then asks oneself whether the translation which gives the right meaning is also appropriate to the situation involved : is that really what you say in that kind of context."

La littéralité n'est plus commandée par un choix interprétatif, mais par un souci d'*économie* (déjà évoquée avec la stratégie du Minimax, p. 88). Ces considérations sont intéressantes d'un point de vue heuristique, car elles font ressortir deux types de mécanisme traductionnel, évoqués par Sager (1994 : 203-204) :

1. la traduction comme action réflexe, très largement automatisée, suivant des mécanismes inconscients acquis par l'expérience.
2. la traduction comme le résultat d'une stratégie adaptative sensible au type de texte. Cet effort adaptatif n'est cependant fourni que pour les problèmes consciemment identifiés.

L'étude des mécanismes mnésiques et des automatismes mis en jeu dans l'exercice de l'interprétariat semble révéler deux niveaux similaires. De manière analogue, Seleskovitch (citée par Laplace, 1994 : 241) oppose « traduction littérale » (ou « traduction réflexe ») et « traduction réfléchie ». La traduction littérale est présentée comme un « transcodage », une traduction au niveau des langues et non pas au niveau de l'interprétation du message. De fait, la traduction littérale est le lieu privilégié des interférences entre systèmes :

> « La traduction littérale sévit à tous les niveaux des langues en contact, du transcodage phonologique pur et simple du type *contrôle* pour *control*, à la traduction de la signification courante des termes au lieu de leur signification pertinente, du type « fenêtre » pour *window* alors qu'il s'agit de " créneau ". » (ibid.)

La traduction littérale peut très vite devenir source de contresens et d'erreur d'interprétation, car les termes complètement transcodables sont rares, comme le remarque Laplace (1994 : 241) : « Ce type de traduction littérale intempestive s'oppose à la traduction interprétative qui est assimilation du sens par le jeu des compléments cognitifs et qui est la seule opération traduisante acceptable pour tout ce qui n'est pas termes transcodables, c'est-à-dire pour l'essentiel du discours. »

Le transcodage n'en reste pas moins une composante du travail traductionnel : « Le transcodage, applicable à certains éléments des textes, est important en traduction, il n'est pas *la* traduction. » (Seleskovitch, citée par Laplace, 1994 : 240). Si Seleskovitch reconnaît l'existence de mécanismes réflexes, c'est pour les subordonner à la vigilance de

l'interprétation réfléchie. Laplace (1994 : 242) décrit ainsi le fonctionnement de cette forme d'automatisme débrayable, qui s'enrichit graduellement des solutions trouvées par une interprétation active :

> « Tant que les termes présents dans la chaîne sonore sont engrammés dans la mémoire, le traitement s'opère de façon réflexe. Si par contre nous rencontrons un terme inconnu, ou simplement un terme connu, mais dans des acceptions que la structure collocative existante rend improbable, le travail réflexe devient conscient et passe du même coup au niveau supérieur, puisqu'il va falloir faire jouer les compléments cognitifs pour comprendre la signification de ce terme nouveau ou sa signification pertinente s'il s'agit d'une nouvelle acception d'un terme connu. Cette opération une fois effectuée, la connaissance fraîchement acquise ira s'engrammer avec les autres connaissances linguistiques et pourra dès lors être utilisée de façon réflexe. »

Dans la perspective de la traduction automatisée, la prise en compte de ces deux types de mécanisme est essentielle. En effet, dans l'état actuel des développements informatiques, le transcodage (ou la traduction littérale, et non mot à mot) représente la seule tâche raisonnablement automatisable. Avant d'étudier les prérequis et le champ d'application des outils d'aide à la traduction (cf. § I.2), il est nécessaire de circonscrire soigneusement cet aspect de l'activité traduisante, de façon à articuler rationnellement traduction automatisée et traduction humaine.

L'analyse contrastive des systèmes linguistiques, abordée d'un point de vue très général, va nous permettre d'esquisser les problèmes posés par la mise en œuvre du transcodage.

I.1.3 D'une langue à une autre: l'analyse contrastive

L'analyse contrastive s'oppose à l'exégèse en ce qu'elle se situe sur le seul plan des transformations linguistiques (cf. p. 23). L'analyse se concentre sur les procédés habituels utilisés dans deux langues pour exprimer des contenus similaires ; elle est contrastive dans la mesure où elle s'attache à comparer ces procédés pour en faire ressortir les similitudes et les différences.

Ce type d'analyse s'intéresse donc aux *signifiés*, tels qu'ils sont codifiés par les systèmes, à l'instar de la grammaire et de la lexicologie. De ce point de vue, le passage d'une langue à l'autre n'est plus à proprement parler une traduction mais plutôt, pour reprendre le terme de Seleskovitch, une forme de *transcodage* :

Transcodage : Langue de départ → Langue d'arrivée

Avec l'analyse contrastive, on ne considère plus les unités linguistiques comme les parties d'un message, mais comme éléments occupant une *place* au sein d'un système, duquel ils tirent une *valeur* déterminée. Pour reprendre l'opposition saussurienne, l'exégèse concerne des occurrences de *parole* tandis que l'analyse est centrée sur l'étude des structures en *langue*.

Est-ce à dire que l'analyse doit porter sur des unités linguistiques abstraction faite de tout contexte ? Evidemment non. Toutes les unités présentent des virtualités sémantiques différemment actualisées en fonction des contextes d'emploi. On ne peut aborder la polysémie sans supposer la variété des contextes, comme l'explique Benveniste (La forme et le sens dans le langage : 38) : « Ce qu'on appelle la polysémie n'est que la somme institutionnalisée, si l'on peut dire, de ces valeurs contextuelles, toujours instantanées, aptes continuellement à s'enrichir, à disparaître, bref, sans permanence, sans valeur constante ». Mais dans la mesure où il s'agit de « somme institutionnalisée », les contextes auxquels s'intéresse l'analyse ne sont plus du ressort de la parole : ce sont des contextes abstraits, reconstruits *in vitro*, qui n'ont qu'une valeur générale. Ces contextes institutionnalisés sont assimilables aux « domaines sémantiques » décrits par Rastier :

> « A chaque type de pratique sociale est associé un type d'usage linguistique que l'on peut appeler discours : ainsi des discours juridiques, politiques, médicaux, etc. Les discours ainsi entendus correspondent à ces formations paradigmatiques

> que sont les domaines sémantiques. Au sein d'un domaine sémantique, il n'existe pas, en règle générale, de polysémie. » (Rastier, 1989 : 39)

Nous retrouvons la distinction précédemment établie entre *contexte* et *situation* : le contexte concerne des codifications générales, qui englobent des sémiotiques extra-linguistiques, comme les domaines de Rastier. Or, même si ces domaines ne peuvent se réduire au seul plan linguistique, ils se manifestent dans les structurations sémantiques de la langue : c'est pourquoi les projections du monde sur la langue sont au cœur de l'analyse contrastive.

I.1.3.1 Contrastes

Ce n'est pas le lieu de rentrer dans le détail d'une étude contrastive entre deux systèmes linguistiques. Nous ne chercherons ici qu'à dégager des principes généraux permettant de circonscrire les phénomènes contrastifs, afin d'identifier quels rôles ils peuvent jouer dans l'exercice de la traduction.

I.1.3.1.1 Deux types de relations

D'après Gideon Toury (1980, trad. in Nergaard, 1995 : 106)[77] trois relations interdépendantes sont en cause dans la traduction d'une entité linguistique dans une autre langue :

1. La relation de chaque entité avec son propre système. Pour la traduction, ce niveau concerne l'acceptabilité du texte cible.

2. La relation des deux entités (source et cible) entre elles, déterminée par la conservation d'une constante. Cette relation se pose en terme d'équivalence, de correspondance, etc.

3. La relation entre les deux codes ou systèmes sous-jacents.

---

[77] Texte original : G. Toury (1980) « Communication in Translated Texts. A Semiotic Approach », in *In Search of a Theory of Translation*, Tel Aviv, The Porter Institute of Poetics and Semiotics, pp. 11-18

Dans la mesure où il n'est fait aucune mention du *message*, puisqu'il s'agit d'établir des relations entre unités linguistiques et systèmes, nous pensons que la relation 3 subsume les relations 1et 2. L'équivalence structurale entre deux unités lexicales, indépendamment de tout contexte, concerne les deux systèmes lexicaux tout entiers.

Ce que cette catégorisation montre mal, c'est qu'il existe deux types de systématicité :

- sur le plan *intralinguistique*, la systématicité du code linguistique, d'où dérivent les unités de la langue. Les grammaires et dictionnaires décrivent ce premier niveau.

- sur le plan *interlinguistique*, la systématicité liée à la pratique de la traduction, ou s'établissent des équivalences entre deux codes. Ce type de relation est décrit par les grammaires contrastives et les dictionnaires bilingues : il inclut néanmoins les deux codes pris indépendamment, et englobe par conséquent le premier niveau. Par rapport aux niveaux d'équivalence traductionnelle précédemment dégagés, qui se situaient au niveau du *sens*, il s'agit d'un autre type d'équivalence, relatif à des *significations*. Nous l'appellerons *équivalence sémantique*.

I.1.3.1.2 Principes de comparaison

Comparer des systèmes linguistiques, sur le plan du lexique, des parties du discours, des catégories grammaticales ou de la syntaxe, revient à effectuer un travail d'*analogie*. En effet, les langues, en tant que « système où tout se tient », sont des systèmes *sui generis* où la totalité et les éléments s'interdéfinissent de manière circulaire. De fait, les unités de deux langues se correspondent rarement terme à terme, comme le note Pergnier (1993 : 79) :

> « [Le] phénomène de non-coïncidence des signes de deux langues différentes, qui était considéré autrefois comme étant le fait d'un nombre limité de mots ou de formes, a maintenant pris sa juste place dans l'étude linguistique en tant que phénomène central, et ce sont maintenant les cas de coïncidence absolue, qui sont, à juste titre, considérés comme des exceptions. »

Cette clôture, même relative, constitue un obstacle à la comparaison directe, car les systèmes ne sont pas isomorphes entre eux. Il n'y a donc jamais relation d'*identité* entre des unités signifiantes, mais plutôt rapport d'*analogie* : par exemple l'anglais *college* s'oppose à *high school* comme le français *université* s'oppose à *lycée*. On a donc :

$$\frac{college}{high\ \ school} \equiv \frac{université}{lycée}$$

ce qui n'implique nullement que *college* ait tout à fait la même signification et / ou désigne les mêmes réalités que *université*.

Nous distinguerons deux types de comparaisons, selon les deux aspects de la forme et de la substance :

– *comparaison des formes*

Par exemple, on peut confronter les 15 voyelles du système vocalique français avec les 7 voyelles de l'italien. Mais les formes de « première articulation » présentent plus d'intérêt vis-à-vis des problèmes traductionnels que les formes phoniques (sauf bien sûr en ce qui concerne le style, le rythme, la poésie, etc.).

Par exemple, si l'on compare la distribution des genres entre le français et l'italien, on peut en dériver des règles contrastives purement formelles d'une grande généralité :

mot féminin français avec le suffixe - *eur*
↔ mot masculin italien avec le suffixe - *ore*

Exemples :

fr.: une odeur
it.: *un odore*

fr.: une saveur
it.: *un sapore*

fr.: une douleur
it.: *un dolore*

Ces règles peuvent se situer à un niveau de généralité grammatical :

- en français, on accorde l'adjectif avec le substantif en genre et en nombre.
- en anglais l'adjectif est invariable.

A un niveau encore plus général, pour pouvoir comparer, une telle règle suppose l'établissement préalable d'un rapport d'analogie entre les adjectifs français et les adjectifs anglais.

La comparaison des formes concerne aussi les contenus : là où le français opère une distinction entre *langue* et *langage*, l'anglais n'emploie qu'un mot : *language*.

– *comparaison des substances*

Les comparaisons précédentes se situaient au plan formel, mais la traduction se joue essentiellement au niveau de la *substance* du contenu (concepts, représentations, etc.), et c'est généralement cette substance qui permet de confronter et d'assimiler, en contexte, deux unités. Pergnier (1993 : 35) résume assez clairement cette dialectique de la différence et de l'identité, qui se joue sur des niveaux hétérogènes :

> « Il est à la fois vrai et faux de dire, comme on l'a beaucoup dit dans le cadre structuraliste, que les unités d'une langue sont intraduisibles dans une autre langue. Cela est vrai car, en effet, *terre* n'est pas équivalent à *land*, puisque ce mot recouvre des notions que *land* ne recouvre pas (et qu'inversement *land* recouvre des notions qui ne sont pas désignées par *terre* : pays, territoire, etc.). Mais cela est faux dans la mesure où, si les conditions contextuelles sont réunies, *terre* et *land* peuvent recouvrir exactement la même notion, pourvu que le seul trait sémantique spatial – inscrit implicitement dans le signifié de *terre*, et explicitement dans le signifié de *land* – soit le point d'interférence entre les deux mots. »

Là encore, l'identification découle d'un rapport d'analogie, mais non plus seulement entre des unités linguistiques : ce sont les liens entre les unités linguistiques et leurs contextes qui sont analogues. Dans l'exemple de Pergnier, il y a identité notionnelle dans le contexte maritime, pour désigner une terre vue d'un navire :

$$\frac{land}{/\,contexte\ \ maritime\,/} \equiv /\,notion\ \ de\ \ terre\ \ \ ferme\,/ \equiv \frac{terre}{/\,contexte\ \ maritime\,/}$$

I.1.3.1.3 Trois niveaux d'étude entrelacés : *lexique*, *grammaire*, *idiome*

De nombreux linguistes se sont penchés sur les problèmes de traduction découlant des contrastes linguistiques. Jakobson (1963 : 83) donne l'exemple de la phrase anglaise *I hired a worker*. Pour la traduire en russe, il faut spécifier l'aspect perfectif ou imperfectif de la prédication ainsi que le genre masculin ou féminin correspondant à la personne désignée par *worker*. Or ces informations ne figurent pas dans l'énoncé anglais : il y a en quelque sorte une « *sous-spécification* » du système anglais par rapport au russe. Pour résoudre ce problème, on peut éventuellement déduire les informations additionnelles du contexte textuel et extra-textuel (et c'est ce que font les traducteurs, pour qui ce genre de divergence est rarement problématique). Si ces informations ne sont pas disponibles, le traducteur est contraint de faire un choix arbitraire, ce qui peut aboutir à une forme de *surtraduction*, cas de figure où l'énoncé traduit contient plus d'information que l'énoncé original.[78]

Jakobson (1961 : 84) en déduit qu'en ce qui concerne les traductions, les problèmes proviennent plus souvent des contraintes imposées par les systèmes, que de la limitation de leur capacité expressive : « Les langues diffèrent essentiellement par ce qu'elles *doivent* exprimer, et non par ce qu'elles *peuvent* exprimer ».

L'exemple de Jakobson concerne des contrastes au niveau grammatical. D'autres problèmes surgissent avec le lexique. Certaines langues semblent présenter des béances, des lacunes dans l'organisation de leur lexique : il n'est pas rare que le « bon » mot fasse défaut. Par exemple, l'anglais n'a pas d'équivalent précis pour désigner les « bouquins » vendus sur les quais de la Seine. Ces « défaillances » peuvent toucher des mots très courants, comme certains mots outils appartenant au vocabulaire de base de la langue. Jakobson (1963 : 82) donne l'exemple du samoyède, qui possède une conjonction signifiant « et / ou », mais pas de *et* ni de *ou*, comme en français.

---

[78] Nida (in Nergaard, 1995 : 167), soulève ce problème en parlant de traduction biblique. Il en cite deux illustrations : dans le dialecte des Zapotec (sud du Mexique), on fait la distinction entre les actions qui s'effectuent pour la première fois et celles qui se répètent. Quand Jésus arrive à Capharnaüm, il faut donc préciser s'il y est déjà passé ou non, ce qui n'est pas précisé dans le texte. De même, quand la langue d'arrivée requiert des formules honorifiques suivant la position sociale (par exemple en coréen), il faut faire un choix de traduction, lourd de conséquences sur le plan de l'interprétation : considérer Jésus comme un rabbin (ces disciples le considéraient comme tel) ou comme un homme du peuple.

Sur la base des problèmes que nous avons évoqués, il serait tentant d'analyser séparément le lexique et la grammaire des deux langues comparées. Ceci n'est possible que dans certaines limites, pour deux raisons :

1. D'une part, il existe, comme le signale Fuchs (1981 : 41), « (…) un continuum entre faits de syntaxe et faits de lexique ».

De manière très informelle, on peut dire que le lexique concerne l'ensemble des mots d'une langue, et la grammaire désigne l'ensemble des règles permettant de combiner ces mots. Très schématiquement, on pourrait opposer d'un côté des unités lexicales et de l'autre côté des règles combinatoires. La réalité est un peu plus compliquée, puisqu'un stock fermé d'unités, les morphèmes grammaticaux, sont étroitement liés à cette combinatoire, et sont traditionnellement rangés dans la grammaire.

Par rapport au couple lexique / grammaire, deux points de vue s'opposent :

– certains auteurs pensent qu'il y a un primat de la grammaire sur le lexique. Par exemple, Lucien Tesnière, qui affirme (1959 : 25) :

> « C'est peut-être que trop souvent on *part de la notion de mot* pour arriver à la notion de phrase, au lieu de partir de la notion de phrase pour *arriver à la notion de mot*. Or on ne saurait définir la phrase à partir du mot, mais seulement le mot à partir de la phrase. Car *la notion de phrase est logiquement antérieure à celle de mot*. »

Le mot découlerait donc des relations apparaissant d'abord sur l'axe syntagmatique.

– à l'opposé, certains auteurs considèrent que les propriétés combinatoires sont inscrites au niveau de chaque unité lexicale. Pour Mel'čuk *et al*., la lexie, « unité de base de la lexicologie » (1995 : 15), est une entité trilatérale ayant :

> « - un sens (le *signifié* saussurien),
> - une forme phonique ou graphique (le *signifiant* saussurien),
> - et un ensemble de traits de combinatoire (...) » (1995 : 16)

Or, ces traits de combinatoire sont censés renfermer et déterminer les règles de grammaire : « Les règles qui réunissent les lexies en syntagmes, les syntagmes en phrases, et les phrases en discours sont donc nettement secondaires par rapport aux lexies – en ce sens que leur nature et leur forme sont déterminées par les lexies. » (1995 : 17).

Nous pensons qu'il s'agit deux de façons différentes d'appréhender les mêmes phénomènes. Comme dit Bernard Pottier (1992 : 35), « L'un fait obligatoirement référence à l'autre. Une lexie entraîne un certain nombre de pressions sur son entourage (rections, sélections, affinités…). » Il y a une co-détermination des deux plans, certaines règles combinatoires de niveau lexical étant susceptibles de faire émerger des règles générales d'un point de vue grammatical et réciproquement. Plus précisément, nous pensons qu'il y a coexistence dynamique de plusieurs niveaux de systématicité. Les règles grammaticales forment des prescriptions (flexions, accords, agencement) ou des interdictions (incompatibilités). Mais ces règles ne sont ni complètes ni consistantes, et le lexique impose des combinaisons originales, irrégulières du seul point de vue grammatical, comme dans certaines unités composées : *un à-valoir, à la va comme je te pousse, faire la part belle,* ou des expressions semi-figées définissant leur propre système combinatoire : *c'est grave de chez grave,* etc.

La seule différence entre les combinatoires lexicales et les règles grammaticales tient dans leur degré de généralité, dans le fait qu'elles sont attachées à des formes particulières ou de petits paradigmes plutôt qu'à des classes aussi générales que les parties du discours.

2. D'autre part, s'il est possible de comparer indépendamment les organisations lexicales et les systèmes grammaticaux pris dans leur ensemble, le transcodage de constructions linguistiques particulières fait toujours intervenir à la fois lexique et grammaire.

Si l'on reprend l'exemple de Jakobson : *I hired a worker,* le problème de surtraduction n'a de sens qu'au niveau très général des grammaires comparées de l'anglais et du russe. En français, où la spécification du genre de *worker* est aussi requise, il suffirait de traduire : *j'ai embauché quelqu'un*. De même pour le problème de la traduction de *et* et

de *ou* en samoyède, Jakobson donne une solution qui fait intervenir une construction syntaxique plus élaborée :

> *Jean et Pierre viendront* se traduit littéralement par [Jean et / ou Pierre viendront tous deux]
>
> *Jean ou Pierre viendra* se traduit littéralement par [Jean et / ou Pierre, l'un des deux, viendra]
>
> En définitive, le lexique pallie la syntaxe là où elle fait défaut, et réciproquement.

Il apparaît qu'on ne peut comprendre correctement les rapports complexes qu'entretiennent les deux pôles du lexique et de la grammaire si l'on ne tient compte d'un troisième type de détermination, essentiel pour le transcodage : l'*idiome*.

Par *idiome* nous désignons l'ensemble des réalisations du *système* : il est le produit de l'usage et reflète les habitudes langagières des locuteurs d'une communauté linguistique, dont il manifeste une volonté de convergence. Ce niveau concerne un grand nombre de phénomènes linguistiques qu'on ne peut décrire qu'en terme de stéréotypes et d'habitudes, comme les exemples donnés par André Clas (1994 : 576) :

> « Toute langue est une série d'habitudes : les " formules langagières " nous sont données. Bien sûr elles peuvent être plus ou moins riches, mais dans de nombreux cas, elles sont contraignantes. Ainsi quelqu'un qui a " une bonne mémoire " a une *mémoire d'éléphant*, une *mémoire prodigieuse*, une *mémoire excellente*, une *mémoire étonnante*, … en français et *ein gutes Gedächtnis*, *ein ausgezeichnetes Gedächtnis*,… en allemand. »

L'idiome se situe à la fois au-delà et en deçà du système, car d'une part il est susceptible de rentrer en conflit avec les règles morphosyntaxiques (on peut donner l'exemple des syllepses *c'est eux, elle a l'air belle*) et d'autre part il n'exploite qu'une très faible part de la combinatoire autorisée par le système : les expressions figées, les locutions idiomatiques ou idiotismes (angl. *how do you do ?, comment ça va ?*), les collocations (*faire le plein d'essence, prendre de l'essence*, it. *fare benzina* sont idiomatiques, mais pas * *prendre le plein d'essence, * faire de l'essence*, it.* *prendere benzina*), sont autant de traces de ces conventions partagées mais difficilement systématisables au niveau grammatical.

Les productions de l'idiome sont partiellement incluses dans l'ensemble des réalisations prévues par le système, pourtant l'idiome n'est pas un *sous-système* dont les

prescriptions s'ajouteraient à celles du système. On ne peut le décrire par l'ajout de *règles* supplémentaires : sinon on pourrait le caractériser de manière systématique, et il coïnciderait avec le système fonctionnel. L'idiome est par essence non systématique. Il s'établit à la frontière du système et se nourrit de son incomplétude. Les usages qu'il privilégie découlent de choix arbitraires. Enfin, même s'il véhicule des stéréotypes et des « clichés » (Clas, 1994), l'idiome ne bride pas la liberté créatrice du système : il écarte des combinaisons sans détruire la fécondité de la combinatoire.

L'identification des constructions idiomatiques est fondamentale pour la traduction, comme pour les opérations de transcodage, puisqu'elle garantit le respect du « naturel » dont parlait Nida. C'est la connaissance de l'idiome qui permet d'établir l'équivalence suivante :

angl. : *to be left unemployed*
fr. : *perdre son emploi*

Pour désigner ce type d'équivalence, à la fois sémantique et idiomatique, nous parlerons simplement d'« *équivalence linguistique* ».

En définitive, on ne peut réduire les phénomènes langagiers à leurs structures purement formelles : la mise en évidence structurale, presque algébrique, d'unités discrètes définies par des réseaux d'opposition ne doit pas nous conduire à négliger de très nombreux phénomènes scalaires, admettant des degrés d'intensité. Le point de vue *continuiste* s'applique à nombre d'aspects du langage : l'usage d'une forme linguistique peut être plus ou moins répandu, plus ou moins fréquent, plus ou moins neuf ou ancien ; la substance de l'expression ou du contenu s'inscrit dans des espaces repérables par des axes de variation continus ; liée à cette substance, l'opacité ou la motivation d'une expression linguistique sont perçues par le locuteur à des degrés variables ; un composé peut être plus ou moins figé, plus ou moins libre ; une règle de grammaire peut être plus ou moins admise (p. ex. la concordance de l'imparfait du subjonctif), plus ou moins générale dans son application, etc.

Sur la base de cette vision continuiste, on peut essayer de synthétiser les rapports entre *système* et *idiome* à partir d'une vision « topologique », sur la base de trois continuums de phénomènes reliant les deux pôles opposés du lexique et de la grammaire :

1. le continuum des unités de base de la première articulation, ou morphèmes. Dans l'ordre, du lexique à la grammaire, on trouve par exemple : les mots simples, les racines (intervenant dans les opérations de dérivation), les affixoïdes, les affixes, les mots outils (prépositions, conjonctions, pronoms, etc.) et enfin les morphèmes grammaticaux ou grammèmes. Ce continuum décrit des stocks d'unités allant de l'ouverture à la fermeture, de l'autonomie à la dépendance syntaxique.

2. le continuum des règles combinatoires. Ces règles vont du local au global, du particulier au général. Au niveau lexical, on parle du *régime* de la lexie (Mel'čuk *et al*., 1995), imposant parfois des combinatoires sans généralité (p. ex. les syntaxes spécifiques, comme dans l'opposition *vulgaire photo* / *photo vulgaire*, ou les paradigmes restreints *ça ne vaut pas un sou* / *clou* / *radis* / *kopeck*). Du côté grammatical, on trouve les règles de la morphosyntaxe, souvent de grande portée, comme les flexions et les accords. Entre les deux, il existe des règles de portée plus réduite qui ne sont toutefois pas liées au régime particulier d'une lexie. Considérons les exemples suivants :

fr. : *les mains dans les poches*
angl. : *with his hands in his pockets*

fr. : *un pistolet au poing*
angl. : *with a gun in his hand*

On observe une règle relative à l'usage du déterminant concernant les parties du corps : là où l'anglais emploie l'adjectif possessif, le français recourt à l'article défini. Mais cette règle est de portée réduite et connaît des réalisations variables suivant le degré de proximité des possessions :

il s'est brûlé la main	*vs*	** il a abîmé sa main*
il s'est brûlé la veste	*vs*	*il a abîmé sa veste*
** il s'est brûlé la voiture*	*vs*	*il a abîmé sa voiture*

Ces règles intermédiaires, situées à mi-chemin entre les catégories générales et les spécificités lexicales (p. ex. en français, la règle de l'emploi du subjonctif pour certaines conjonctions de subordination ; en italien, les verbes esclaves *sapere*, *volere*, *potere* héritant de l'auxiliaire du verbe modalisé, etc.) sont nombreuses et variées.

Il est évidemment difficile de quantifier nettement le degré de généralité des différentes règles d'un système, même si l'étendue des classes concernées (superparties du discours, parties du discours, sous-classes de parties du discours, paradigmes restreints à quelques unités) constituent une indication.

3. corrélé à l'axe précédent, le continuum allant du figement à la liberté combinatoire, définissant le degré de blocage ou d'actualisation des règles de composition morpho-syntactico-sémantiques.

La figure 7 synthétise, de manière extrêmement simplifiée, les relations entre les trois niveaux du lexique, de la grammaire et de l'idiome.

Nous pouvons désormais étudier les contrastes selon ces trois niveaux, en gardant à l'esprit qu'ils sont entrelacés.

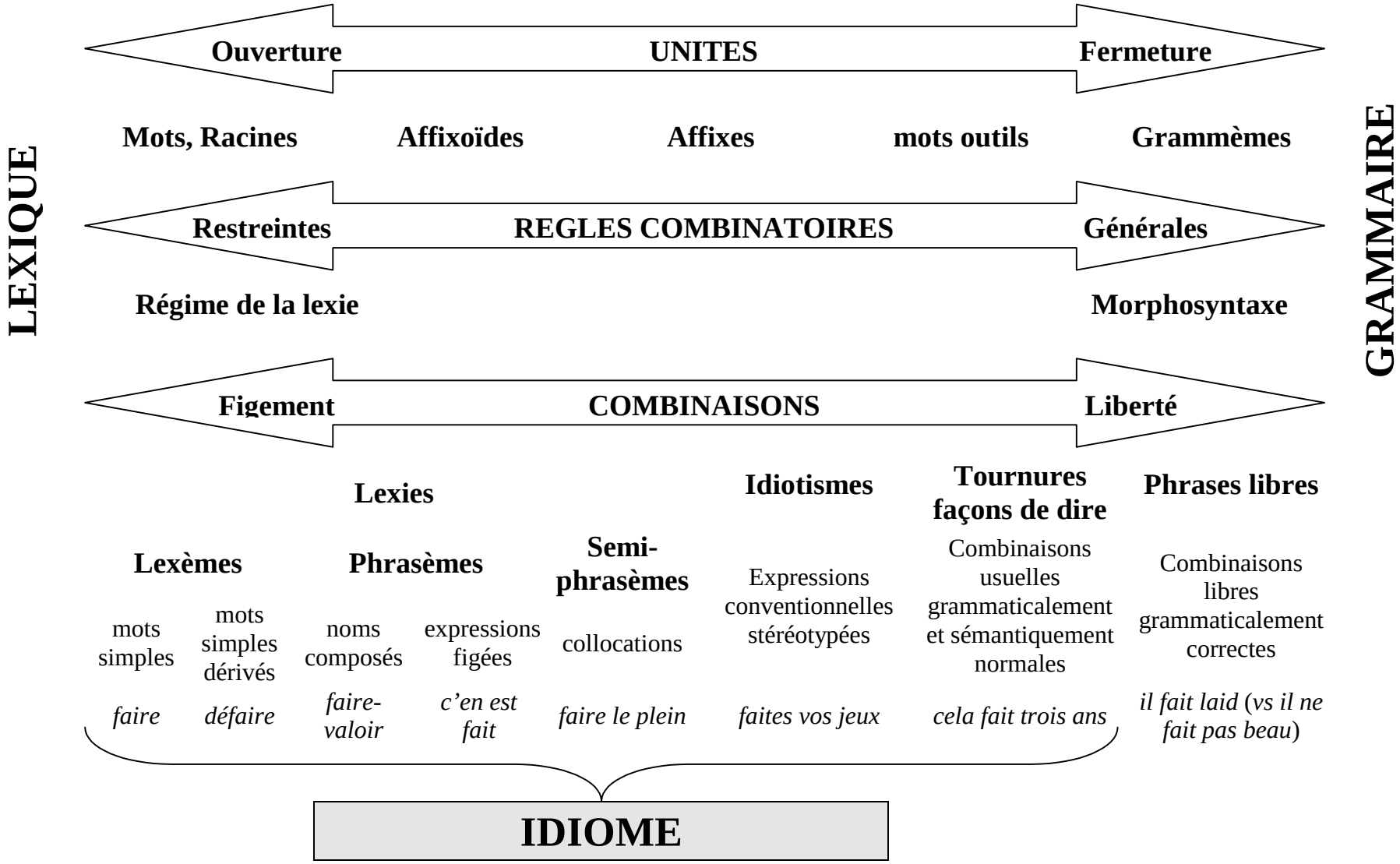


figure 7 : *lexique, idiome, grammaire* : *les trois continuums de phénomène*

I.1.3.2 Contrastes lexicaux

Les phénomènes de contrastes lexicaux occupent une position centrale dans l'étude du transcodage, puisque c'est la possibilité de trouver des lexies équivalentes, sur le plan sémantique, qui conditionne une grande part des transformations dans le passage d'un code à l'autre. En outre, les structures lexicales, tant sur le plan sémantique que sur celui de la combinatoire, prédominent dans la détermination des irrégularités et des préférences idiomatiques. On s'éloigne ainsi de la conception naïve de « lexique répertoire » dénoncée par Mounin (1963 : 26) selon laquelle le lexique serait assimilable à une collection d'objets inertes agencés et fléchis par les seules règles de la grammaire.

Pour aborder de manière économique ces organisations, nous articulerons notre étude suivant les deux axes orthogonaux et complémentaires du paradigme et du syntagme :

– *Structurations syntagmatiques*

Le long de cet axe, l'unité lexicale entretient deux types de rapport : des relations externes avec les autres constituants de la phrase ; des relations internes avec ses propres constituants. Ces deux types de relations syntagmatiques définissent d'une part le *régime* de la lexie, conditionnant ses possibilités combinatoires, et d'autre part sa *motivation* et son *identité* même, puisqu'il peut être question de décomposer une unité en unités lexicales autonomes plus petites. Les valences d'une lexie ressortissent donc au plan de la combinatoire externe, tandis que les phénomènes de dérivation et de composition, associés au plan sémantique de la motivation, intéressent sa constitution interne.

Ces deux types de combinatoire sont bien sûr étroitement liés puisqu'ils dépendent de ce qu'on entend par unité : du point de vue de l'unité *porte*, l'expression *porte de garage* est une combinaison externe déterminée par son régime ; du point de vue de *porte de garage*, prise comme une unité en tant que telle, *porte* est un constituant qui en explicite la signification. Nous verrons que cette question constitue un enjeu majeur de l'exploitation des corpus bi-textuels.

– *Structurations paradigmatiques*

A ce niveau, les unités lexicales s'organisent en classes à l'intérieur desquelles les significations se définissent et se délimitent. Les structures émergentes comportent deux aspects : une organisation formelle des espaces de signification, ou champs sémantiques, et les rapports dénotatifs liant les signifiés aux objets extra-linguistiques, référents ou concepts. Or, d'un point de vue contrastif, le transcodage des signifiés n'est pas sans problème, puisqu'il est reconnu que dans l'exercice de la traduction certaines lexies n'ont pas d'équivalent. Faut-il y voir un obstacle insurmontable ?

I.1.3.2.1 Aspects paradigmatiques

La conception naïve renvoie une image statique du lexique, semblable un répertoire de mots étiquetant des objets préexistants du monde. Saint Augustin, dans sa description de l'apprentissage de la parole, nous donne une illustration de ce que Wittgenstein, dans ses *Investigations philosophiques* (1961), appelle la *définition démonstrative* des mots :

> « Alors, je captais par la mémoire les noms que j'entendais donner aux choses, et qui s'accompagnaient de mouvement vers les objets ; je voyais et je retenais que l'objet avait pour nom le mot qu'on proférait, quand on voulait le désigner. »[79]

Depuis Saussure, la simplicité de cette relation allant de la chose au mot s'est dispersée dans la complexité du *système* : les mots ne désignent plus rien de manière isolée, mais de par leur *place* au sein du système de la langue. Une unité lexicale est signifiante dans la mesure où elle se différencie des unités voisines, par certains traits, et prend ainsi sa *valeur* au sein du système.

Pour reprendre une métaphore de Mounin, le lexique est structuré comme un « filet », aux mailles plus ou moins resserrées. Et l'étendue sémantique de chaque unité lexicale y est délimitée par les unités voisines.

Du point de vue contrastif, les problèmes d'équivalence surgissent dès que les mailles des filets ne se superposent pas correctement. Comme le note Nida (1969 : 19), ce n'est pas l'étendue de ces systèmes classificateurs qui est en cause, car ils peuvent tous prétendre à l'universalité, mais leur structure interne :

> « 1) Chaque langue couvre la totalité de l'expérience avec un ensemble de signes linguistiques, les mots, qui désignent différents aspects de l'expérience, et 2) chaque langue est différente des autres dans la manière de classer ces différents aspects à partir de l'ensemble de ses signes. »[80]

La valeur des signifiés découlant de l'ensemble du système, une première question se pose : des systèmes différents peuvent-ils dégager des signifiés similaires ? Mais la métaphore du filet laisse pressentir de nombreuses distorsions, et le véritable problème mérite peut-être d'être posé différemment : si la plupart des signifiés lexicaux sont différents, sur quel critère peut-on établir l'équivalence lexicale ? Pergnier (1993 : 155) en fait une question centrale : « Le problème théorique plus général de la traduction est (...) celui de la possibilité de la convergence des champs lexicaux. »

Pour commencer à y répondre, nous allons considérer dans un premier temps, ce que révèle la forme des « filets ».

I.1.3.2.1.1 Plan de la forme

Sur le plan de la forme, l'antonymie est la relation d'opposition sémantique minimale. Il est aisé, dès ce stade, de relever des oppositions qui n'existent qu'au sein d'un système linguistique donné (bien entendu, en rapport avec des contenus culturels spécifiques). Le *Yin* et le *Yang*, par exemple, qualifient en Chine une opposition canonique qui ne prend son sens qu'au sein d'un système complexe d'oppositions et de bipolarités. En Italie, on a coutume de distinguer la *pasta in bianco* (sans tomate) et *in rosso* (avec tomate). En Grande Bretagne, on distingue entre *black shoes* et *brown shoes*, les premières désignant les chaussures « de ville » et les secondes désignant toutes les autres (indépendamment de la couleur). Ces expressions antonymes manifestent des oppositions absentes de la langue française.

On peut élargir la notion de couple antonyme à la notion plus générale de *taxème*, ainsi définie par Rastier (1989 : 55) « Le taxème est la classe minimale où les sémèmes

---

[79] Saint Augustin (réédition de 1964), « L'apprentissage de la Parole » in *Les confessions*, Paris, GF, pp. 23-24.

[80] "1) each language covers all of experience with a set of verbal symbols, i.e., words to designate various features of experience, and (2) each language is different from all other languages in the ways in which the sets of verbal symbols classify the various elements of experience."

sont interdéfinis : par exemple, ‘cigarette’, ‘cigare’, ‘pipe’ s’opposent au sein du taxème //tabac// ». En se plaçant sur le plan sémantique, et non lexical, nous nommerons *champ sémantique* l’espace des variations sémantiques couvert par un tel taxème. Pour prolonger la métaphore spatiale, le champ sémantique, défini par son noyau sémique, correspond à la surface quadrillée par les unités du taxème, de manière plus ou moins fine.

Notons que la minimalité de ces classes pose problème, puisque, comme le montrent les études psycholinguistiques, toutes les catégorisations linguistiques sont soumises à des phénomènes prototypiques (E. Rosch *et al*., 1976 ; G. Kleiber, 1990) :

- il existe une hiérarchie interne des membres de la classe, certains étant plus centraux, plus saillants que les autres (p. ex. *cigarette*) ;
- de même les traits sémantiques décrivent un continuum du plus au moins saillant (/avec du tabac/, /qui se fume/, /présence de nicotine/, /cancérigène/, etc.) ;
- les frontières sont floues : l’appartenance des éléments périphériques est incertaine (p. ex. *cigarette à* l’*eucalyptus*, *tabac à priser*, *cannabis* …) ;
- les niveaux de généricité ou de spécificité (hyperonymie / hyponymie) sont mêlés (*cigarillos* est-il un hyponyme ou un cohyponyme de *cigare* ?).

Cette vision prototypique permet de nuancer la structuration en classes sans la remettre en question : même si les classes sont floues, tant sur le plan horizontal de leurs frontières que sur le plan vertical des hiérarchies, elles existent intuitivement pour tout locuteur et présentent des différences d’organisation dans chaque langue.

Or, les taxèmes sont très variables en taille et en richesse suivant les langues et les cultures matérielles. Par exemple, on relève une cinquantaine d’expressions pour désigner le pain dans la région d’Aix-en-Provence en 1959 (Mounin, 1963 : 65) :

> *la baguette, le boulot, la chenille, le chemin de fer, le coupé, la couronne, l’épi, le fendu, le fil de fer, la ficelle, la flûte, la fougasse, le fuseau, la fusée, le gressin, le grichon, le kilomètre, le longuet, la main, le marseillais, le pain d’Aix, le pain de*

mie, le pain mousseline, le restaurant, la rosace, le roulé, le saucisson, le seiglon, la tête d'Aix, la tière, la tresse, la torsade, le tordu, la tomate.

Tous ces termes renvoient à des pains différents et sont encore d'usage courant. On peut citer bien d'autres exemples, devenus des « classiques » : les gauchos argentins utilisent deux cents termes pour désigner les différents aspects de la robe des chevaux ; certaines langues africaines dénomment différemment une soixantaine d'espèces de palmiers ; les Inuits distinguent plus d'une centaine de formes de neige, etc. Ces exemples sont populaires parce qu'ils frappent l'esprit, et semblent ériger un mur d'intraduisibilité entre des langues qui se « spécialisent » dans des domaines différents.

Mais insistons encore, avec Nida, Jakobson, Mounin et bien d'autres, sur le fait qu'aucun système linguistique n'est limité *a priori*. Lorsqu'il n'y a pas d'équivalent lexicalisé dans la langue d'arrivée, le traducteur dispose de toute une palette de solutions pour pallier le manque de vocabulaire :

– *le calque* : pour Vinay & Darbelnet (1959 : 47-52), c'est le fait d'« emprunte[r] à la langue étrangère le syntagme, mais [de] tradui[re] littéralement les éléments qui le composent » :

angl. :	*politically correct*	→ fr. :	*politiquement correct*
angl. :	*to take into account*	→ fr. :	*prendre en compte*
angl. :	*honeymoon*	→ fr. :	*lune de miel*
angl. :	*non-sense*	→ fr. :	*non-sens*
angl. :	*freethinker*	→ fr. :	*libre-penseur*[81]

On peut aussi calquer un procédé de dérivation :

angl. : *merchandising* → fr. : *marchandisage*

– *l'emprun*t : on conserve le mot d'origine, en l'adaptant éventuellement sur les plans phonémiques et graphémiques. En traduction, les emprunts non encore attestés sont souvent accompagnés de notes explicatives.

angl. :	*mouse*	→ it. :	*mouse*
fr. :	*toilette*	→ it. :	*toalet* (graphie répandue mais non officielle)

[81] cf. Henriette Walter (1997 : 192)

angl. : *film* → malais : *filem*

Il n'est pas rare qu'on emprunte des mots existant déjà dans sa propre langue, mais avec un sens légèrement différent (alors qu'il existe déjà un équivalent lexicalisé) :

angl. : *to realise* → fr. : *réaliser* (« prendre conscience de … »)
angl. : *to control* → fr. : *contrôler* (« maîtriser la situation … »)
angl. : *to support* → fr. : *supporter* (« soutenir une équipe … »)[82]
angl. : *application* → fr. du Québec : *application* (« candidature »)

– la néologie :

angl. : *marketing* → fr. : *mercatique*

– une circonlocution explicative : une tournure périphrastique permet d'expliciter le sens du mot à traduire.

fr. : *bouquins* → angl. : *old books*
angl. : *root beer* → fr. : *boisson pétillante non-alcoolisée aux extraits de plantes*

Notons que les frontières entre les procédés énumérés ci-dessus sont floues. L'exemple suivant mêle à la fois la néologie, le calque et l'emprunt :

angl. : *computational linguistics* → fr. : *linguistique computationnelle*.

Finalement, comme pour la grammaire (cf. l'exemple de Jakobson à propos du genre de *worker*), les problèmes de spécification sont peut être plus épineux lorsqu'on traduit vers une langue qui impose plus de distinction, par manque d'un terme générique. C'est le cas dans l'exemple cité par Mounin (1963 : 66) : « Comment (...) traduire frère et sœur en maya, lorsque cette langue n'a pas de mots pour l'extension de ces notions chez nous, mais des termes distincts pour frère plus jeune, et frère plus âgé ? ». Des langues plus proches de nous présentent les mêmes difficultés. Le latin a quatre termes pour la distinction *tante* / *oncle* :

latin : *avunculus* → fr. : *oncle frère de la mère / maternel*
latin : *patruus* → fr. : *oncle frère du père / paternel*
latin : *amita* → fr. : *tante sœur du père / paternel*
latin : *matertera* → fr. : *tante sœur de la mère / maternelle*

[82] *cf*. Henriette Walter (1997 : 193)

A chacun de ces termes correspond un sens plein avec des implications pratiques, juridiques et sociales au sein du monde romain. Le passage du latin au français peut nécessiter des circonlocutions. Mais dans le sens inverse, du français au latin (à supposer qu'on traduise dans ce sens, au Vatican par exemple), et si le contexte n'y supplée pas, on retrouve les problèmes de *sous-spécification*.

De même que certaines distinctions sont imposées par le système grammatical, d'autres sont inscrites dans le lexique. L'assertion de Jakobson (1961 : 84), déjà citée, s'applique donc aussi sur le terrain du lexique : « Les langues diffèrent essentiellement par ce qu'elles *doivent* exprimer, et non par ce qu'elles *peuvent* exprimer ».

La finesse des mailles du « filet » sémantique est étroitement liée à la dimension verticale de l'organisation hiérarchique des lexies. Bien souvent, les découpages d'un même champ sémantique réalisent des configurations multidimensionnelles, à la manière d'un empilement de strates de plus en plus finement quadrillées. Par exemple le taxème //tabac// donné en exemple par Rastier peut être traversé par une nouvelle série d'oppositions : /tabac brun/ vs /tabac blond/, /avec filtre/ vs /sans filtre/, /cubain/ vs /toscan/, etc.

Nous empruntons à Nida (1969 : 20) le schéma de la figure 8, qui représente clairement cet aspect hiérarchique :

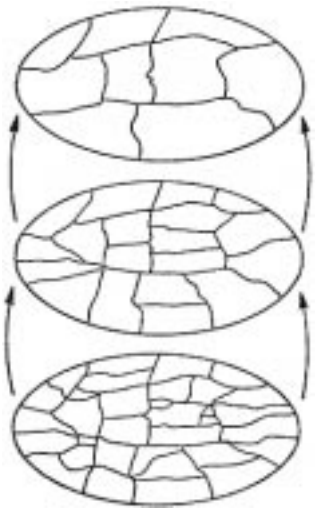
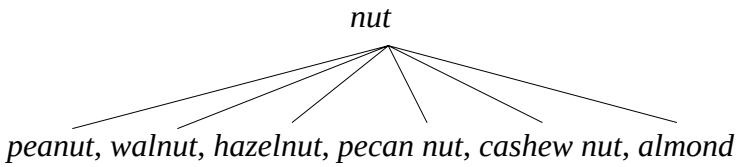


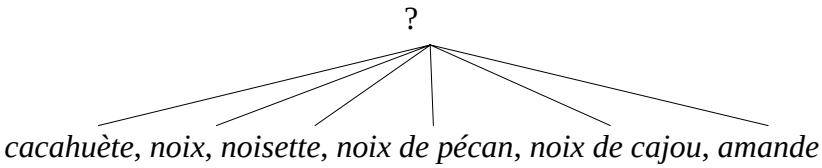
figure 8 : *dimensions horizontales et verticales des découpages sémantiques*

Bien sûr, ce type de schéma donne une vision idéalisée et simplifiée des phénomènes : la généricité, de même que l'appartenance à une classe, ne connaît pas de palier aussi net, et les éléments prototypiques jouent souvent le rôle de générique en subsumant des éléments de même niveau. Par exemple, le caractère générique de *blé* dans le taxème //céréales// est manifesté par des constructions composées, comme pour *blé noir* (sarrasin), ou *blé d'Inde* (« maïs », au Canada).

Ces organisations hiérarchiques, même floues et inconsistantes, ne se laissent pas transposer si facilement d'une langue à l'autre. Là encore on observe des places vides et des dissymétries. La langue bulu, parlée au Cameroun, possède au moins 25 termes pour désigner différents types de panier, mais aucun terme générique ne correspond au français *panier* (Nida, in Nergaard, 1995 : 173). En anglais le lexème *nut* subsume les éléments suivants :



Cette structure devient en français :



On est bien en peine de trouver un équivalent français de l'hyperonyme anglais *nut*. Le mot *noix* joue en partie ce rôle, puisqu'il sert de base pour de nombreux composés ou dérivés. Cependant, employé seul, il prend un sens spécifique : « j'aime les noix » se traduit par « *I like walnuts* » ; *noix* est un prototype de sa classe, mais pas un véritable générique.

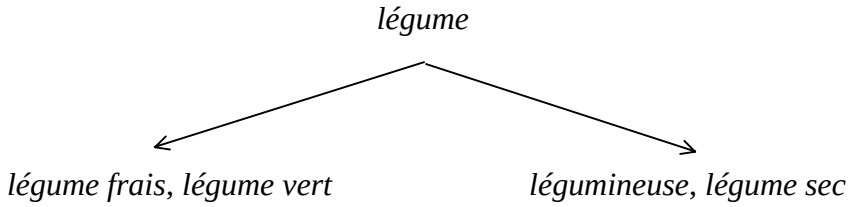
Si l'on considère que les opérations de dérivation et de composition sont révélatrices des classifications spontanées opérées par la langue on note d'autres dissymétries : *chestnut* en anglais est rapporté, de près ou de loin, au même paradigme, ce qui n'est pas le cas pour *châtaigne*.

Notons que ces classifications, pertinentes sur le plan linguistique, n'ont pas nécessairement d'assise scientifique. Elles se fondent sur l'observation de traits sémantiques sélectionnés par la langue : peu nous importe (et peu importe au locuteur) que le sarrasin ne soit pas une céréale où que l'amande soit une graine et non un fruit. Les codes linguistiques, en tant que systèmes taxinomiques intuitifs, présentent des variétés d'organisation et de structures qui sont en grande partie arbitraires (et partiellement motivées). On peut donner l'exemple de classements bizarres ou aberrants, de notre point de vue, du fait de leur éloignement et leur « étrangeté » culturelle, comme en témoigne la description suivante tirée de la *Pensée sauvage* (Lévi-Strauss, 1962 : 55) :

> « Les indiens Navaho, qui se proclament eux-mêmes « grands classificateurs », divisent les êtres vivants en deux catégories, selon qu'ils sont ou non doués de la parole. Les êtres sans paroles comprennent les animaux et les plantes. Les animaux se répartissent en trois groupes : « courants », « volants » ou « rampants »; chaque groupe est, à son tour, recoupé par une double division : celle entre « voyageurs sur terre » et « voyageurs sur eau » d'une part, et d'autre part, celle entre « voyageur de jour » et « voyageur de nuit ». Le découpage des « espèces » obtenu par cette méthode n'est pas toujours le même que celui de la zoologie. Il arrive ainsi que des oiseaux groupés en paires sur la base d'une opposition : mâle/femelle appartiennent en fait au même sexe, mais à des genres différents ; car l'association est fondée, d'une part, sur leur taille relative, d'autre part sur leur place dans la classification des couleurs, et sur la fonction qui leur est assignée dans la magie et le rituel. » [83]

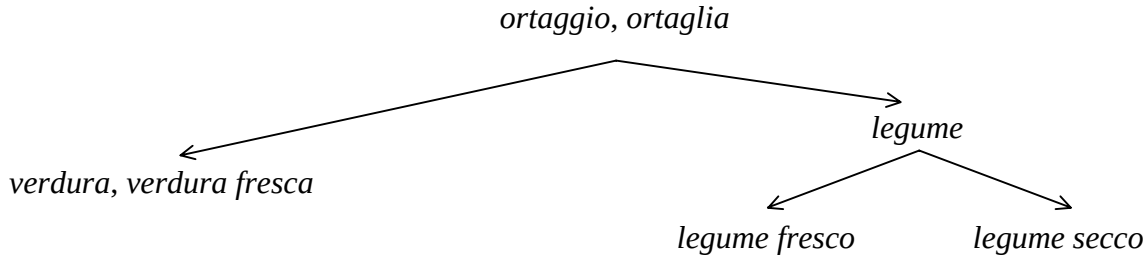
Même entre des cultures apparentées, les hiérarchies peuvent présenter des contrastes qui empêchent l'établissement d'équivalences pleines. Par exemple, en comparant deux champs sémantiques équivalents, correspondant aux unités génériques (fr.) *légume* et (it.) *ortaggio*, on constate que le flou inhérent à ces classes brouille le jeu des correspondances. Nous nous sommes basé sur les classifications proposées par le *Petit Larousse* 1996 et le *Zingarelli* 1998, pour tenter de décrire l'organisation du taxème //légume// dans les deux langues :

[83] Ici, l'auteur ne précise pas si ces classifications sont d'ordre linguistique ou mythique : de toute façon celles-ci influencent celles-là, et réciproquement.



Bien sûr les sous-classes définies par ses unités sont floues, se chevauchent et comportent de nombreuses incohérences. La *pomme de terre* est un *légume* sans être un *légume vert*. Les *légumes verts* sont plus spécifiquement les légumes de couleur verte (*épinards*, *blettes*, *poireaux*, …), même si la *carotte* (cf. le *petit Larousse*, 1999), est aussi un *légume vert* ; le *haricot vert* rentre aussi bien dans *légume frais*, *légume vert*, que *légumineuse*, tout comme le *petit pois* ; mais le *pois cassé* n'appartient qu'au deux catégories *légumineuse* et *légume sec*.

La représentation hiérarchique arborescente convient bien mal à ce genre de classification, mais elle permet de donner une représentation schématique simplifiée des relations hiérarchiques. En italien on trouve une organisation légèrement différente :



On constate que la symétrie du générique est brisée en italien : *ortaggio* n'entre dans aucun composé. Sa formation est d'origine métonymique, par référence au potager, *orto*. *Verdura fresca* correspond à *légume vert* mais pas exactement à *légume frais*. L'opposition /frais/ *vs* /sec/ n'est définie en italien que pour les légumineuses. Tout se complique si l'on tient compte de la polysémie : *verdura* peut aussi désigner des crudités, à la différence de *légume*. En outre, *légume* peut être pris comme générique désignant la *garniture* ou l'*accompagnement* d'un plat, que l'italien désigne par *contorno*. Mais cette garniture, en français, inclut les pâtes et le riz, à la différence de l'italien.

Quoi que l'on choisisse pour traduire *légume* en italien, il faut donc admettre qu'il n'y aura pas équivalence exacte au point de vue du signifié. D'une manière générale, toutes

les lexies héritant leur signifiés de leur *valeur* au sein de l'ensemble du système, il est rare d'obtenir des signifiés de valeur identique entre des systèmes non isomorphes.

Puisque ces contrastes lexicaux aboutissent presque systématiquement au non-recouvrement des signifiés, l'équivalence lexicale se situe donc ailleurs que dans la congruence des organisations formelles de la signification. En fait, de même qu'avec l'équivalence traductionnelle au sens large, la possibilité de transcoder ne se situe pas directement dans les *significations*, mais au niveau du rapport de *désignation* : « Dans tout acte de traduction (comme dans tout acte de « reformulation » linguistique), on changera de signification alors même que le passage d'un signifié à un autre aura manifesté la possibilité de rendre deux signes équivalents au niveau du désigné » (Pergnier, 1993: 113). Si *légume* et *ortaggio* peuvent être considérés comme équivalents, c'est parce qu'ils ont des points de contact : ils peuvent, dans certains contextes, désigner les mêmes référents. La notion de correspondance, illustrée par les vedettes de dictionnaires multilingues, implique donc des équivalences virtuelles, actualisées dans des contextes généraux, plutôt que des identités dans le découpage des champs sémantiques.

La prise en compte des contextes nous oblige à intégrer les phénomènes de polysémie, et nous contraint de sortir de la perspective strictement structurale, pour aboutir à une sémantique élargie. Comme le note Mounin (1963 : 138), « la sémantique est la partie où la formule de Saussure est fausse, la partie où la langue ne peut pas être envisagée en elle-même, parce que c'est la partie par où l'on passe incessamment de la langue au monde, et du monde à la langue ». Ce va-et-vient entre monde et langue concerne la traduction à deux titres : dans la formation des contenus sémantiques des lexies à l'intérieur d'une langue, et dans la possibilité d'enjamber les systèmes linguistiques pour établir des correspondances. Car si la comparaison est possible, c'est sur la base des contextes extra-linguistiques où se tissent les équivalences, comme le précise Pergnier (1993 : 118) : « Si le signifié était uniquement structural, c'est-à-dire pure forme, rien ne permettrait d'établir quelque rapport que ce soit entre ces deux éléments. Rien n'autoriserait, par exemple, à comparer *tableau* et *picture* plutôt que *tableau* et *brown*. »

I.1.3.2.1.2 Plan de la substance

En constatant que l'espagnol emploie deux lexèmes, *pez* et *pescado*, là où le français n'emploie que le mot *poisson*, on peut être tenté de conclure sur la plus grande finesse de découpage du système espagnol, du moins pour ce qui concerne ces unités. L'exemple de la traduction de *mouton* par (angl.) *sheep* et (angl.) *mutton*, est souvent cité dans la littérature depuis Saussure. Cependant, dans ces deux cas, la métaphore du « filet » plus ou moins resserré risque d'induire en erreur, car en contexte l'espagnol n'apporte pas une information plus précise que le français : quand on va chez le poissonnier, il n'y a pas d'ambiguïté sur ce que *poisson* désigne – et on ne trouvera jamais un *banc de poisson* sur un étalage.

Bien souvent, la variété des traductions possibles d'une même unité n'est autre que la manifestation de sa *polysémie*, c'est-à-dire de sa capacité à désigner *différemment* suivant les contextes où elle est employée. Plus précisément, une unité est polysémique lorsqu'elle s'insère dans différents taxèmes : par exemple *poisson* peut s'insérer dans le taxème //animal aquatique//, au côté de *mollusque* et de *mammifère marin* mais aussi dans le taxème //aliment//, par opposition avec *légume* et *viande*.

La polysémie d'une lexie n'est donc pas directement liée à son degré de généricité dans un même champ sémantique : elle indique seulement que cette unité croise *plusieurs* champs sémantiques.

Examinons différentes traductions de *bois* en italien, en ne retenant que la signification *A* notée par le *Zingarelli* (1998). Pour chaque équivalent proposé, nous avons relevé les indications contextuelles données par le dictionnaire (cf. tableau 2).

A chacune de ces acceptions on peut rapporter un taxème spécifique, que nous étiquetons au moyen d'expressions linguistiques explicites (mais en partie arbitraires, car la description des significations nécessiterait la construction d'un métalangage rigoureux et précis) : //forêt//, //combustible//, //matériau - matière//, //matériau de gros œuvre//, //gravure//, //arbre//. Or il est rare que le contexte d'emploi ne soit pas suffisant pour qu'on puisse décider de quelle acception de *bois* il s'agit.

Indications contextuelles	*Traduction*
se promener au bois	*bosco*
bois de chauffage, poêle à bois, charbon de bois, scier du bois, fendre du bois	*legna*
(technologie) *bois aggloméré, bois amélioré, bois armé, bois contre-plaqué, sciure de bois, bois de papeterie*	*legno*
bois débité, bois de charpente, de construction, bois de marine, bois d'industrie, bois d'œuvre	*legname*
un livre agrémenté de quelques beaux bois	*incisione*
(sylviculture) *bois chablis, bois mort*	*albero*

tableau 2 : *quelques équivalents italiens de* bois

Entre *bois* et ses différentes traductions, on observe deux types de relation : le vague du mot *bois*, par rapport à des nuances très voisines comme *legna*, *legno* ou *legname*, n'est pas de même nature que le vague du mot *bois* par rapport à *bosco* ou *incisione* : dans le premier cas il s'agit de *vague référentiel*, la lexie *bois* couvrant un champ sémantique plus général //bois = matière ligneuse//, que l'italien subdivise en trois taxèmes ; dans le deuxième cas il s'agit d'ambiguïté, car *bois* a une distribution plus large que ses traductions italiennes, et couvre des taxèmes hétérogènes et non compatibles. On peut dire que *bosco* ou *incisione*, à la différence de *bois*, portent en elles une partie de leurs contextes d'emploi, et sont donc plus redondantes par rapport à ceux-ci. On retrouve ainsi la distinction entre *vague* et *ambigu* décrite par Victorri & Fuchs (1996). Pour illustrer cette distinction, Mel'čuk *et al*. (1995 : 60) donnent les exemples suivant :

– le mot *tante* est vague car il « correspond alternativement à plus d'un référent extralinguistique, alors qu'elle-même correspond à une seule lexie ». On peut appliquer par exemple le critère de « cooccurrence compatible » (critère de Green - Apresjan) : la phrase *J'ai vu tante Anna et tante Cécile, la sœur de mon père et celle de ma mère* est parfaitement correcte. De même, on peut employer

simultanément lexie *bois* dans les sens de *legna*, *legno* ou *legname* : *je vends du bois pour le chauffage et la construction*.

– le verbe *peindre* est ambigu car il « correspond alternativement à plus d'une lexie [=soit L1, soit L2 soit à…] ». Cette fois, le critère de cooccurrence compatible ne peut s'appliquer, car les acceptions sont exclusives les unes des autres : sans jeu de mot, on ne peut dire **j'ai peint sa femme et sa chambre*. De même, *j'aime le bois de Boulogne et le bois de hêtre* n'est pas une « phrase normale » au sens de Mel'čuk *et al*.. D'autres critères confirment ces conclusions : la phrase *je vais au bois* est ambiguë car elle répond au critère « d'interprétation multiple », les interprétations « aller se promener dans le bois » et « aller acheter du bois » étant mutuellement exclusives.

La polysémie est une donnée fondamentale du transcodage dans la mesure où chaque langue organise les signifiés dans un jeu de polysémie qui lui est propre. A titre d'illustration, comparons les emplois contrastés de deux lexies polysémiques : *note* en français, et *conto*, un de ses équivalents possibles en italien. Le tableau 3 donne les principales équivalences proposées par le *Zingarelli*.

(*fr.*) *note*	**(*it.*) *conto***
1 *nota*: note manuscrite, *nota manoscritta*; donner la note, dare il la # carnet de notes, *taccuino per appunti* 3 *conto* (m.), *nota*: régler sa note d'hôtel, *saldare il conto dell'albergo* 4 *bolletta*: la note du gaz, du téléphone, *la bolletta del gas*, del telefono 5 *voto* (m.), *votazione*: une bonne note, une mauvaise note, *un bel voto, un brutto voto* 6 (raro) *segno* (m.), *notula*: mettre une note en regard d'une phrase, *fare un segno in calce a una frase* 7 (raro) *tasto* (m.): les notes du piano, *i tasti del pianoforte* 8 †*ignominia, infamia*	1 compte: *fare un conto*, faire un compte; *conto bancario*, compte en banque; *libro dei conti*, livre de comptes; 2 note (f.): *il conto della sarta, dei fornitori*, la note de la couturière, des fournisseurs 3 addition (f.): *chiedere il conto in un ristorante*, demander l'addition dans un restaurant 4 estime (f.): *avere in buon, in gran conto q.*, avoir en grande estime q.; *persona da, di conto*, personne digne d'estime 5 nella loc. *fare conto*, compter: *fare conto su q.*, su qc., compter sur q., sur qc.; 6 nella loc. *fare conto*, supposer, imaginer # *faccia conto di essere a casa sua*!, faites comme chez vous!

tableau 3 : *traductions comparées de* (fr.) note *et* (it.) conto

La comparaison de ces lexies nous montre des emplois se recouvrant partiellement et des emplois divergents : s'agit-il d'un décalage entre deux « filets » sémantiques, dont les maillages non congruents recouvriraient de manière différente une même surface sémantique ? A l'évidence non. Chacune des deux unités peut se ranger dans un éventail de taxèmes, qui sont autant d'espaces sémantiques plus ou moins disjoints. Ainsi la comparaison structurale de *conto* et *note* n'est valide qu'à l'intérieur des champs sémantiques qui leur sont communs : au sein d'un taxème qu'on pourrait nommer //somme due//, on constate que *note* a une signification plus large que *conto*, puisque *conto* s'oppose à *bolletta* (suivant l'opposition /abstrait/ vs /concret/, *bolletta* correspondant à une quittance, un bulletin ou un bordereau). En dehors de ce point de convergence, les polysémies des deux lexies ne sont pas parallèles, et les acceptions voisines s'« éloignent » dans des directions différentes dans chacune des langues. Comme le note Pergnier (1993 : 84) :

> « La confrontation de signes appartenant à deux langues différentes révèle à la fois la polysémie de chacun (c'est-à-dire la diversité interne de leurs signifiés considérés du point de vue des concepts désignés) et la non-coïncidence de ces signifiés, c'est-à-dire le fait qu'ils sont polysémiques *différemment*. »

Ces divergences expliquent la variabilité des équivalents traductionnels, et la sensibilité au contexte, dans la mesure où c'est le contexte qui détermine le taxème. Si l'on résume les équivalents donnés pour *note* et *conto* on a :

fr. : *note* → it.: *nota*, *appunti*, *conto*, *bolletta*, *voto*, *votazione*, *segno*, *notula*, *tasto*

it.: *conto* → fr.: *compte*, *note*, *addition*, *estime*

Mais à travers ces variations polysémiques se dégage pourtant, pour chaque lexie, une certaine *signification*. Pour caractériser cette signification, nous éludons ici la question de la frontière entre polysème et homonyme, l'évolution diachronique des signifiés montrant qu'il existe un continuum de phénomène entre ces deux pôles : nous considérons qu'il y a polysémie lorsque le locuteur peut percevoir un fil conducteur, reliant les différentes acceptions par des mécanismes de motivation, métaphoriques ou métonymiques. De même, nous laissons de côté les hypothèses d'*un* noyau sémique commun, qui constituerait la base du signifié de la lexie, ou d'*un réseau de* sous-noyaux

reliés deux à deux par une « ressemblance de famille » (Wittgenstein, *Investigations*, 1961 ; Kleiber, 1999 : 64).

Ce qui nous intéresse, d'un point de vue contrastif, c'est que chaque unité renferme un *contenu* positif cimenté par une force de cohésion interne, contenu que nous appelons *signification*. Or, on ne peut comparer des lexies comme *note* et *conto* que si l'on articule la description du contenu sur deux niveaux :

- au niveau de la forme linguistique, par les liens sémantiques entretenus avec des unités voisines : par exemple *note* est clairement relié au verbe *noter* et *conto* au verbe *contare*. Ce type de dérivation régulière, implique une co-détermination sémantique : *une note* ↔ *ce qu'on note*, *un conto* ↔ *quel che si conta*. Le contenu sémantique peut donc être caractérisé formellement par des aspects structuraux liés aux distributions, aux paradigmes dérivationnels, aux commutations possibles, etc.

- au niveau de la substance, relative aux concepts et aux référents désignés par les unités linguistiques. En comparant tous les objets conventionnellement désignés par une même lexie (c'est-à-dire dans les usages les plus fréquents), on peut aisément abstraire un ou plusieurs dénominateurs communs : d'un point de vue général, le mot *note* désigne ce qui est consigné, écrit, enregistré. Le mot italien *conto* a trait à ce qui est additionné ou soustrait : le *conto*, dans les emplois cités, correspond toujours à un bilan résultant d'un calcul où interviennent des plus et des moins.

Pour comprendre le rapport entre ces deux modes de description, structural et référentiel, il faut mettre de côté les tentations normatives, en abandonnant par exemple le principe fregéen selon lequel le sens détermine la référence, et sortir du réductionnisme découlant de certains partis pris méthodologiques, comme le béhaviorisme de Bloomfield assimilant le sens au contexte référentiel.

D'une part, il faut bien admettre que toute unité véhicule une signification conventionnelle, stable, reconnue par les locuteurs, qui préexiste à son usage dans une situation particulière. Même si le sens est *construit*, il se sert des éléments fournis par le

code partagé. Comme le remarque Kleiber (1999 : 36), « non seulement la construction dynamique du sens d'un énoncé n'est pas incompatible avec le fait qu'elle s'effectue des éléments de sens stables ou conventionnels, mais bien plus encore elle l'exige : sans sens conventionnel ou stable, il n'est guère de construction sémantique possible. »

D'autre part, cette signification préexistante n'est pas vierge de tout influence extralinguistique, et il n'y a pas de signification « immaculée », indépendante des objets auxquels elle réfère. Une vision purement structurale de la signification conduit à une impasse, puisqu'elle empêche de rendre compte de phénomènes sémantiques aussi importants que motivation et polysémie : « plus on postule un sens abstrait, détaché ou dégagé d'une gangue référentielle jugée étouffante et opacifiante, et plus on éprouve de la peine à faire le raccord entre le sens non descriptif et le sens abstrait postulé et le référent finalement désigné. » (Kleiber, 1999 : 45)

Pour nous, la signification est donc un contenu positif constitué d'un ou plusieurs noyaux sémantiques abstraits, émergeant du rapport conventionnel entre des usages linguistiques et des classes de référents. Sa description n'est envisageable que *dans* et *par* l'interaction du structural et du référentiel. Ce rapport de rétroaction peut être schématisé ainsi : à tout moment, la signification d'une lexie rend possible la désignation motivée d'une nouvelle classe de référent non encore nommée ou nommée différemment, par des processus aussi généraux que l'explication, la métaphore, la métonymie ; réciproquement, cette nouvelle classe référentielle enrichit la signification de nouveaux traits potentiels, qui deviennent susceptibles de « remonter » dans le noyau de la signification. C'est ainsi que le mot *cadeau*, qui signifiait « lettre capitale avec enjolivure » à d'abord été étendu, de façon métaphorique, à des « divertissements offerts à une dame »[84] ; cette nouvelle désignation a progressivement modifié la signification globale du mot, et le trait /enjolivure/ a finit par s'estomper devant le trait /présent/ devenu prépondérant dans la signification moderne.

Après avoir séparé, et articulé, ces deux aspects de la *désignation* et de la *signification*, ont peut désormais mettre en lumière la notion d'équivalence dans le transcodage des lexies. Dans l'exemple des deux unités *note* et *conto*, la tension entre divergence et convergence se résout sur chacun des deux niveaux :

– leurs *significations* sont différentes : à la signification de *note* on peut rapprocher les concepts « écrire », « enregistrer » tandis que *compte* implique les opérations de « soustraire », « additionner », « calculer ».

– leurs *designata* coïncident partiellement : les notes de la teinturière et du restaurateur sont à l'intersection des designata de *note* et de *conto*.

La différence de signification explique que les deux unités soient polysémiques différemment ; et les divergences dans leurs polysémies permettent que leurs désignations puissent se croiser. La figure 9 représente schématiquement ce type de configuration :

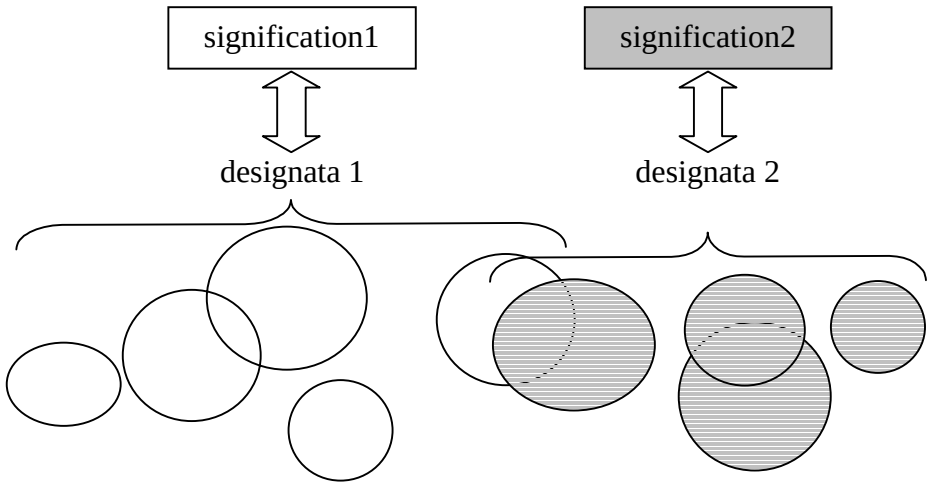


figure 9 : *différence des significations vs coïncidence des désignations*

Ainsi, dans la polysémie de leur lexique, les langues élaborent des systèmes de classification qui regroupent des *designata* autour de contenus positifs, ce que Pergnier nomme des *ressemblances* :

> « Au niveau de son organisation interne, on pourrait dire que le signifié saisit les choses qu'il désigne non par leurs différences, mais par leurs ressemblances. Le passage de l'anglais au français n'a pas seulement pour effet de changer le signifiant ; il a pour effet de le faire changer de système classificateur. » [85]

On peut parler dans ce cas de classification *analogique*, par opposition aux classifications différentielles issues des modèles hiérarchiques d'inspiration

[84] Larousse, *dictionnaire étymologique*, 1997

aristotélicienne. Ces derniers modèles, parfois appelés modèles des conditions nécessaires et suffisantes lorsqu'ils explicitent les définitions (Kleiber, 1990), sont caractéristiques des usages scientifiques de la langue, qui débouchent sur l'établissement des nomenclatures et des terminologies. Ils conviennent assez bien aux ontologies locales, limitées à un domaine, et sont compatibles avec le schéma de Nida représentant des « filets sémantiques » plus ou moins resserrés, correspondant à différents niveaux de généricité. S'il n'y avait que ce type de classification, il serait possible de se baser sur un simple système de conversion d'unité à unité pour trouver des équivalents sémantiques satisfaisants, moyennant quelques variations vis-à-vis du degré de spécificité.

Mais le phénomène de la polysémie fait éclater ce type d'organisation hiérarchique, en recyclant les mêmes unités à travers des champs sémantiques très différents. Les classifications s'organisent autour de significations abstraites, enjambant les domaines par le mouvement imprévisible des tropes, métaphores et métonymies :

> « En fait, la conception que l'on peut se faire de la polysémie change radicalement dès lors qu'on prend en considération le fait que le signe n'est pas un simple substitut symbolique d'une fraction d'univers (de même que le phonème n'est pas un simple cadre délimitant une « zone » de sons), mais qu'il est une rubrique à l'intérieur d'un système classificateur. » (Pergnier, 1993 : 112)

On passe ainsi d'une conception quelque peu passive de la sémantique, qui présentait l'organisation des significations comme une opération de quadrillage sur des portions inertes de la réalité, à une vision plus dynamique de l'économie du sens. Dénommer ce n'est plus seulement étiqueter, différencier, mais c'est aussi rapprocher, abstraire, synthétiser, établir des liens analogiques : échafauder, déjà, une conceptualisation du monde. C'est ce que Pergnier appelle la « fonction cognitive » du langage (1993 : 66) :

> « Concevoir le signifié comme un fragment de la réalité globale, perceptible ou intelligible, fragment dont les frontières sont déterminées par les autres signifiés de la langue, c'est admettre que le réel ainsi désigné est saisissable indépendamment du signe qui le signifie, et qu'en changeant de système linguistique on change seulement de frontières. Ce serait aussi nier la fonction cognitive du langage. Or pas plus que le signifiant qu'étudie la phonologie, le signifié n'est une donnée immédiate. »

[85] Pergnier, ibid., p. 109

Dans cet effort de réduction du monde, chaque langue réorganise les objets suivant sa libre fantaisie. Du coup les classifications analogiques interdisent le transcodage de lexie à lexie, car les unités ne désignent plus directement, mais doivent transiter par deux niveaux de codification : 1) une signification abstraite (par exemple une sélection de traits référentiels) et 2) des classes conventionnelles de *designata* qui sont susceptibles de rentrer sous la rubrique de la lexie. Or dans une perspective référentielle de la sémantique, il faut bien admettre que le niveau 2) fait aussi partie de la langue : c'est un ensemble de conventions partagées par les locuteurs, qu'on peut mettre en évidence par des études de type sociologique. C'est ce qui nous fait appeler un album de musique un *disque* et non un *enregistrement*, alors qu'en Grande-Bretagne on emploiera le mot *record* et non *disc*. Là où le français dénomme l'objet par référence à son apparence circulaire, l'anglais choisit de le désigner par sa fonction (*to record = conserver*). La *désignation* fait donc partie du code, au même titre que la *signification*.

D'ordinaire, les locuteurs ne perçoivent pas, dans l'usage de leur propre langue, la contingence du rapport qui lie signification et désignation : à chaque fois que ce rapport est motivé, il apparaît subjectivement comme nécessaire. Il faut véritablement sortir de son système et aborder les dénominations dans une autre langue pour réaliser ce que la motivation implique d'arbitraire, c'est-à-dire de contingent, comme le montrent les exemples donnés par Marianne Lederer (citée par Laplace, 1995 : 195) :

> « Nul ne contestera à première vue que (angl.) *height* signifie *hauteur* en français, mais chacun constatera que pour désigner la notion que l'anglais nomme *depth* (of a tank), nous disons en français *hauteur* (d'une cuve) […] Pour *offshore drilling*, l'expression française est *forage en mer* ; le forage s'effectuant à la fois loin des côtes en mer, la dénomination, différente dans chacune des langues, est dans chacune elliptique. L'anglais dit *outlet*, le français dit *prise* (de courant) […] »

La leçon de ces quelques exemples est d'importance : d'une part le rapport entre désignation et signification est largement arbitraire, d'autre part les équivalences sont établies au niveau du *designatum*, et non du signifié.

Ceci appelle une véritable redéfinition du transcodage : le transcodage d'une lexie ou d'une phrase consiste à rechercher, non pas l'équivalence au niveau des *significations*, qui sont des contenus abstraits débouchant sur un faisceau d'interprétations potentielles, mais au niveau des *désignations*, c'est-à-dire des réalités extralinguistiques qui sont dénotées par

les unités linguistiques au sein d'un contexte déterminé. Si plusieurs désignations mutuellement exclusives sont envisageables, nous dirons que l'expression est ambiguë car le contexte est insuffisamment spécifié.

Parfois le co-texte suffit à lever toute forme d'ambiguïté : dans la phrase *as-tu écouté ce disque ?* on peut supposer, avec une très forte probabilité, qu'il ne s'agit pas du frein d'une voiture : on traduira sans hésiter par *record* plutôt que par *disc*. Cela nous permet-il de caractériser linguistiquement ce que nous avons jusqu'à présent appelé contexte ? En d'autres termes, le contexte peut-il être réductible au co-texte, dans le filtrage des acceptions ? Non, car rien ne nous empêche d'imaginer des co-textes ambigus dans des contextes qui sont suffisamment spécifiés : la phrase « je voudrais changer de disques » s'interprète différemment chez le mécanicien ou dans une médiathèque.

Nous pensons que le co-texte est seulement source d'*indices* contextuels. Ces indices donnent parfois des informations très sûres pour différencier les acceptions : comme le note Véronis (2000b), il est possible, dans une certaine mesure, de déterminer empiriquement la structure des définitions des unités lexicales dans un dictionnaire, en se basant sur l'observation de leurs « classes d'usage » à travers des corpus.

Prenant l'exemple du mot *barrage*, il montre que les informations syntaxiques et collocationnelles permettent, en grande partie, d'en différencier les acceptions : les trois constructions *barrage de X par Y*, *barrage sur X* et *barrage à X*, suffisent à distinguer les trois principales acceptions : 1. « l'action de barrer », 2.1 « l'ouvrage » 2.2 « l'obstacle », dans un sens abstrait. Par ailleurs Véronis constate, en observant les collocations, que les prédicats s'appliquant à *barrage* selon l'acception 2.1 forment deux ensembles disjoints : *construire, édifier, démolir,* etc., pour le barrage sur un cours d'eau et *dresser, franchir, démanteler,* etc., pour le barrage sur une route. Ce qui lui permet de subdiviser 2.1 en deux sous-acceptions.

Mais dans cette dernière opération, l'organisation des données co-textuelles découle de l'identification préalable du contexte référentiel : il se pourrait très bien que le *barrage* qu'on *dresse* et le *barrage* qu'on *franchit* ne soient pas les mêmes. Seule une connaissance référentielle permet de les assimiler. En théorie, une même acception peut très bien correspondre à des distributions différentes, et des acceptions différentes peuvent avoir des distributions identiques. Certes, le co-texte est le meilleur guide du linguiste, car il est

facilement objectivable : mais il n'y a pas de raison de supposer *a priori* que ce soit le seul, et que les informations co-textuelles suffisent toujours à déterminer l'interprétation du sens.

Nous pensons donc que le contexte n'est pas toujours spécifié linguistiquement : le transcodage lexical n'est donc possible que moyennant l'adjonction d'informations contextuelles extérieures aux systèmes linguistiques.

I.1.3.2.2 Aspects syntagmatiques

L'étude de la polysémie nous révèle une propriété majeure du langage : les unités lexicales ne signifient pas de manière isolée, mais par leur insertion au sein d'un contexte. Dès lors, le statut d'indépendance de l'unité se brouille et le co-texte syntagmatique devient un élément important de sa mise en œuvre.

Ce qui se joue à ce niveau, c'est la question de la définition de l'unité lexicale elle-même : lorsqu'une lexie entretient des rapports privilégiés avec d'autres lexies, sur les plans syntaxiques et sémantiques, assiste-t-on à l'émergence d'une autre unité, une unité *polylexicale* pour reprendre le terme de Gaston Gross, dont le degré d'autonomie serait plus affirmé ?

Cette question est importante d'un point de vue traductionnel : pour traduire correctement *pomme de terre* ou *pomme de pin* en anglais, il faut évidemment repérer le mot composé comme une unité autonome dont les éléments ont perdu leur indépendance sémantique et syntaxique.

Cependant, le problème de la définition des unités polylexicales peut donner lieu à des solutions très variables, et il est difficile de trouver des critères linguistiques permettant de trancher sans équivoque :

- D'une part, les critères purement formels ne sont pas suffisants : l'intégration des signifiants graphiques ou phoniques, le figement interne (l'impossibilité d'insérer d'autres éléments), la variabilité, présentent des configurations très différentes, comme dans les exemples ci-dessous :

Phrasèmes figés (sans insertion possible) dont certaines unités n'ont pas d'existence autonome: *au fur et à mesure, à vau-l'eau, à la queue leu leu, à potron minet, tandis que*, etc.

Phrasèmes figés soudés au niveau graphique : *portemanteau, portefeuille*, etc.

Phrasèmes figés invariables : *turlututu chapeau pointu, or noir, entre chien et* loup, etc.

Phrases figées invariables (sauf ellipse) : *quand le chat n'est pas là, les souris dansent*, etc.

Phrasèmes figés invariables avec tiret : *c'est-à-dire, monte-en-l'air, coupe-papier, porte-clefs, porte-monnaie, dessus-de-lit, sot-l'y-laisse*, etc.

Phrasèmes figés avec variation graphique interne : *des pommes de terre*, des *sacs à main*, etc.

Phrasèmes figés avec tiret et variation graphique interne : *œils-de-bœuf, œils-de-chat, dure-mère*, etc.

Phrasèmes figés avec variation graphique et phonique interne : *des chevaux-vapeur, une bonne à rien*, etc.
Phrasèmes figés autorisant des insertions et incluant un paradigme flexionnel : *avoir maille à partir avec, prendre la poudre d'escampette, en finir avec*, etc.

– D'autre part, la prise en compte des relations syntactico-sémantiques fournit des critères précieux, mais qui présentent là encore une grande variété de cas de figure. Examinons quelques-uns de ces critères.

I.1.3.2.2.1 Le figement syntactico-sémantique

On peut définir les unités polylexicales sur la base du *figement* des propriétés combinatoires des unités qui les constituent. A l'instar de G. Gross (1996 : 6), nous caractérisons le figement par opposition à la notion de liberté combinatoire :

> « Nous définissons un groupe (syntagme ou phrase) *libre* comme une séquence générée par les règles combinatoires mettant en jeu à la fois des propriétés syntaxiques et sémantiques, comme, par exemple, les relations existant entre les prédicats et leurs arguments. »

Le figement d'une expression polylexicale se manifeste donc par le gel des principes combinatoires, au plan syntaxique comme au plan sémantique, qui règlent la construction

libre du sens à partir des unités lexicales et du système grammatical. Gross insiste à plusieurs reprise sur l'aspect scalaire du phénomène. Pour différencier les unités polylexicales figées des constructions libres, Gross (1996 : 27) propose de décrire dans le détail l'ensemble des traits liés au phénomène du figement. Par exemple il suggère de « lister les noms composés sur la base de leur structure de surface et décrire avec précision l'ensemble des paramètres qui séparent les suites figées des suites libres, et montrer par là que le figement est un phénomène scalaire ».

L'auteur relève un certain nombre de traits, dont nous avons sélectionné les plus remarquables :

– L'opacité sémantique, qui elle-même « est un phénomène scalaire : elle peut être totale (*la clé des champs*), partielle (*clé anglaise*) ou inexistante (*clé neuve*). » (1996 : 11)

– Le « blocage des propriétés transformationnelles », délimitant le champ des paraphrases possibles d'un énoncé. Par exemple, pour une construction verbale libre, telle que « l'enfant a lu ce livre » on peut opérer un certain nombre de transformations syntaxiques « normales », n'altérant pas le noyau sémantique de l'assertion (1996 : 12) :

  passivation : Ce livre a été lu par l'enfant
  pronominalisation : *L'enfant l'a lu*
  détachement : *Ce livre, l'enfant l'a lu*
  extraction : *C'est ce livre que l'enfant a lu*
  relativation : *Le livre que l'enfant a lu*

  Ces transformations ne sont plus possibles, sans jeu de mots, avec une expression telle que *l'enfant a pris la mouche*. Pour les constructions nominales, adjectivales, adverbiales, on assiste au même phénomène : des transformations paraphrastiques sont inhibées lorsque le composé est figé (p. ex. l'expression *vache folle* ne peut aboutir, dans un usage normal, à * *la folie des vaches*).

– La « non-actualisation des éléments » qui composent l'expression figée, i.e. la perte de « ce qui les lie à une situation donnée » (1996 : 144). Cette non-

actualisation se manifeste par l'impossibilité de déterminer les éléments séparément, dans la mesure où ils n'ont plus d'autonomie sémantique : dans l'expression *prendre une veste*, le mot *veste* ne réfère à aucun vêtement et n'hérite pas de ses déterminations usuelles – on ne dit pas **prendre une veste en tweed*. Gross appelle *locution* les suites à éléments non-actualisés, et désigne par là un premier degré de figement, concernant par exemple des expressions semi-figées dont le sens est compositionnel et désignant un concept précis, comme *cordon électrique*. D'après lui, la création terminologique se situe fréquemment, à ce niveau.

- L'absence de « prédicativité ». Les relations normales de prédications sont inhibées : en usant d'une expression figée, on n'affirme rien, on utilise une séquence « préconstruite […] qui fait partie du stock lexical au même titre que les noms simples. » (1996 : 33) Par exemple dans le cas d'adjectif typologisant, dans la suite *Sub. + Adj.*, il faut considérer les adjectifs « comme des étiquettes et non des qualités » (1996 : 51), comme dans *bolet jaune, casque intégral, accent circonflexe*. N'étant pas des prédicats, ces adjectifs ne peuvent faire l'objet d'une nominalisation ni d'une quantification : pour une *arme blanche* on ne parlera pas de ** la blancheur de cette arme, * une arme très blanche* (1996 : 51).

- Le « blocage des paradigmes synonymiques ». Gross donne les exemples suivants, dont l'effet comique est un indice de « défigement » artificiel : ** une poire de terre, * briser sa pipe, * avoir des graines à moudre, * une clé britannique, * une caisse noire, * à fin de force, * sourd comme un vase, * un autorail de mesure*…

- L'« impossibilité d'insertion ni d'adjonction », qui interdit par exemple les constructions suivantes : ** une arme blanche et lourde, * un col très vert, * une pomme bonne de terre, * il tourne de l'œil gauche*. De même, l'effacement d'un élément n'est pas toujours envisageable : *un cordon-bleu* n'est jamais *un cordon* tout court…

- La structure syntaxique atypique par rapport à la partie du discours (absence de translatif). Ce sont les expressions figées les plus facilement repérables, dans la

mesure où elles ne sont pas construites selon les règles syntaxiques normales, comme un *en avant*.

Notons que ces traits sont étroitement entrelacés (et par conséquent quelque peu redondants) dans la mesure où ils reflètent tous la même propriété, sous différents angles : la perte d'autonomie syntactico-sémantique d'une ou plusieurs unités. Ils présentent en outre l'intérêt de ne pas séparer les phénomènes syntaxiques de l'arrière plan sémantique, qui est ici prééminent.

Ces propriétés, telles que l'aspect scalaire du figement, la multiplicité et la variabilité des traits qui le caractérisent, se prêtent très naturellement à une caractérisation prototypique (E. Rosch *et al*., 1976). La classe des expressions figées a en effet tous les traits des classes définies par la théorie des prototypes :

– On ne peut la définir à partir d'un noyau de traits nécessaires et suffisants, mais plutôt par la convergence de propriétés variables et hétérogènes, pouvant mêler des considérations syntaxiques, sémantiques, pragmatiques ou culturelles. J. Charteris Black (1999), proposant une définition prototypique des unités phraséologiques, énumère certains de ces traits :

> « Une unité prototypique au centre du système phraséologique est une formule opaque et non-compositionnelle. En d'autres termes, elle se comporte comme un mot au niveau syntaxique ; elle est fixée et institutionnalisée au niveau lexical ; d'un point de vue conceptuel, elle est imagée, connotée et véhicule une signification ancrée dans une culture spécifique. »[86]

– Ces propriétés ne sont pas exclusives les unes des autres, et se recouvrent partiellement. En outre, elles n'ont pas toutes le même statut, certaines étant plus représentatives que d'autres : par exemple l'opacité sémantique est une propriété plus saillante que l'impossibilité d'insertion.

[86] "A prototypical unit at the centre of the phraseological system is opaque, formulaic and non-compositional. In linguistic terms, it is word like, syntactically and lexically fixed, and institutionalised ; in conceptual terms, it is figurative, evaluative and has culture-specific meaning."

- L'appartenance à la catégorie des unités polylexicales connaît des degrés divers : certains éléments sont centraux, prototypiques, et cumulent un grand nombre des traits catégoriels (p. ex. l'expression *à la queue leu leu*), et d'autres sont périphériques et leur appartenance à la classe est sujette à caution (p. ex. la locution *porte d'entrée*, citée par Gross, dont le figement est faible).

- Corollaire du point précédent, les frontières de la classe sont floues : il n'est pas possible de tracer une démarcation nette entre figement et liberté.

Sur le plan contrastif, la question de l'identification des unités polylexicales est fondamentale : un seul mot peut très bien correspondre, du point de vue de l'équivalence sémantique (et donc, en contexte, de la désignation) à une phrase entière, et réciproquement. Pour le transcodage, l'identification des unités de sens est essentielle pour éviter les interférences aboutissant à des constructions incompréhensibles. Pour analyser correctement l'équivalence suivante :

fr. : *je t'en veux*
it. : *ce l'ho con te*

il faut d'abord identifier les deux unités polylexicales :

fr. : *en vouloir à …*
it. : *avercela con …* (littéralement : « l'avoir avec quelqu'un »)

Nous verrons plus loin, dans la définition des *unités de traduction*, qu'il est important de déceler les phénomènes de figement syntactico-sémantique afin de différencier ces unités des combinaisons libres. Les critères énumérés nous seront précieux dans la mise en œuvre de l'extraction de correspondances lexicales : nous constaterons néanmoins que pour tenir compte des contrastes idiomatiques, ils ne sont pas suffisants.

I.1.3.2.2.2 Régime de la lexie

La définition des unités lexicales concernait les relations syntagmatiques internes à l'unité lexicale. Le régime de la lexie, au sens de Mel'čuk *et al*. (1995), concerne l'aspect externe de ses propriétés distributionnelles. Comme le suggèrent ces auteurs (1995 : 17),

on constate qu'une partie de la combinatoire des lexies doit être définie au niveau lexical : « Les instructions d'assemblage sont obligatoirement écrites en fonction des pièces à assembler. »

Ainsi, la *valence* est une propriété fondamentale de la lexie, aussi importante que son contenu sémantique, puisqu'elle en détermine la correcte « mise en œuvre ». Tesnière (1959 : 287) insiste avec force sur ce point : « un verbe dont on connaît le sens, mais dont on ignore la structure actancielle est *inutilisable*. » Ainsi, la traduction, ou le transcodage, ne peut se limiter à un simple transfert lexical : les unités transférées transportent avec elles leurs propriétés structurales, et des unités sémantiquement équivalentes ne véhiculent pas nécessairement les mêmes structures actancielles. Changer les composants implique bien souvent de modifier tout le « câblage » permettant de relier les unités entre elles, transformation « délicate » que Tesnière (1959 : 283) appelle *métataxe* :

> « Il s'agit donc ici de traductions particulièrement délicates puisqu'il y a lieu, non seulement de procéder à l'opération en quelque sorte *mécanique* qui consiste à remplacer un mot par un autre mot, mais de substituer une structure différente à celle qui se trouve dans la phrase à traduire et par conséquent de *repenser* cette phrase dans la langue dans laquelle il s'agit de la traduire. »

Nous empruntons à Tesnière quelques-uns des exemples suivants, où la métataxe est commandée par le choix du noyau verbal :

– Interversion prime actant / second actant :

all. : *I like Mary*
esp. : *me gusta Maria*.

– Interversion tiers actant / second actant :

all. : *sein Knecht half him*
fr. : *son valet l'aida*.

– Permutation second actant / tiers actant :

latin : *doceo pueros grammaticam*
fr. : *j'enseigne la grammaire aux enfants*.

– Interversion circonstant / tiers actant

all. : *etwas von jemanden kaufen*
fr. : *acheter quelque chose à quelqu'un*

On peut considérer certaines prépositions comme faisant partie du verbe, dans la mesure où elles sont totalement intégrées à sa structure actancielle et y perdent leur autonomie :

fr. : *substituer X à Y*
it. : *sostituire X con Y*
angl. : *to substitute X for Y, to replace Y by X*

Cette intégration de la préposition dans la structure verbale explique le statut intermédiaire des verbes transitifs indirects, où la préposition ne joue plus de rôle d'introducteur de circonstant :

angl. : *I will attend this conference*
fr. : *j'assisterai à cette conférence*

Notons que la métataxe n'est pas restreinte au nucleus verbal. Par exemple dans les structures :

fr. : *la transformation de X en Y*
angl. : *the transformation from X into Y*

à un niveau superficiel, l'actant *X* devient circonstant : *from X*

En général, l'interprétation des unités est largement tributaire des structures argumentales. C'est ce que montre l'exemple de Véronis (2000b) déjà cité : aux trois constructions *barrage de X par Y*, *barrage sur X* et *barrage à X* correspondent trois acceptions différentes du mot *barrage*. Des structures aussi fréquentes que SN + *en bois*, SN + *de bois* permettent d'éliminer, avec une forte probabilité, les acceptions « bois = forêt », « bois = gravure ».

Le régime spécifique à chaque lexie concerne donc à la fois ses relations syntaxiques et son interprétation sémantique : il est évident que ce type d'information est incontournable pour mettre en œuvre tout type de traduction.

I.1.3.3 Contrastes grammaticaux

Il apparaît que la traduction, même réduite au simple transcodage, c'est-à-dire aux seules transformations linguistiques, abstraction faite d'un ancrage situationnel particulier, ne peut se limiter à des systèmes de transformation séparant lexique et grammaire : des règles spécifiques, inscrites au niveau lexical, conditionnent une bonne partie des relations structurales internes à la phrase.

Il est néanmoins possible de comparer les systèmes grammaticaux à un niveau général, non pas pour en tirer directement des règles de transformation entre structures de phrase, mais pour évaluer les problèmes posés par des systèmes de contraintes hétéromorphes.

Comme le notait Jakobson, les langues diffèrent surtout par ce qu'elles « doivent » exprimer. En premier chef, ce « devoir » est de nature grammaticale : la grammaire, par la sélection de traits catégoriels qu'elle opère, impose un certain « découpage » de la réalité. Or chaque langue distribue ces traits de manière originale, comme le signale Tesnière (1959 : 284) : « Toute langue établit entre les catégories de la pensée et les catégories grammaticales qui les expriment, certaines correspondances qui lui sont propres. De telle sorte que, pour exprimer tel ou tel genre de notion , elle fera plus volontiers appel à telle catégorie grammaticale qu'à telle autre. »

Chaque langue a en effet sa manière de *grammaticaliser*, i.e. d'inscrire dans un système de contraintes les contenus sémantiques plus ou moins abstraits tels que actance, genre, nombre, personne, temps, aspects, mode, modalités, etc. Sur la base des significations plus ou moins équivalentes attachées aux catégories grammaticales, on peut parfois établir des correspondances très générales, du type :

- dans une conditionnelle introduite par (it.) *se*, le subjonctif imparfait italien correspond à l'imparfait de l'indicatif en français.

- à la personne du *on* indéfini français, correspond la forme du pronom réfléchi italien avec diathèse passive :

fr. : *on parlait beaucoup de cette affaire*
it. : *si parlava molto di questa faccenda*

– la construction espagnole *estare* + gérondif, exprimant l'aspect progressif, n'a pas de correspondant grammaticalisé en français, où l'on utilise plutôt une circonlocution du type *être en train de*...

– en bats, langue du Caucase, les cas ergatif et nominatif permettent de distinguer une action volontaire d'une action involontaire, ce que le français ne fait pas (Pottier, 1992a :171) :

bats : *so v-ože* (*so* : *je* nominatif)
fr. : *je suis tombé* (involontairement)

bats : *as v-ože* (*as* : *je* ergatif)
fr. : *je suis tombé* (par ma faute)

Comme l'illustre l'exemple de Jakobson (relatif au genre du mot *worker*), les découpages catégoriels sont rarement congruents, même lorsqu'il existe des équivalences. On aboutit alors à la nécessité d'opérer des choix, comme dans les exemples ci-dessous :

– Genre

fr. : *Je le lui ai décrit*
angl. : *I described it to him* / *I described it to her* / *I described him to her*

– Nombre

fr. : *Il*(*s*) *mange*(*nt*) (à l'oral)
angl. : *He eat* / *They eat*

– Temps verbal

angl. : *I took my car and went to see him*
fr. : *Je* (*pris* / *ai pris*) *ma voiture et* (*j'allai* / *je suis allé*) *le voir*

– Modalité

fr. : *Il se peut qu'il soit allé au cinéma*
angl. : *He* (*may* / *might*) *have gone to the picture*

fr. : *Il* (*se peut qu'il prenne* / *a l'autorisation de prendre*) *le bus*
angl. : *He may take the bus*

A un niveau plus général, le problème de l'équivalence se pose au niveau des classes qui structurent le lexique en fonction de leur comportement morphosyntaxique : les parties du discours. Les quelques règles transformationnelles données en exemple se fondent implicitement sur la possibilité de transposer des catégories générales telles que verbe, pronom, etc. Quand on examine la plupart des langues européennes, on retrouve une répartition commune, correspondant peu ou prou aux huit catégories déjà identifiées par Aristarque (1ère moitié du IIe siècle av. J.-C.) et Denys de Thrace (~170 - ~90) : article, nom, pronom, verbe, participe, adverbe, préposition, conjonction.

Tesnière (1959 : 63), se limitant aux catégories les plus générales, énumère quatre classes fondamentales pour les mots pleins : « Les quatre espèces de mots pleins sont donc en définitive le substantif, l'adjectif, le verbe, l'adverbe. Ces quatre éléments sont les pierres angulaires du discours. » Mais comme le signale Alain Lemaréchal (1989 : 61) cette vision, qui convient assez bien aux langues européennes, ne peut être généralisée *a priori*, sous peine de verser dans une forme d'ethnocentrisme. Une langue comme le palau (du sud-ouest de la Micronésie), rentre mal dans les cadres établis pour les langues d'origine indo-européenne :

> « Comment rendre compte, par exemple des déterminants du nom dans les langues ou les équivalents de nos adjectifs (…), constituent une sous-classe de verbes – les verbes statifs – qui, comme les autres verbes, ne fournissent de déterminations épithétiques que par le biais d'une relativisation ? »

En outre la distinction entre *mots pleins* et *mots outils* n'a rien d'étanche. Lemaréchal (1989 : 61) note que « dans certaines langues (…) les relateurs, ou au moins une partie d'entre eux, sont d'anciens noms ou sont même homonymes de noms fonctionnant encore comme noms (…) ». Dans le cas du chinois, les prépositions font partie de la catégorie verbale. Pour cette langue, Lemaréchal (1989 : 94) distingue trois sous-classes décrivant un continuum, des verbes ne connaissant qu'un emploi prépositionnel aux verbes purs, en passant par les verbes-prépositions susceptibles de remplir les deux fonctions (1989 : 91) :

	Préposition	*Verbe*
(chinois) : *gēn*	avec	suivre
(chinois) : *xiàng*	comme	ressembler
(chinois) : *gěi*	à	donner

D'après Lemaréchal (1989 : 88) ce type d'inventaire en trois sous-catégories est très répandu pour les langues à verbes-prépositions (ou postpositions), à noms-prépositions (comme le vai en Sierra Leone et au Liberia) ou à adverbes-prépositions (comme le français avec *après*, *avant*, *devant*, *derrière*, *avec* etc. et les usages anciens ou dialectaux de *dessus*, *dedans*, *dessous*, etc.).

Ainsi, la non-congruence des parties du discours compromet la possibilité même de les comparer. Sur quoi fonder le principe analogique permettant d'établir des équivalences entre deux systèmes ? A l'intérieur d'un système les parties du discours sont définies par des critères concomitants : les propriétés distributionnelles, les caractéristiques morphologiques et les fonctions syntaxiques qu'elles peuvent remplir. Dans la perspective traductionnelle, ce sont les aspects concernant la « sémantique de la syntaxe », selon l'expression de Claude Hagège, qui peuvent fournir un terrain à la notion d'équivalence : ce sont donc les *fonctions* syntaxiques qui fourniront un principe d'analogie.

Bien entendu, les fonctions jouent à un niveau très abstrait. Dans son analyse des parties du discours du tagalog[87], Lemaréchal (1989 : 39) s'appuie sur des « fonctions fondamentales », pour dégager quatre grandes classes, ou « superparties du discours » à l'intérieur desquelles se rangent, par analogie fonctionnelle, les parties du discours au sens classique :

- superpartie « Qualificatif » à fonction de prédicat : nom, adjectif, verbe
- superpartie « Circonstanciel » à fonction d'adverbe : adverbe
- superpartie « Substantif » à fonction d'actant : démonstratif, pronom personnel
- superpartie « Nom personnel » à fonction de vocatif : nom propre

Cette analyse s'appuie sur des observations syntaxiques, et notamment sur la distribution des *translatifs*, relateurs dont le rôle est de transformer, selon Tesnière (1959 : 365), « n'importe quelle espèce de mots en n'importe quelle autre ». Le concept de

[87] langue du groupe malayo-polynésien parlée aux Philippines.

translation est central dans l'étude des parties du discours et Lemaréchal, partant de la définition de Tesnière, en donne une version purement syntaxique centrée sur la présence ou l'absence de translatifs spécifiques, permettant à une partie du discours d'exercer une fonction donnée. Par cette redéfinition « la translation fournit des fondements solides pour l'étude : 1) des rapports entre parties du discours ; 2) des rapports des parties du discours les unes avec les autres ; 3) de la situation, au sein du système de la langue, d'un certain type de marques, les translatifs ; » (1989 : 65)

La translation permet également de suppléer les parties du discours existantes, lorsque certaines fonctions ne correspondent à aucune classe spécifique : « Dans les langues, comme le palau, sans adjectif ou sans adverbe, il existe des fonctions (en l'occurrence les fonctions épithétique et circonstancielle) qu'aucune partie du discours ne peut remplir sans l'adjonction d'une marque segmentale, c'est-à-dire des fonctions qui ne sont les fonctions fondamentales d'aucune partie du discours. » (1989 : 65)

Notons que la translation n'est pas seulement un phénomène syntaxique : les opérations de dérivation et de composition aboutissent aussi au changement catégoriel. Par exemple les séries :

fr. : *facile, facilité, facilement, faciliter*
fr. : *plat, platitude, platement, aplatir*
angl. : *wide, width, widely, to widen*

correspondent aux translations syntaxiques suivantes :

X= adjectif, le fait d'être X, de façon X, rendre X

Ainsi la translation met en jeu différents niveaux de *segmentation*, pour lesquels elle réalise l'intégration des unités par différents procédés, car « elle met en rapport – des catégories de mots ; - des catégories de syntagmes ; - des catégories de propositions ; - des catégories d'affixe ou de clitiques. » (Lemaréchal, 1989 : 72-73)

Le phénomène de la translation intéresse donc l'analyse contrastive dans la mesure où il révèle les possibilités transformationnelles inscrites à l'intérieur même de tout système linguistique. Il montre que les structurations en partie du discours n'ont rien de nécessaire, puisqu'un même concept peut correspondre aux fonctions grammaticales du nom (*facilité*), du verbe (*faciliter, rendre facile*), de l'adverbe ou du circonstant (*facilement, avec facilité, en facilitant …, rendant facile …*), de l'adjectif (*facile, facilité, d'une grande facilité*), etc.

Nous n'avons pas encore évoqué les contrastes sur le seul plan syntaxique, qui aboutissent pourtant à des règles transformationnelles simples, du type : l'adjectif en anglais est antéposé par rapport au nom qualifié, tandis qu'il est (le plus souvent) postposé en français ; en allemand l'infinitif est rejeté après son c.o.d., etc. C'est que ces règles présupposent, de manière implicite, la conservation des parties du discours, voire le transcodage mot à mot. Or même si ce transcodage est parfois possible entre des systèmes très proches, il ne constitue aucunement une règle. La projection syntagmatique étant une opération finale, dépendant de tous les choix préalables au niveau du lexique (cf. supra, 45 : le « schème d'entendement »), de la prédication (cf. le « schème prédiqué »), des catégories grammaticales (cf. le « schème résultatif »), les contrastes syntaxiques seuls peuvent difficilement fournir de règles transformationnelles pertinentes au niveau du transcodage.

En ce qui concerne les aspects grammaticaux, les équivalences du niveau contrastif se ramifient et se dispersent à travers les importantes possibilités paraphrastiques déployées par les systèmes linguistiques : au bout du compte, ce n'est pas la conservation de tel ou tel trait grammatical ou partie du discours qui doit guider le transcodage, mais la plus forte adéquation de la construction finale avec l'idiome d'arrivée.

I.1.3.4 Contrastes idiomatiques

Nous avons examiné deux aspects des systèmes linguistiques où se manifestent des contrastes susceptibles d'engendrer des transformations traductionnelles : du côté de la grammaire avec les sous-systèmes morphosyntaxiques et la définition des parties du discours, et du côté du lexique avec la structuration des champs lexicaux et la définition des unités polylexicales. Il faut y ajouter un troisième ordre de contraintes : celui de l'*idiome* relatif à *l'usage* propre à une langue donnée.

Ce qui caractérise l'idiome, c'est qu'il n'exploite qu'une faible part de la combinatoire autorisée par le *système*. De fait, l'idiome établit la correspondance entre des formules toutes faites et des contextes-types, sans pour autant aboutir au figement de ces expressions. De telles expressions idiomatiques, comme les proverbes et les dictons, sont à

prendre en bloc mais ne sont pas assimilables à des unités lexicales. Bien que décomposables, elles sont rarement traduisibles mot à mot :

fr. : (au téléphone) *qui est à l'appareil* ?
it. : *chi parla* ?

fr. : *vous avez l'heure* ?
it. : *che ore sono* ?

fr. : (au magasin) *combien je vous dois* ?
it. : *quanto viene* ?

fr. : (au magasin) *ce sera tout*.
it. : *va bene così* / *basta così*.

Certains phénomènes lexicaux sont de nature idiomatique, dans la mesure où le lexique n'enregistre pas seulement des significations abstraites, mais aussi des usages locaux, des combinatoires spécifiques, des paradigmes restreints. Par exemple, en France le rossignol *chante* tandis qu'en Allemagne on l'entend *appeler* (*die Nachtigall schlägt*). D'une certaine manière, la polysémie, qui manifeste la prégnance du contexte sur l'interprétation de l'unité, est aussi un phénomène idiomatique : elle n'est pas prévisible à partir de la seule valeur du signifié lexical, car elle est le résultat des usages établissant le lien entre le *designatum* et l'usage de l'unité dans un contexte donné.

I.1.3.4.1 Système, idiome et paraphrase

De façon générale, l'idiome se manifeste par des tournures considérées comme plus « naturelles ». Par exemple, dans l'analyse contrastive de l'anglais et du français, on donne fréquemment l'exemple classique des verbes de mouvement combinés à des locutions prépositionnelles :

angl. : *He swims across the river*
fr. : *Il traverse la rivière à la nage*

angl. : *He swam back to the shore*
fr. : *Il regagna la rive à la nage*

angl. : *He ran up the stairs*
fr. : *Il monta les marches en courant*

angl. : *He limped down the hill*
fr. : *Il descendit la colline en boitant*

angl. : *He rushed out of the house*
fr. : *Il sortit de la maison précipitamment*

angl. : *The snake crawled out of the hole*
fr. : *Le serpent sortit du trou en rampant*

Ces équivalences peuvent se ramener à une matrice commune :

angl. : verbe de mouvement + SP (1)
fr. : verbe de déplacement + SN / SP + expression adverbiale

où SN désigne un syntagme nominal, et SP désigne un SN introduit par une préposition.

Etant donné la régularité de ces mécanismes de transformation, il est tentant d'y voir des faits découlant de contrastes grammaticaux. En ce qui concerne les verbes de déplacement français, pris dans ces contextes, leurs équivalents anglais impliquent pratiquement tous des particules entrant dans des locutions prépositionnelles, car ce sont elles qui expriment le type de déplacement :

fr. : *descendre* (la colline)
angl. : *to go down, to come down*

fr. : *sortir* (de la maison)
angl. : *to go out, to come out*

fr. : *monter* (l'escalier)
angl. : *to go up, to come up*

fr. : *traverser* (la rivière)
angl. : *to cross, go across*

A cette contrainte lexicale s'ajoute l'absence d'expression adverbiale permettant d'exprimer le mouvement. Là où le français possède des adverbes (comme *précipitamment*), une locution adverbiale (*à la nage*), une construction grammaticale (le gérondif *en* + part. présent), l'anglais a beaucoup moins de choix (la construction *by* + *-ing* ne convient pas, car elle signifie de façon plus précise : *by means of*). Tout au plus

pourrait-on traduire *précipitamment* par *hurriedly*. Mais la formule *He went out the house hurriedly* semble un peu lourde, pour des raisons idiomatiques.

En revanche l'anglais possède un grand nombre de verbes permettant d'exprimer le mouvement. Toutes ces contraintes, à la fois lexicales, grammaticales et idiomatiques expliquent la régularité de la construction anglaise.

Pour le français, la réalité est légèrement plus contrastée. Si l'on calque les expressions anglaises, on obtient par exemple :

fr. : *Il nagea d'un bord à l'autre de la rivière*
fr. : * *Il nagea de retour vers la rive*
fr. : *Il courut jusqu'en haut de l'escalier*
fr. : *Il boita jusqu'en bas de la colline*
fr. : *Il se précipita hors de la m*aison
fr. : *Le serpent rampa hors du trou*

On constate qu'il est difficile de trouver les équivalents prépositionnels exacts : *back to* n'a pas d'équivalent strict en français lorsqu'elle indique un déplacement – en français, *de retour* signifie un état, le résultat du déplacement, et *de retour vers* n'est pas très idiomatique.

Pour le reste, *jusqu'en haut de, jusqu'en bas de, d'un bord à l'autre de* ne rendent pas tout à fait le sens voulu, même s'ils sont plus proches que *en haut de, en bas de, à travers*. Ces changements subtils entraînent un déplacement de la topicalisation, désormais centrée sur les verbes de mouvement *boita, nagea, courut*, ce qui ne rend plus exactement le sens initial. Ceci s'explique par le fait que ces prépositions insistent sur le déroulement de l'action plus que sur son résultat.

Seule la préposition *hors de* semble conserver la nuance de *out of,* et les deux dernières phrases sont tout à fait satisfaisantes sur le plan idiomatique.

Ainsi, ce qu'on aurait pu analyser comme le résultat de contraintes grammaticales est dû, en grande partie, à des organisations lexicales. Ces observations nous montrent que l'idiome est la résultante complexe d'habitudes langagières conditionnées par des valeurs lexicales particulières et des règles morphosyntaxiques de plus ou moins grande portée.

C'est en ce sens qu'un grand nombre de règles de transformation du type (1) n'ont pas de pertinence grammaticale, mais relèvent de phénomènes idiomatiques. Quand elles existent, ces règles permettent seulement de choisir l'expression la plus « proche » et la plus « naturelle » (cf. Nida), parmi un grand nombre de paraphrases possibles, toutes également grammaticales.

A l'intérieur d'une même langue, la paraphrase montre en effet qu'on peut observer des fluctuations idiomatiques pour des constructions grammaticales identiques et des significations analogues :

1. *il rampa dans la maison*
2. *il entra dans la maison en rampant*

1. *il sautilla dans la maison* (peu idiomatique, ou sens différent)
2. *il entra dans la maison en sautillant*

1. *il bondit dans la maison*
2. *il entra dans la maison en bondissant* (peu idiomatique)

Ces trois groupes d'exemples montrent différents types de configuration lexicale :

– *ramper* peut être suivi de la préposition *dans* pour signifier l' « entrée », tandis que *sautiller* n'admet pas ce genre de combinaison, car ce dernier verbe n'implique pas l'idée de déplacement dans une direction.

– *bondir* suivi de la préposition *dans* n'a pas exactement le même sens que *bondir* sans argument. Peut-être est-ce dû au sens figuré : *cela m'a fait bondir*. En outre le participe présent n'est guère usuel, l'action de bondir étant subite.

On pourra objecter que ces préférences ne sont pas de nature idiomatique, mais conditionnées par la sémantique particulière des unités lexicales. Cependant la réciproque est aussi vraie : si l'usage permettait d'employer *l'oiseau sautilla dans la maison*, le verbe *sautiller* aurait une signification différente. D'ailleurs, en anglais on peut sautiller en avançant : *the bird hopped along in the house*.

Ce qui apparaît dans ces derniers exemples, c'est la difficulté à distinguer entre lexique et idiome. Il n'est pas étonnant que les expressions dites idiomatiques, comme *à bouche que veux-tu*, *tomber la chemise* / *la veste* / *le pull*, soient tantôt rangées dans l'idiome, tantôt dans le lexique.

En effet, le lexique a un caractère largement non systématique, et la sémantique lexicale est étroitement liée à l'usage : usage et signification sont en constante rétroaction. Quand en Gironde, on dit *mettre des cheveux blancs* / *des rides*, cet usage enrichit la signification de *mettre* : /présenter (des signes de vieillesse)/. Véronis (2000b :6) note qu'il est bien plus difficile de s'accorder sur les différentes acceptions d'une entrée de dictionnaire, qui énumère une liste de significations abstraites, que d'établir des « classes d'usage » en fonction de critères cohérents touchant aux distributions syntaxiques des mots. En outre, ces classes semblent correspondre à des contenus cohérents d'un point de vue cognitif. L'idiome inclut donc tous les phénomènes lexicaux si l'on entend le lexique de manière dynamique, non pas comme un stock d'éléments inertes passivement livrés aux mouvements grammaticaux, mais comme l'ensemble des usages de ses éléments.

Les rapports entre grammaire et idiome sont plus faciles à distinguer, car moins étroits : l'idiome se manifeste soit par des préférences vis-à-vis de formulations tout aussi grammaticalement correctes, soit par des infractions aux règles générales (p. ex. avec des tournures archaïsantes *il faut raison garder*, des régionalismes *il est comme ça grand*, etc.).

La morphosyntaxe modèle la structure de la phrase, et ,naturellement, celle-ci est sujette à de nombreuses restructurations dans le passage à la traduction. Comme l'écrit Nida (1969 : 112) : « Certaines caractéristiques comme la longueur de la phrase et sa structuration syntaxique n'ont rien de sacro-saint, et trop souvent la tentative de refléter les aspects formels de l'original alourdit la communication et entrave la compréhension du lecteur. »[88] Nida (1969 : 113) énumère les caractéristiques structurales qui lui paraissent les plus instables : « (a) l'ordre des mots et des phrases (b) les doubles négatives, (c) les

[88] "There is nothing sacrosanct about such features of structure as sentence length and phrase structure patterns, and too often the effort to reflect the source in these formal aspects results in badly overloading the communication and thus making it very hard for the reader to understand"

accords singulier / pluriel, (d) les voix passives et actives, (e) la coordination et la subordination, (f) l'apposition, (g) l'ellipse, et (h) la spécification des relations. »[89]

Mais ceci ne signifie pas que ce sont les contrastes grammaticaux qui déterminent *seules* les transformations. Tous les systèmes autorisent un grand nombre de réalisations différentes, également grammaticales, et véhiculant des significations équivalentes. Aux transformations découlant des différences grammaticales s'ajoutent donc les transformations paraphrastiques, car « des constructions différentes peuvent exprimer les mêmes relations de sens entre les différentes parties »[90] (Nida, 1969 : 48)

Nida (1969 : 49) en donne quelques exemples :

1. *Jesus rebuked Peter.*
2. *Peter was rebuked by Jesus*
3. *Jesus' rebuking of Peter*
4. *Peter's rebuke by Jesus*
5. *the rebuke of Peter by Jesus*
6. *Peter's rebuke by Jesus*
7. *The rebuking of Peter by Jesus*
8. *It was Jesus who rebuked Peter*
9. *It was Peter who was rebuked by Jesus*

Pour Tesnière (1959 : 283) ces changements syntaxiques manifestent l'indépendance du niveau sémantique : « La métataxe n'est qu'une application du principe de *l'indépendance du structural et du sémantique* qui a été signalée ci-dessus, puisqu'il s'agit d'exprimer une idée sémantiquement identique par une phrase structuralement différente. »

Dans toutes les langues, les possibilités paraphrastiques sont très importantes. Apresjan (1973 : 173-175) distingue deux types de paraphrases :

– les transformations lexicales. Il y compte sept rubriques :

(a) passage à la converse (p. ex. *C'est plus un planificateur qu'un rêveur* / *C'est moins un rêveur qu'un planificateur*)
(b) passage à l'antonyme (p. ex. *Tout le monde se souvient* / *Personne n'a oublié*)
(c) passage à un terme générique suivi d'un modificateur (p. ex. *Il murmure* / *Il parle dans un murmure*)

[89] "(a) word and phrase order (b) double negatives (c) singular and plural agreement (d) active and passive structures, (e) coordination and subordination, (f) apposition, (g) ellipsis, and (h) specification of relationship."
[90] "Different constructions may express the same meaningful relationship between the parts"

(d) passage à un dérivé (p. ex. *Il m'aide*, *Il me donne de l'aide*)
(e) passage à un mot relié sémantiquement (p. ex. : *Il régna trente ans* / *Il occupa le trône trente ans*)
(f) passage au « code sémantique » (p. ex. : *Personne ne vint, à part lui* / *Il fut le seul à venir*)
(g) passage à un synonyme lexical (p. ex. : *La distance entre eux fut réduite de moitié* / *La distance entre eux fut diminuée de moitié*).

– les transformations syntaxiques à lexique constant, comme dans les exemples donnés par Nida. Elles concernent l'étape de *prédication* dans le parcours de verbalisation décrit par Pottier (cf. supra, p. 45).

Comme le remarque Fuchs (1981 : 45), ces dernières paraphrases sont censées représenter une structure commune : « Cet invariant commun correspond en général au contenu propositionnel logique ; il porte différents noms, selon les auteurs : " structure profonde " de la grammaire générative, " structure très profonde " chez Martin, " formule sémantique " pour Mel'čuk, " schème conceptuel " de Pottier, " lexis " pour Culioli. »

A toutes ces paraphrases, on peut certes attribuer des nuances différentes, suivant le « postulat structuraliste selon lequel tout changement de forme induit nécessairement un changement de sens, si minime soit-il. » (Fuchs, 1981 : 51).

Ainsi, pour le transcodage, on peut supposer que certaines paraphrases ou reformulations seront plus proches de l'original sur les plans sémantiques et idiomatiques, tandis que d'autres seront plus éloignées. Mais rien n'impose que la « meilleure » paraphrase soit structuralement la plus proche de la phrase originale. Les transformations paraphrastiques s'ajoutent donc aux aménagements grammaticaux.

Par conséquent, pour coller au plus près à la signification de l'original et à l'idiome d'arrivée, la structure de l'original se dissout dans un *double* filtrage, avec les restructurations minimums imposées par le système cible d'une part, et les remaniements paraphrastiques voulus par l'idiome d'autre part.

Les systèmes de TA à approche directe étaient bâtis sur le principe de deux systèmes de transformations séparés, sur les plans du lexique et de la morphosyntaxe. Mais on constate que le transcodage ne peut être réduit à ces deux plans, car il est soumis à plusieurs systèmes transformationnels complexes et entrelacés, surdéterminés par les

contraintes idiomatiques : transformations au niveau des combinatoires lexicales, transformations au niveau des règles grammaticales, transformations paraphrastiques.

Notons que les opérations de transcodage peuvent encore gagner en complexité, si l'on tient compte d'autres formes de codification, comme les prescriptions de *style*. Prenons l'exemple de la traduction suivante, extraite du corpus JOC :

> angl.: *they received assurances that the Bishops would be in no danger and free to move about and to conduct Church services*
> fr. : *ils ont reçu l'assurance que les évêques ne seraient pas en danger, qu'ils bénéficieraient de la liberté de mouvement et de celle de célébrer les offices religieux*.

Dans chacune des langues, on peut donner des paraphrases correctes, tant sur le plan de la grammaire que de la signification, en calquant sur la forme traduite :

> angl.: *they received assurances that the Bishops would not be in danger and would enjoy the freedom of movement and the one of conducting Church services*
> fr. : *ils ont reçu l'assurance que les évêques ne seraient pas en danger, qu'ils seraient libres de se déplacer et de célébrer les offices religieux*.

Cette fois la formulation anglaise paraît un peu alambiquée, tandis que la version française reste correcte. Nous touchons à une forme de détermination encore plus délicate que l'idiome, quoiqu'elle procède du même type de restriction : le style. Pour le français écrit standard, il existe des prescriptions telles que : éviter les répétitions, les constructions trop complexes (relatives en cascades), les phrases trop longues, etc. Mais ces prescriptions ne sont pas généralisables à toutes les langues.

Pour illustrer ce type de préférences stylistiques, Vinay & Darbelnet (1963) donnent les exemples suivants :

> angl. : *As they covered mile after mile, ...*
> fr. : *A mesure que les kilomètres s'allongeaient derrière eux, ...*

> angl. : *The right way was to accept the happiness presented by life itself day after day, year after year*

> fr. : *La sagesse consistait à accepter le bonheur tel qu'il se présente au fil des jours et des ans*.

angl. : *... and the still solitude had echoed and rëchoed in the reports of his gun*
fr. : *... et les calmes solitudes avaient retenti à plusieurs reprises des détonations de son fusil.*

D'après ces deux auteurs, le français aurait tendance à esquiver les répétitions binaires prisées par l'anglais : *mile after mile, day after day*, etc. [91]

I.1.3.4.2 Idiome et structures de la phrase

Nous avons établi la nature des systèmes de transformation engagés dans le transcodage : d'abord, les correspondances lexicales ne sont pas envisageables termes à termes, mais nécessitent la prise en compte d'informations contextuelles et l'identification d'unités dépassant le mot ; ensuite, de ces choix lexicaux, qui déterminent ce que Pottier appelle le « schème d'entendement », dérive le « schème prédiqué » choisi parmi un éventail de solutions paraphrastiques, dont les structures dépendent directement des lexies mises en jeu. Et à chacun de ces niveaux, se manifestent les préférences idiomatiques.

Dans ce parcours transformationnel, certains chemins peuvent avoir une relative généralité : dépendant étroitement des contrastes linguistiques et de la proximité des langues concernées, l'application parallèle des filtres lexicaux, grammaticaux et idiomatiques permet d'observer des régularités locales : notamment, il est possible, lorsque des équivalents lexicaux aux régimes similaires ont été trouvés, de mettre en œuvre localement des règles de transformation de structure à structure.

Envisager ce type de transformation, à la manière de Vinay & Darbelnet (1963), nécessiterait une analyse approfondie des deux langues concernées, ce qui ne rentre pas dans le cadre du présent travail. A titre d'illustration, et sans chercher à être exhaustif ni systématique, nous donnons ci-dessous quelques exemples[92] de transformations structurales, non réductibles aux contrastes grammaticaux :

---

[91] Mais il n'est pas certain que ces prescriptions soient toujours valides en ce qui concerne la littérature française moderne. *des miles et des miles*, *jour après jour*, paraissent aujourd'hui tout à fait acceptables.

[92] certains de ces exemples sont empruntés à Tesnière (1959), Françoise Vandooren, « Divergences de traduction et architectures de transfert » (in Bouillon & Clas, 1993), Ulrich Heid, « Le lexique : quelques problèmes de description et de représentation locale » (in Bouillon & Clas, 1993 : 169-196), et Vinay & Darbelnet (1963).

– *Changement de parcours diathétique*

D'après Tesnière le passif est relativement peu fréquent en français. Certaines formes actives correspondent de manière systématique à une diathèse passive dans des langues comme l'espagnol ou l'italien, notamment pour la traduction du pronom indéfini « on » :

it. : *questo palazzo è stato ristrutturato*
fr. : *on a rénové cet immeuble*

En ce qui concerne l'expression des modalités (au sens large, incluant l'expression des sensations, sentiments et perceptions du sujet), Pottier (1992 : 142) remarque que le français dispose usuellement des deux parcours diathétiques :

Parcours direct	Parcours inverse
je crois que P	*il me semble que P*
il faut que je parte	*il me faut partir*
je suis ravi de cette idée	*cette idée me ravit*

Tandis qu'un grand nombre de langues optent plus naturellement pour la deuxième solution. Pottier donne les exemples suivants :

all. : *mich friert*
fr. : *j'ai froid*

esp. : *se me hace frio*
fr. : *je sens le froid*

esp. : *me duele la cabeza*
fr. : *j'ai mal à la tête*

esp. : *me gusta el chocolate*
fr. : *j'aime le chocolat*

lat. : *mihi opus est libris*
fr. : *j'ai besoin de livres*

La recherche d'équivalents lexicaux peut aboutir à un renversement d'actance : Nida (in Nergaard, 1995) cite l'exemple de certaines langues nilotiques où *Il est arrivé en ville* se traduit par un verbe inverse à la voix passive *la ville a été arrivée par lui*.

– *Causation*

Une tournure causative peut correspondre à un circonstant :

all. : *Davon zitterten die Fensterscheiben*
fr. : *Cela fit trembler les vitres* [litt. de cela tremblèrent les vitres]

Selon Tesnière (1959) ce genre de métataxe est fréquent entre le français et l'allemand, cette dernière utilisant peu la construction explicite du causatif, tandis qu'en français il n'est pas rare qu'on remplace le circonstant de cause par le prime actant, avec une tournure en *faire + verbe*. Ce même type de transformation peut concerner d'autres formes de circonstant, avec par exemple la construction *voir + verbe* :

all. : *Zu dieser Zeit wurde ein grosser Dichter geboren*
fr.: *Cette époque vit naître un grand poète*

– *Parataxe et hypotaxe* (*coordination et subordination*)

Une parataxe, coordination plaçant des éléments sur le même plan, peut équivaloir sémantiquement à une hypotaxe, établissant une relation hiérarchique entre les éléments (qualification, circonstant, etc.).

latin : *spectator et testis*
fr. : *un témoin oculaire*

angl. : *go and fetch me the book*
fr. : *allez me chercher le livre*

fr. : *un geste et je tire*
it. : *se ti muovi sparo* (hypotaxe sans marqueur)

Ces transformations sont toutefois d'application restreinte : dans le dernier exemple, la coordination exprimant la conséquence (p. ex. : *faites-le et vous êtes un homme mort*) est un usage peu fréquent connotant la menace.

Il est possible, de la même manière, de relever des régularités dans les transformations des parties du discours : ce que Vinay & Darbelnet (1963 : 102) appellent une « stylistique comparée des espèces ». Nous emploierons, par analogie, le terme de *translation* pour ce type de transformation : le phénomène ainsi désigné n'est cependant pas identique à la translation monolingue, qui est une opération grammaticale.

– *Translation verbe* / *adverbe*

La translation découlant du transcodage implique aussi des transformations structurales, dans la mesure où les rapports de rection s'en trouvent modifiés. Ces modifications sont particulièrement manifestes lorsqu'elles affectent le nucleus verbal, occupant le centre de l'édifice structural. Par exemple, on assiste souvent à une redistribution des rôles du verbe et de l'adverbe, découlant d'une différence d'intégration de l'adverbe au sein du nucleus. Nous avons déjà évoqué ce phénomène avec l'expression du mouvement. Sur ce point, l'allemand et l'anglais présentent des profils similaires (les unités translatées ne sont pas en italique) :

angl.: *he* strode off
fr. : *il* s'éloigna à grand pas

angl.: *he* limped up *the stairs*
fr. : *il* monta *les escaliers* en boitant

all.: *Er* kriecht aus *dem Loch*
fr. : *il* sort *du trou* en rampant

all.: *Er* rennt in *den Saal*
fr. : *Il* entre *dans la salle* en courant

all.: *Er* schwimmt durch *des Fluss*
fr. : *il* traverse *le fleuve* en nageant

Mais ce genre de transformations n'est pas spécifique à ce champ sémantique, comme le montre l'exemple suivant :

angl.: *she* filed *the number* off
fr. : *elle* effaça *le numéro* à la lime

En anglais comme en allemand, ce sont des particules adverbiales étroitement intégrées au verbe qui expriment le résultat de l'action, tandis qu'en français c'est le verbe lui-même :

all.: *Ich* mache *die Tür* zu
fr. : *je* ferme *la porte*

angl.: *I* brought *it* back
fr. : *je lui ai* rapporté

Ce type de translation n'est pas limité aux particules adverbiales. Par exemple, un syntagme nominal en fonction de second actant peut aussi exprimer le résultat de l'action :

angl.: *he* burned a hole *in her coat*
fr. : *il* troua *son manteau* en le brûlant

angl.: *she* smiled *him* thanks
fr. : *elle le* remercia d'un sourire

L'anglais autorise aussi l'emploi translaté de l'adjectif remplissant la fonction d'un adverbe résultatif :

angl.: *the door* creaked open
fr. : *la porte* s'ouvrit en grinçant

angl.: *he* hammers *the metal* flat
fr. : *il* aplatit *le métal* au marteau

Dans tous ces exemples, le résultat de l'action est exprimé en français par le verbe, tandis que la manière, ce qui a trait au déroulement du procès, est plus spécifiquement réservée à l'adverbe. Là encore, il faut y voir une tendance idiomatique : il ne s'agit pas de prescriptions grammaticales.

On observe la configuration inverse où l'adverbe anglais est exprimé par un verbe :

angl. : *John* usually goes *home*
esp. : *Juan* suele ir *a casa*

angl. : *He* merely said *yes*
fr. : *Il se* contenta de dire *oui*

Dans cette forme de translation, les adverbes *usually* et *merely* sont rendus par des équivalents verbaux, (esp.) *suele* et (fr.) *contenta*. Dans ce cas, l'équivalence est définie au niveau de constructions lexicales spécifiques :

angl. : *usually* + verbe
esp. : *soler* + verbe infinitif

angl. : *merely* + verbe
fr. : *se contenter de* + verbe infinitif

Ces transformations reflètent des préférences idiomatiques : d'après Tesnière (1959), par rapport à l'allemand, au russe ou au latin, le français préfère placer au centre de la phrase des tournures plus « abstraites ». Ce que des langues plus « concrètes » expriment par un adverbe, le français l'exprime en général par un verbe.

Il est difficile de définir linguistiquement cette notion intuitive de degré d'abstraction. Contentons-nous de remarquer que des catégories aussi générales que la modalité, ou l'aspect, sont souvent placées au centre du nucleus verbal, en français, dans des constructions de type verbe + infinitif :

all.: *er zieht sich schnell an*
fr. : *il se dépêche de s'habiller*

all.: *shreiben Sie weiter*
fr. : *continuez à écrire*

all.: *der Fluss steigt unaufhörlich*
fr. : *le fleuve ne cesse de monter*

it.: *Antonio è appena partito*
fr. : *Antoine vient de part*ir

all.: *Ich trinke gern Wein*
fr. : *j'aime boire du vin*

all.: *wir fahren besser*
fr. : *nous ferons mieux de prendre le tramway*

Tous ces exemples correspondent à des constructions lexicalisées en français :

all.: verbe+ *schnell*
fr. : *se dépêcher de* + infinitif

all.: verbe + *weiter*
fr. : *continuer à* + infinitif

all.: verbe + *unaufhörlich*
fr. : *ne* (*pas*) *cesser de* + infinitif

it.: verbe + *appena*
fr. : *venir de* + infinitif

all.: verbe + *gern*
fr. : *aimer* + infinitif

all.: verbe + *besser*
fr. : *faire mieux de* + infinitif

– *Translation verbe* / *substantif*

Ce type de translation peut avoir des origines purement lexicales : par exemple Nida (in Nergaard, 1995 : 166) remarque que dans certaines langues on ne peut dire, littéralement, « Dieu est amour », car le mot « amour » n'existe que sous forme verbale. On doit traduire alors par, littéralement, « Dieu aime ».

Mais ces transformations manifestent parfois des tendances générales : par rapport aux traductions bibliques, Nida remarque que le grec a une prédilection pour les tournures nominales indiquant des actions, mais sans faire référence aux personnes ou objets qui participent à ces actions. Par exemple, là où le grec formule, littéralement : « Jean prêchait un baptême de pénitence pour la rémission des péchés. », Nida (in Nergaard, 1995 : 165-166) pense que l'anglais « préfère » une tournure verbale moins elliptique, littéralement : « Jean prêchait que les gens devaient se repentir et se faire baptiser, afin que Dieu pardonne le mal qu'ils ont fait »

Entre l'anglais et le français, les translations de verbe à substantif sont fréquentes, le français manifestant une préférence pour les noms abstraits :

angl.: *I'm waiting for the postman to pass*
fr. : *J'attends le passage du facteur*

angl.: *A proposal to pay for the equipment*
fr. : *Une proposition de paiement de matériel*

angl.: *after he comes back*
fr. : *après son retour*

angl.: *as soon as he arrives*
fr. : *dès son arrivée*

angl. : *to collide*
fr. : *entrer en collision*

La translation peut être amorcée dans la langue source, comme dans les exemples ci-dessous où le gérondif est déjà une forme de nominalisation :

angl.: *He is responsible for exporting the goods*
fr. : *Il est responsable de l'exportation des marchandises*

angl.: *programming language*
fr. : *langage de programmation*

– *Translation substantif / adjectif*

Ces translations sont fréquentes entre le français et l'anglais, qui emploie des substantifs en position de modifieur :

angl. : *university degree*
fr. : *diplôme universitaire*

angl. : *world market*
fr. : *marché mondial*

angl. : *horse show*
fr. : *concours hippique*

Ceci dit, les translations inverses, même si elles sont moins fréquentes, sont toujours possibles :

angl. : *he is in hurry*
fr. : *Il est pressé*

angl. : *a sweaty smell*
fr. : *une odeur de sueur*

I.1.3.4.3 Idiome et lexique

Dans une certaine mesure, les exemples de transformations précédemment donnés, centrés sur les changements structuraux, présupposent l'existence de lexies équivalentes dans la langue d'arrivée. Or il n'est pas rare que la conservation globale du contenu sémantique, en contexte, soit assurée par des lexies de contenu différent imposant un autre mode de construction de la signification : lorsque ces divergences ne découlent pas du manque d'équivalents lexicaux, mais d'une autre manière d'élaborer le sens, jugée plus naturelle, nous y voyons la conséquence d'un contraste idiomatique.

I.1.3.4.3.1 Idiotismes

Pour qualifier ces *tournures* consacrées par l'usage, ces façons de dire reflétant les habitudes et le caractère propre d'une langue, Gross (1996 : 6) emploie le terme d'*idiotisme* :

> « A partir de là nous appelons *idiotisme* (gallicisme, anglicisme, germanisme) une séquence que l'on ne peut pas traduire terme à terme dans une autre langue, sans pour autant qu'elle soit contrainte dans la langue en question ni sur le plan syntaxique (les transformations habituelles sont possibles) ni sur le plan sémantique (le sens est compositionnel et non opaque). »

Gross donne l'exemple de l'expression *c'est lui-même* répondant à la question *Allô, M. Untel* ?, tandis qu'en allemand on répondrait *Am apparat*. Les formules précédemment citées p. 145 sont des exemples typiques d'idiotisme (p. ex. *qui est à l'appareil* ?, it. *chi parla* ?, etc.).

On ne peut cependant pas restreindre l'idiome à un phénomène contrastif, comme le laisserait entendre la définition de Gross. Bien que le passage à la traduction *manifeste* les

préférences idiomatiques, nous pensons que celui-ci se définit dans un cadre unilingue, à travers la fréquence des usages et le sentiment subjectif des locuteurs, sentiment qui les amène à juger du caractère « conventionnel » ou « normal » d'une formulation comme *c'est lui-même*, ou au contraire de l'« étrangeté » ou de l'« originalité » d'une autre tournure tout aussi correcte et compréhensible, comme *à l'appareil*.

Nous avons résumé, figure 10, les systèmes de contraintes qui régissent la liberté combinatoire des unités linguistiques : les groupes représentés schématisent des ensembles de constructions linguistiques appartenant virtuellement au code ; les frontières de ces ensembles sont floues ; les contraintes lexicales occupent une position intermédiaire entre les prescriptions générales de la grammaire et les formulations spécifiques de l'idiome ; enfin, les usages spécialisés ne définissent pas un « sous-langage » : ils correspondent à l'usage de *la* langue, impliquant toutes ses richesses combinatoires, polysémiques, etc., mais en y ajoutant des prescriptions spécifiques, notamment aux niveaux terminologique et stylistique.

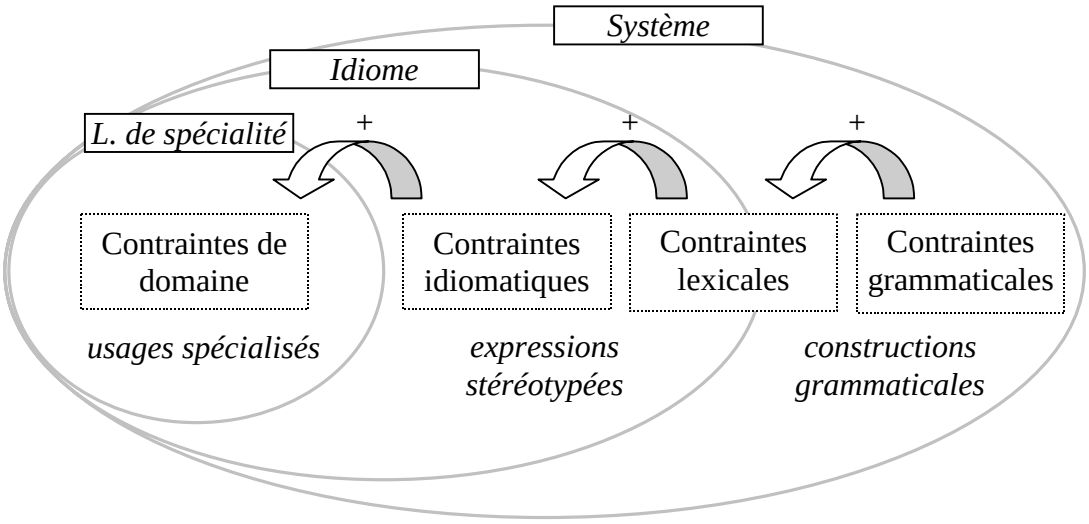


figure 10 : *contraintes et liberté combinatoire*

La particularité de l'idiome est de ne pas faire système : il échappe à toute caractérisation structurale dans la mesure où il ne crée pas de différence sémantique dans la dénotation. D'un point de vue structural, l'idiome est « dénotativement » neutre. En revanche, dans la mesure où il définit un certain *style* linguistique, une certaine naturalité, étroitement liée aux variations dialectales, l'idiome véhicule des connotations : tandis qu'à Paris on dit *grand comme ça*, *haut comme ça*, à Bordeaux, on *préfère* dire *comme ça*

grand, comme ça haut. Or il est clair que l'idiome déborde le phénomène du figement, qui décrit un stade intermédiaire, transitoire et continu, entre idiome et lexique. Dans cette acception, il est impropre de qualifier d'idiotisme une expression figée telle que *haut comme trois pommes*.

L'exemple donné par Gross ne doit pas laisser entendre que les idiotismes sont limités à des unités dont le contenu est essentiellement pragmatique, du type (fr.) *enchanté*, (it.) *piacere*, (angl.) *how do you do* ?, etc. Dans le corpus JOC, nous avons relevé un grand nombre de traductions mettant en correspondance des tournures différentes, reflétant les habitudes linguistiques de chaque langue sans pour autant répondre aux critères de figement énumérés par Gross :

angl. : *race riots*
fr. : *violences racistes*

angl. : *employment*
fr. : *activité professionnelle*

angl. : *measures for its protection*
fr. : *mesure de protection*

angl. : *response preparedness*
fr. : *la faculté de faire face à*

angl. : *are well aware of*
fr. : *n'ignorent pas*

angl. : *to be left unemployed*
fr. : *perdre son emploi*

angl. : *contradicts*
fr. : *est en contradiction avec*

angl. : *cooperate in*
fr. : *unir leurs efforts pour*

angl. : *building a nuclear arsenal*
fr. : *se doter de l'arme nucléaire*

angl. : *have reached maturity*
fr. : *sont arrivés à maturité*

angl. : *have taken place with the support of*
fr. : *ont bénéficié du concours de*

Selon quels critères peut-on confirmer le caractère plus ou moins idiomatique de ces formules ? Une caractérisation rigoureuse devrait être basée sur la fréquence, à l'intérieur d'un corpus, puisque c'est l'habitude et la récurrence qui définit l'idiome. Mais la fréquence absolue ne veut rien dire : il faut pouvoir comparer les fréquences à quelque chose. La seule base de comparaison est le *designatum* : on peut comparer la fréquence relative des expressions qui désignent la même chose, ce qui permet ensuite d'estimer la plus ou moins forte stéréotypie des différentes formules employées. On constate que la notion de « naturel » est vague et relative, même si elle correspond à une réalité.

Une méthode d'enquête consiste à tester, pour une expression donnée, les paraphrases visant le même contenu sémantique : si toutes les paraphrases sont jugées moins « naturelles » que l'expression étudiée, il y a des chances pour que ce soit une tournure plus stéréotypée.

Par exemple, l'expression *se doter de l'arme nucléaire* peut être paraphrasée par de nombreuses tournures sémantiquement équivalentes :

s'équiper de l'arme atomique
se munir de la puissance atomique
fabriquer des armes nucléaires
s'armer de la force nucléaire
etc.

Une enquête auprès d'un échantillon de locuteurs permettrait de déterminer laquelle de ces formulations semble plus fréquente, « conventionnelle », « courante », « naturelle », etc. : si l'expression *se doter de l'arme nucléaire*, ou une autre expression, se détache de manière significative, suivant un seuil qui reste à fixer, on peut alors considérer qu'elle est plus idiomatique.

Ce type de test soulève cependant un problème : dans la mesure où aucune de ces paraphrases n'est totalement identique, l'expression *se doter de l'arme nucléaire* est peut-être plus adaptée au concept sous-jacent parce qu'elle est *sémantiquement* plus juste. En d'autres termes, son utilisation privilégiée ne serait pas le résultat d'une préférence arbitraire (contraintes de l'idiome) mais d'une meilleure adéquation sémantique (contraintes du système). On le voit, la définition de l'idiome est toujours précaire, dans la mesure où les habitudes sont plus ou moins motivées, et plus ou moins arbitraires.

D'une certaine manière, le point de vue contrastif permet de simplifier ce problème d'identification : pour le transcodage, on ne s'intéressera pas aux idiotismes en tant que tels, mais seulement aux expressions dont le transfert mot à mot aboutit à un résultat sémantiquement compréhensible mais non recevable dans l'idiome d'arrivée ; dans ce sens, il apparaît clairement que la plupart des exemples précédents tirés du JOC mettent en jeu des idiotismes français et / ou anglais. Ceci nous amène à réinterpréter la définition de Gross, qui s'ajuste très bien à cette catégorie particulière de tournures idiomatiques, qu'on pourrait appeler « idiotismes contrastifs ».

I.1.3.4.3.2 Collocations

Notre définition des idiotismes couvre un autre type de relation situé à mi-chemin entre le lexique et l'idiome : les *collocations*[93]. Ces relations caractérisent des associations d'unités lexicales ne manifestant ni opacité sémantique ni figement syntaxique, aboutissant pourtant à des expressions fortement stéréotypées.

Par exemple, si un français peut librement choisir entre *demander des renseignements, poser des questions,* ou bien *faire une demande* on lui tiendra grief d'essayer de **demander une question* ou de **poser une demande*. Ces deux dernières tournures, quoique compréhensibles et grammaticalement correctes, ont une consonance bizarre. Et si l'on s'autorise à traduire certaines tournures mot à mot, un anglais peut fort bien **demander une question*, calque français de l'expression *to ask a question* : bien souvent, on reconnaît les locuteurs non natifs à ce genre d'interférence.

Aurélien Sauvageot[94] note qu'« il existe une énorme partie du vocabulaire qui est constituée de mots qui n'ont qu'un rayon d'action limité. Ces vocables ne se construisent qu'avec un nombre plus ou moins réduit d'autres mots, toujours les mêmes, avec lesquels ils forment des locutions stéréotypées ». Sauvageot donne de nombreux exemples de tels stéréotypes :

---

[93] On trouve le terme de *cooccurrence lexicale restreinte* dans I. Mel'čuk (1984) ; d'autres auteurs parlent de *phraséologie* ou *phraséologisme*.

[94] A. Sauvageot (1964), *Portrait du vocabulaire français*, Paris, Larousse. Cité par Clas (1994 : 577).

> « perpétrer un crime, un forfait, une mauvaise action ; commettre une faute, un crime, un péché, une erreur ; élever, émettre une protestation ; prononcer un discours, une allocution ; proférer une parole, une injure, une plainte ; effectuer une opération, un calcul, un voyage ; toucher, percevoir un traitement, une solde, des émoluments, des honoraires. »

Nombre de ces unités sont des prédicats nominaux à verbe support. Gross remarque que ces expressions sont souvent confondues avec des locutions verbales (1996 : 72), alors qu'elles ne présentent pas les caractères du figement (blocage des transformations, non-actualisation, etc.). Très souvent, le sens des verbes support est voisin de « faire » : outre leur fonction de support des marques aspectuo-temporelles, ils n'apportent pas de sens précis à la construction. Etant donné le degré d'abstraction sémantique des verbes support, le problème de l'opacité ne se pose même pas. Et pourtant il y a bien un lien de dépendance, qui fait de ces expressions des unités à traduire comme un tout :

fr. : *prendre une décision*
angl. : *to make a decision*

fr. : *tirer une conclusion*
angl. : *to draw a conclusion*

Ces lexies réunies par un solide lien d'affinité aboutissent donc à des unités polylexicales méritant d'être notées dans les dictionnaires, comme le remarque Clas (1994) avec insistance. De même que l'identification des unités lexicales composées, le repérage et la maîtrise des collocations sont d'une importance cruciale dans l'exercice de la traduction : « Les traducteurs partagent avec les enseignants de langue étrangère, les lexicographes et les terminologues une conviction : la maîtrise d'une langue passe par celle des collocations des mots. » (Lerat, 1995 : 102).

Clas (1994 : 578) distingue divers groupes de collocations, dont il donne un classement en fonction de leurs structures syntaxiques :

1) verbe + nom, où le verbe a un contenu sémantique très général proche simplement de " faire " (*prononcer un discours*) ;
2) nom + adjectif (*rude épreuve*, *marque distinctive*) ;
3) adverbe + adjectif (*vachement bon*) ;
4) verbe + adverbe (*boire goulûment*) ;
5) nom (sujet) + verbe (*la cloche sonne*, *le chat miaule*, *l'abeille bourdonne*) ;

6) marquage de la quantité (unité ou collectif) du nom (*essaim d'abeilles*, *troupeau de vaches*, *pincée de sel*, *barre de chocolat*).

Notons que les *collocations* énumérées par Clas ne sont pas toutes de même nature :

- expression figée : *caresser un espoir*
- idiome : *passer un examen*, (angl.) *sincerely yours*
- détermination de la polysémie par le contexte : *la cloche sonne*, …

Mais toutes ces combinaisons ont un point commun : dans tous les cas la relation d'affinité est antisymétrique, et c'est la combinatoire spécifique d'une des deux unités qui conditionne l'association. Dans les exemples suivants, c'est à chaque fois l'unité à gauche du signe > qui détermine la collocation :

douche > prendre :	*prendre une douche*
épreuve > rude :	*une rude épreuve*
malade > gravement :	*gravement malade*
tancer > vertement :	*tancer vertement*
hibou > boubouler :	*le hibou bouboule*
bois > stère :	*un stère de bois*
test > batterie :	*une batterie de test*

Pour décrire ces phénomènes, Mel'čuk (1984) introduit le concept de *fonction lexicale*, dont il répertorie 21 variétés syntactico-sémantiques, du type :

Magn signifiant « très, intense… »	p. ex. *Magn* (*désir*) = *ardent*
Bon signifiant la louange	p. ex *Bon* (*conseil*) = *précieux*
Func pour la relation sujet - verbe	p. ex. *Func* (*silence*) = *règne*
Mult signifiant « ensemble de … »	p. ex. *Mult* (*chien*) = *meute*
Real signifiant « réaliser »	p. ex. *Real* (*piège*) = *tendre*
…	

L'unité des collocations apparaît très nettement lorsqu'on aborde le point de vue contrastif, car ces groupements d'unités se traduisent rarement de façon compositionnelle, même entre des langues étroitement apparentées :

fr. :	*rude épreuve*
angl. :	*terrible ordeal*

fr. : *gravement malade*
angl. : *critically ill*

fr.: *présenter ses vœux*
québécois : *offrir ses vœux*

fr. : *prendre un verre*
angl. : *to have a drink*

fr. : *prendre une douche*
it. : *fare la doccia*

fr. : *prendre de l'essence*
it. : *fare benzina*

fr. : *serrer la main de quelqu'un*
angl. : *to shake one's hand*

fr. : *passer un examen*
it. : *dare un esame*

fr. : *passer le permis*
it. : *prendere la patente*

Nous pensons avec Clas (1994 : 577) que la lexicographie doit signaler ces combinaisons, car elles « ne sont pas prévisibles ». Mais nous n'en déduirons pas, comme ce dernier, qu'« elles ne sont pas motivées ». Tout comme les différentes organisations de la polysémie (avec lesquelles les collocations entretiennent évidemment des rapports étroits), tout comme les idiotismes élus parmi les multiples possibilités paraphrastiques inscrites dans les codes linguistiques, le phénomène de la collocation manifeste une *préférence* (plus ou moins marquée et contraignante) au sein d'un éventail de possibilités également motivées du point de vue du système. C'est ce *choix* qui n'est pas motivé, non les expressions elles-mêmes : c'est pourquoi nous rangeons plutôt les collocations du côté de l'idiome, même si elles concernent directement le régime des lexies.

I.1.3.4.3.3 Idiome et dérivation

Sur le plan de la lexicogenèse, on peut effectuer le même type d'analyse : la distribution des affixes est parfois soumise à des choix aléatoires parmi des combinaisons

également motivées. En effet pourquoi, si *petit* donne *petitesse*, *grand* ne donnerait-il pas **grandesse* ? Cet arbitraire se manifeste clairement entre des langues apparentées, comme l'italien et le français, qui possèdent un grand stock de racines, préfixes et suffixes correspondants, mais les distribuent différemment :

fr. : *sécurité* ↔ it. : *sicurezza*
fr. : *invitation* ↔ it. : *invito*
fr. : *élevage* ↔ it. : *allevamento*
fr. : *sécheresse* ↔ it. : *siccità*
fr. : *grandeur* ↔ it. : *grandezza*
fr. : *insertion* ↔ it. : *inserimento*

A ce niveau, bien entendu, les unités sont enregistrées dans le lexique et ces combinaisons n'ont pas à être répertoriées séparément. Cependant ces phénomènes méritent d'être notés dans l'étude de la traduction, puisque la néologie fait partie des outils à la disposition du traducteur, en l'absence d'équivalent lexical attesté. La question de l'« idiomaticité » d'une expression nouvelle peut se poser, la préférence entre deux expressions n'étant plus une affaire d'habitude, mais de goût : par exemple, cela peut expliquer qu'on ait préféré *marchandisage* à **marchandisique* mais pas **marchétage* ou **mercatage* à *mercatique*.

Outre les processus de transformation structurale et lexicale, le transcodage doit ainsi réaliser un mouvement de « désidiomatisation - réidiomatisation » pour que le résultat final soit satisfaisant du point de vue de la langue d'arrivée. Or, comme nous l'avons montré, ce mouvement n'est pas neutre, et n'épargne ni les combinaisons lexicales ni les structures grammaticales. Pergnier (1993 : 190) y voit, à juste titre, un véritable « arrachement » :

> « Toute traduction est donc une " désidiomatisation''. La levée des ambiguïtés structurales (...) n'est pas la seule condition de l'équivalence de la traduction. Il faut encore que la langue arrache l'énoncé à l'univers idiomatique auquel il appartient pour l'insérer dans un autre. »

I.1.3.5 Les unités de traduction à l'intersection du lexique, de l'idiome et de la grammaire

La mise en évidence des idiotismes et des collocations nous révèle l'importance d'un nouveau type de lien associatif, qui déborde largement les frontières de la lexie et la portée

des phénomènes de figement. Pour le transcodage, il apparaît que les unités lexicales ne sont pas « recombinables » indépendamment, mais forment des groupes plus ou moins intégrés qui se comportent *comme* des unités polylexicales.

Ce type d'unité correspond à ce que Vinay & Darbelnet (1959 : 37) ou Sager (1994 : 212) appellent simplement « *unité de traduction* ». La minimalité de ces unités découle du critère de *non-compositionnalité traductionnelle* : « On pourrait encore dire que l'unité de traduction est le plus petit segment de l'énoncé dont la cohésion des signes est telle qu'ils ne doivent pas être traduits séparément. » (Vinay & Darbelnet, 1959 : 37). Il ressort de cette définition que l'unité de traduction est identifiée de manière contrastive, et tient sa cohésion des différences sous-jacentes au couple de langues impliqué.

Vinay & Darbelnet (1959 : 37) tentent par ailleurs de caractériser l'unité de traduction sur le plan cognitif : « Nous considérons comme équivalents les termes : unité de pensée, unité lexicologique et unité de traduction. » Cette deuxième définition a malheureusement tendance à brouiller le concept d'unité : comme nous l'avons montré avec les collocations et les idiotismes, d'une part l'unité de traduction dépasse la définition de la lexie, et d'autre part elle peut rassembler des complexes conceptuels. Dans *gravement malade*, on peut sans mal extraire les concepts de « maladie » et de « gravité ». Ce qui fait de *gravement malade* une unité de traduction potentielle, c'est qu'un lien d'affinité antisymétrique est sous-jacent aux deux lexies, le choix de *malade* conditionnant celui de *gravement* (préféré à *fortement*, *profondément*, *dramatiquement*, de *façon critique*, etc.). Mais en définitive, si *gravement malade* constitue une *unité de traduction*, c'est pour une raison contrastive : il ne se traduit pas en anglais de manière compositionnelle, *critically ill* étant préférable à *seriously ill*.

L'identification des unités de traduction est déterminante sur le plan de l'efficacité et de la correction du transcodage, car elle assure une meilleure économie des transformations et elle garantit contre les interférences découlant du transfert mot à mot. Comme le note Sager (1994 : 212) : « La vitesse de travail est largement tributaire de la capacité à reconnaître et à mettre en œuvre de telles unités comme un tout. »[95]

[95] "Speed of work is largely determined by the ability to recognise and process such units as a whole."

Ainsi, sur la base de la définition contrastive, on peut distinguer plusieurs classes d'unités de traduction, qui permettent de synthétiser la plupart des phénomènes de non-compositionnalité relevés jusqu'ici :

– *Groupes polylexicaux figés*

Ce sont des phrasèmes qui ne peuvent être traduits mot à mot. Il peut s'agir de mots composés :

fr. : *clé anglaise*
angl. : *adjustable spanner*

ou de toute autre locution :

fr. : s*ur le point de*
angl. : *about to*

fr. : *avoir le diable au corps*
angl. : *to be like someone possessed*

Dans ce dernier cas la traduction anglaise n'est pas un phrasème, mais une périphrase libre. Notons que le caractère non-compositionnel de cette dernière traduction disparaît entre un autre couple de langues :

fr. : *avoir le diable au corps*
it. : *avere il diavolo in corpo*

– *Prépositions rattachées*

De la même manière, on constate que certaines prépositions, jouant un rôle d'indicateur d'argument, sont *dépendantes* du prédicat. Ces prépositions sont inscrites dans le régime de la lexie, et elles lui sont d'autant plus liées que leur contenu sémantique est relativement vide :

commencer > à
fr. : *commencer à* ↔ angl. : *to begin to*

intention > de
fr. : *l'intention de* ↔ angl. : *the intention to*

prêt > à
fr. : *prêt à* ↔ angl. : *ready to*

ressortir > à
fr. : *ressortir à* ↔ angl. : *to pertain to*

to result > in
angl.: *to result in* ↔ fr. : *avoir pour résultat*

to result > from
angl.: *to result from* ↔ fr. : *résulter de*

Comme le remarque Gross (1996 : 123), les prépositions jouant le rôle d'indicateurs d'argument ne constituent en général pas de paradigme : dans le cas de *to result from* / *in* on a affaire à deux acceptions différentes du verbe *to result*. Ainsi, lorsque la préposition n'est pas en rapport étroit avec la valence du prédicat, on peut considérer qu'elle est en combinaison libre, comme dans l'exemple suivant :

j'ai lutté contre / *pour* / *avec untel*

La préposition, du fait de son rôle de relateur, peut aussi entretenir un lien de dépendance avec l'élément qui la suit :

fr. : *sous condition*
angl. : *on condition*

fr. : *en voiture*
angl. : *by car*

fr. : *à cheval*
angl. : *on horseback*

fr. : *en congé*
angl. : *on holidays*

fr. : *par avion*
angl. : *by plane*

Dans ces derniers cas, les prépositions ne sont pas liées à une structure prédicative : il s'agit de locutions plus ou moins figées. Par exemple, on ne peut dire sans jeu de mot *sous condition et sous rien d'autre*.

– *Collocations*

Les collocations révèlent un type d'affinité lexicale impliquant une dépendance, ce qui permet de dépasser le cadre étroit du figement syntaxique ou de l'opacité sémantique. Fréquemment, ces unités ne se traduisent pas de façon compositionnelle, et forment donc des unités de traduction :

fr. : *un ardent défenseur*
angl. : *a passionate defender*

fr. : *tirer une conclusion*
angl. : *to draw a conclusion*

– *Autres idiotismes*

Les idiotismes incluent les formules d'usage (p. ex. fr. *mille regrets* vs angl. *I'm terribly sorry*), les dictons et proverbes (p. ex. fr. *les petits ruisseaux font les grandes rivières* vs angl. *great oaks from little acorns grow*), ou encore des tournures stéréotypées (p. ex. fr. *se doter de l'arme nucléaire* vs angl. *building a nuclear arsenal*). Ces unités représentent des unités de traduction lorsqu'elles constituent des « idiotismes contrastifs », manifestant des façons de dire divergentes à l'intérieur du couple de langues impliqué.

– *Unités terminologiques*

Aux prescriptions linguistiques peuvent s'ajouter d'autres formes de contraintes : c'est le cas pour tout ce qui touche aux usages normatifs de la langue, où l'on impose artificiellement des règles s'appliquant au lexique ou à la grammaire. Les nomenclatures et terminologies scientifiques ou techniques obéissent à de telles prescriptions, le but étant d'atteindre une plus grande biunivocité dans le rapport de désignation.

Dans une telle perspective, les pluritermes doivent être considérés comme des unités de traduction, car ils doivent être autant que possible transférés tels quels dans les termes correspondants, qui ne sont pas nécessairement formés de la même manière. Par exemple on trouve, dans le corpus JOC :

angl. : *Community Support Frameworks*
fr. : *Cadres communautaires d'appui*

Sur la base de l'équivalence linguistique, on aurait pu traduire autrement :

angl. : *Community Support Frameworks*
fr. : *Cadres d'appui de la Communauté*

mais la contrainte terminologique de biunivocité n'aurait pas été respectée.

Linguistiquement, il n'y a pas de différence entre des composés libres et des pluritermes contraints pas des normes conventionnelles, et Sager (1994 : 225) note que « dans les textes techniques, il est particulièrement difficile de segmenter les syntagmes nominaux en distinguant les unités terminologiques des combinaisons libres ». Le transcodage des termes nécessite donc l'accès à une base de données terminologique répertoriant les équivalences d'une langue à l'autre.

– *Indices co-textuels*

Il peut être avantageux d'enregistrer, pour le transcodage, un autre type de relation entre unités, dépassant le cadre du figement, des contraintes idiomatiques et des collocations : il s'agit de la sélection contextuelle d'une acception, dans le cas d'unités polysémiques.

Comme on l'a vu, la traduction d'une lexie se fait en fonction de sa *désignation*, liée à une acception précise déterminée par un certain contexte. Or, il arrive que le *co-texte* fournisse des indices suffisants pour déterminer précisément, avec de fortes présomptions, l'interprétation correcte de la lexie. Il est alors possible d'établir des équivalences entre des couples d'unités cooccurrentes, ce qui rend en quelque sorte le transfert plus direct.

Pour reprendre un exemple déjà donné, on a :

fr. : *j'ai écouté ce disque*
angl.: *I've listened to this record*

Cette relation de détermination de l'acception par le co-texte permet donc d'établir une forme d'équivalence entre les couples (contexte : unité) :

fr. : (*écouter* : *disque*)
angl. : (*listen to* : *record*)

Notons que les exemples fournis par les dictionnaires bilingues recèlent un grand nombre de ces couples, qui permettent de repérer rapidement le contexte. Par ailleurs, l'automatisation du transcodage peut tirer parti de ce type de cooccurrences, stockées dans un dictionnaire de transfert.

– *Périphrases libres*

Enfin, il faut noter que le concept d'unité de traduction peut s'étendre à des formulations périphrastiques libres, qu'aucun critère unilingue ne permettrait de reconnaître comme unités. Nous avons déjà donné la correspondance suivante :

fr. : *avoir le diable au corps*
angl. : *to be like someone possessed*

L'expression anglaise n'a le statut d'unité qu'en tant que traduction de la locution française. Cette équivalence doit cependant pouvoir être enregistrée, puisqu'elle fonctionne dans le transcodage comme n'importe quel autre couple d'unités.

Cette typologie des unités de traduction est générale mais ne prétend pas être exhaustive : la plupart des exemples présentés concernaient le couple français - anglais : pour d'autres langues, il est possible que les spécificités lexicales et morphologiques débouchent sur d'autres formes d'unités de traduction.

I.1.3.6 Bilan

Nous avons mis en évidence différents niveaux de contraste susceptibles de conditionner les opérations de transcodage. En chemin, nous avons montré que les correspondances parallèles de règles grammaticales et d'unités lexicales n'étaient pas suffisantes pour mettre en œuvre les transformations aboutissant à des traductions valides dans l'idiome d'arrivée.

Malgré les apparences, le problème le plus aigu du transcodage n'est pas celui de l'absence d'équivalence : bien au contraire, les possibilités paraphrastiques et synonymiques de la langue d'arrivée engendrent parfois une profusion de solutions valides sur les plans morphosyntaxique et sémantique, même après avoir résolu les problèmes d'ambiguïté syntaxique et lexicale générateurs d'interprétations divergentes. Le problème du transcodage devient alors une question de *choix*.

C'est pourquoi la notion d'*unité de traduction* est au centre de l'économie des transformations impliquées par le transcodage : d'une part le choix des unités détermine une combinatoire spécifique, et d'autre part l'identification de groupes polylexicaux ayant une cohésion au niveau de l'idiome permet d'appliquer directement des opérations de transfert d'unité à unité (et non mot à mot), ce qui rétrécit d'autant le champ des transformations structurales à envisager. Si en outre on arrive à inscrire au sein de l'unité de traduction les indices co-textuels susceptibles d'en déterminer de façon univoque la correcte interprétation, la relation d'équivalence devient biunivoque car elle shunte les divergences polysémiques.

L'unité de traduction devient le pivot de ce système de transformation : elle enregistre une combinatoire spécifique qui intègre les spécificités idiomatiques difficilement systématisables au niveau de la grammaire ; et elle assure la possibilité d'une équivalence non équivoque, en contexte, respectant à la fois le contenu sémantique interlingue (communauté de désignation) et les idiosyncrasies idiomatiques de la construction du sens.

En d'autres termes, la saisie de segments transportables en bloc, sur la base d'équivalences dépassant le simple niveau de l'unité lexicale, autorise une mise en œuvre modulaire des transformations, ce qui en simplifie la complexité combinatoire. Plus les unités permettront de capter les phénomènes idiomatiques, mieux on pourra appliquer par la suite des règles de transformation structurale d'un niveau général.

I.2 Les corpus bi-textuels et l'aide à la traduction

Nous avons abordé les différents aspects de la traduction en insistant sur la dimension communicative de l'acte de traduire, les opérations « strictement » linguistiques étant surdéterminées par la composante pragmatique (au sens de « pragmatique linguistique »).

L'étude des outils d'aide à la traduction requiert elle aussi une approche « pragmatique », mais dans le sens général du terme, relatif aux applications pratiques de ces techniques. Dans la mesure où ces outils sont destinés à servir des besoins concrets, les questions d'adéquation à la demande, de coût et de viabilité commerciale ne peuvent être éludées. Les critères d'évaluation de ces outils doivent intégrer à la fois leurs performances et le coût de leur mise en œuvre. Comme le proposent K. W. Church & E. H. Hovy (1993 : 241), il faut que « (...) les critères d'évaluation de la TA soient dépendants de l'utilisation que l'on compte faire des systèmes[96] ».

Au cours de la brève histoire de la traduction automatique, l'enthousiasme et l'espoir ont parfois supplanté une véritable vision prospective tenant compte des difficultés théoriques et des contraintes industrielles. Pour paraphraser Somers, on s'est contenté de raffiner les architectures en place « juste pour voir ce qui arriverait[97] ». Répondant à cette lacune, la critique du rapport ALPAC cherchait à positionner la TA à l'intérieur d'un cadre de réflexion élargi, touchant aussi bien les problèmes théoriques que la question de la rentabilité économique. Aujourd'hui, les promesses non tenues, mais aussi certains succès, imposent plus que jamais de reconsidérer les champs ouverts par la TA, la TAO ou l'aide à la traduction, afin de déterminer lucidement ce qu'on peut en attendre, ainsi que les directions les plus intéressantes à explorer.

Sur le plan pratique, les produits de l'automatisation n'entrent pas nécessairement en concurrence avec la traduction humaine : l'automatisation ne fait que produire des outils qui peuvent, le cas échéant, intervenir dans le cadre global d'une situation de communication. Comme le suggère Sager, il faut remettre la TA à sa juste place, en la

[96] "(…) propose that MT evaluation metrics should be sensitive to the intended use of the system"
[97] "just to see what would happen"

considérant seulement comme un élément susceptible de rentrer en jeu à une étape du processus de médiation engendré par la traduction :

> « La nature et les performances de la traduction automatique ont d'abord été définies par comparaison avec la traduction humaine. Une telle posture est fallacieuse car basée sur une conception erronée de la traduction. Qu'elle soit humaine ou non, la traduction est une activité de médiation, dont la forme particulière est déterminée à la fois par le texte et les circonstances communicatives qui demandent cette médiation. »[98]

Ainsi, quel que soit le degré d'automatisation apporté à l'activité traductionnelle, les applications de la TA, ou de la TAO, doivent être situées dans l'horizon de l'aide à la traduction, ou traduction est pris dans le sens général d'activité communicative développé au chapitre I.1.

Après un bref aperçu des approches à base de règles, reposant sur les techniques classiques du TAL, nous verrons comment l'exploitation des corpus bi-textuels introduit une perspective originale, dont les applications pratiques s'inscrivent harmonieusement dans le champ de l'aide à la traduction.

I.2.1 L'automatisation du processus traductionnel

Pour Catherine Fuchs (1993 : 109), le traitement automatique est à la source d'un véritable dilemme, puisqu'il faut trancher entre « le souci de cohérence théorique d'un côté, et la recherche de solutions opérationnelles au moindre coût de l'autre. » De manière atavique, l'automatisation s'est toujours située à la limite de la recherche et du développement. D'une part elle se nourrit de spéculations théoriques ; d'autre part elle est stimulée par les probables retombées industrielles de ses avancées technologiques.

Avant de déterminer pour quelle tâche l'automatisation peut être économiquement avantageuse, on peut dégager sommairement ses conditions d'application théorique.

---

[98] "(…) the nature and quality of machine translation was initially defined only in relation to human translation and by the degree of closeness they achieved to human translation. Such assessments are inherently fallacious because they are based on an erroneous conception of translation. Whether human or otherwise, translation is a mediating activity, the particular form of which is determined both by the text to be mediated and the communicative circumstances which call for this mediation."

I.2.1.1 Limites théoriques de la TA

Une première question se pose : peut-on automatiser totalement la traduction ? D'après Bourquin (1991 : 117) ce « qui a desservi la traduction automatique depuis toujours c'est qu'on a voulu (et que l'on continue de vouloir) *implémenter* trop vite et implémenter *tout*. La vraie TA ne sera informatisable avant longtemps. » Comme Bourquin (1991 : 113) nous pensons que la question centrale doit être formulée différemment : « Que peut-on automatiser en traduction ? ».

Sur le seul plan théorique, il est possible de donner des éléments de réponse à cette question.

Par exemple, un début de réponse est donné par certains auteurs, qui énumèrent des contraintes *a priori* portant sur les textes susceptibles d'être traités : pour Jean Gordon (cité par Sager, 1994 : 292) les textes doivent être « sous format électronique, formatés de manière adaptée, avec une terminologie stable et limitée, sans polysémie, traitant un sujet étroitement spécialisé ; sans ambiguïté, respectant l'usage, sans erreurs typographiques ou autres, avec des phrases courtes, sans omissions ni ellipses, sans tournures idiomatiques complexes, etc. »[99] Toutes ces conditions tendent vers la simplification du traitement : il est vrai que plus les règles à automatiser sont simples et réduites, moins on risque les traitements erronés.

Mais ces limitations sont d'ordre heuristique et pratique. Sur le plan théorique, l'automatisation rencontre des limites fondamentales qui ne dépendent ni de la richesse ni du raffinement de l'implémentation.

I.2.1.1.1 Restrictions interprétatives

Comme on l'a vu précédemment, l'acte interprétatif suppose la présence d'un sujet, capable de recevoir le sens au sein d'un horizon phénoménologique. Quand même il existe une face objective à la manifestation du sens, la subjectivité est au principe de la compréhension. Pour reprendre les termes de Ricoeur (1994), il ne peut y avoir *explication*

[99] "machine-readable, suitably formatted, limited and stable terminology, without polysemy, of a narrow special subject ; no ambiguity, respectful of usage, no typographical or other errors, short sentences, without omissions or ellipsis, no complex idioms etc."

du sens que dans la mesure où il y a déjà *compréhension*. L'absence de subjectivité se traduit par les limitations suivantes :

– *Figement et réduction du contexte*

L'interprétation du sens demande la saisie du contexte extralinguistique. La machine n'ayant pas directement accès au contexte de la communication, il est donc nécessaire de fixer préalablement les paramètres situationnels et d'expliciter certaines données contextuelles. Or les ambiguïtés lexicales et syntaxiques, l'implicite, la présupposition ne peuvent être pris en compte que de manière rigide et préformée. Certaines interprétations doivent être déterminées à l'avance, de façon univoque, afin de neutraliser les insuffisances du plan formel.

En d'autres termes, il devient nécessaire d'« encoder » le contexte dans le fonctionnement de la machine : il y a *figement contextuel*. Le contexte doit en outre être assez spécialisé pour réduire la polysémie des lexies employées et limiter les possibilités d'interprétations multiples : il y a *réduction contextuelle*.

– *Problème des ambiguïtés d'artefact*

Le sens commun est une notion vague, imprécise et problématique pour le traitement automatique. Il fait appel à un éventail de structures cognitives trop vaste pour être actuellement mis en œuvre. Les recherches en Intelligence artificielle, dans l'élaboration de modèles cognitifs, n'ont su capter de manière satisfaisante que des savoirs d'expert. Ces difficultés sont à la source de ce que Fuchs (1993 : 115) appelle des « ambiguïtés d'artefact », qui n'apparaissent qu'au niveau du traitement automatique. Par exemple, dans l'analyse du syntagme « la période de reconstruction de l'après-guerre », on peut rattacher reconstruction à « période » ou à « après-guerre »[100] : ce type d'ambiguïté ne se présente même pas à l'esprit d'un locuteur humain, puisque l'idée de « reconstruction de l'après-guerre » paraît saugrenue ; pour l'ordinateur, une telle ambiguïté demande un traitement spécifique. Seleskovitch (1984 : 182), note que le problème de l'ambiguïté des

[100] ((période (de reconstruction)) (de l'après-guerre)) *vs* (période (de reconstruction (de l'après-guerre)))

significations ne se pose pratiquement jamais pour le traducteur humain, qui a affaire au *sens* :

> « […] le traducteur humain se rit des difficultés de la machine car la situation qu'évoquent les mots lui fait comprendre ce que les faits de langue ne suffisent pas à éclaircir. *Bulb* peut vouloir dire *oignon* ou *ampoule* ; l'interprète qui s'attacherait au seul contenu sémantique des signes pourrait hésiter. Or dans un contexte donné les deux concepts qui s'attachent à *bulb* ne se présentent jamais ensemble ; l'interprète qui suit le sens n'en entend jamais qu'un. L'esprit humain ne lève pas plus la polysémie qu'il ne dissipe les ambiguïtés ; le contexte verbal, la situation ambiante, les connaissances extralinguistiques font qu'il est rare que l'on attribue plusieurs significations à un même énoncé (à moins que celui-ci ne vise délibérément le jeu de mots). »

L'élimination des ambiguïtés lexicales ou syntaxiques reste un problème épineux pour l'automatisation, car peu de textes en sont exempts, même en domaine spécialisé.

Il existe plusieurs façons de contourner ces problèmes : soit en imposant des contraintes draconiennes sur les textes à traiter (restrictions syntaxiques, etc.), soit en intégrant les informations additionnelles (relations sémantiques, modèles probabilistes, etc.), soit en ayant recours à l'intervention humaine pendant traitement.

I.2.1.1.2 Ouverture vs fermeture

La totalité des énoncés appartenant potentiellement à la langue ne peut être explicitée. Les énoncés potentiels sont en nombre infini, mais ce n'est pas cela qui pose problème pour l'automatisation, car l'ensemble des règles et des éléments permettant de produire ces énoncés peut être fini. Ce qui est source de difficultés, c'est que cet ensemble de règles et d'éléments n'est jamais complètement *défini* : ses limites sont floues dans toutes les directions, tant au niveau géographique, temporel, qu'interindividuel, et l'acceptabilité d'un énoncé découle de la conformité avec un usage dont personne ne possède la référence absolue.

Cette forme d'*ouverture* est un caractère inhérent aux langues vivantes. La machine elle aussi peut évoluer en fonction des variations d'usages, dans la mesure où ses programmes sont susceptibles d'être modifiés. Mais dans un état donné, elle ne pourra recevoir, quel que soit le système, que des textes conformes à sa programmation. L'humain est capable de comprendre des textes « irréguliers » par rapport à son système : car la « grammaticalité » chère aux générativistes n'est pas une condition nécessaire (ni

suffisante) de la compréhension. Lorsqu'on automatise, tout doit donc être explicité et fixé à l'avance, ce qui impose la définition de frontières artificielles entre ce qui est correct et ce qui ne l'est pas, au niveau des contraintes grammaticales, du lexique, des équivalences terminologiques, des tournures idiomatiques, etc. La mise en œuvre d'un système de TA suppose donc une relation d'inclusion entre les textes traités et les spécifications linguistiques enregistrées dans le système : ces textes doivent former un sous-ensemble fermé au sein de l'ensemble des énoncés de la langue.

Ainsi, les limitations théoriques se ramènent toutes au même commun dénominateur : la fermeture *a priori* de tout système de TA, qu'il s'agisse de fermeture au niveau pragmatique, impliquant des interprétations préétablies, le figement et la réduction contextuelle, où au niveau linguistique, les unités et la combinatoire devant être explicités et enregistrés par avance.

La fermeture est une propriété essentielle dans la détermination du champ d'application de la TA ou de tout autre outil d'aide à la traduction : soit elle oblige à se restreindre à un certain type de texte défini par le système, soit elle impose l'intervention de l'humain à un moment ou à un autre du processus traductionnel – à moins qu'on se satisfasse d'une couverture partielle et de résultats médiocres mais suffisants pour remplir leur fonction communicative.

I.2.1.2 Types d'application

Remarquons que sur le plan pragmatique, le recours à des outils informatiques modifie certaines conditions de communication :

- on peut utiliser la traduction automatique à partir d'un poste à domicile, sur Internet ou sur un autre médium, sans faire appel à un tiers et sans connaître la langue source ou cible ;

- à la différence d'une traduction humaine, le temps de traduction est indéfiniment compressible puisqu'il ne dépend que de la puissance de calcul mise en œuvre (et que celle-ci est en constante augmentation avec l'évolution du matériel) ;

– dans le cas de traduction entièrement automatisée, on assiste à l'apparition d'un nouveau type de texte, artificiel, qui peut requérir des traitements spécifiques.

On peut établir une classification des différentes familles d'outils d'aide à la traduction en fonction du type d'interaction homme - machine engagée par le processus, des prérequis nécessaires à la mise en œuvre et du type de résultat obtenu :

– *Traduction automatique*

Comme on l'a vu, les systèmes de TA présupposent la satisfaction d'un certain nombre de contraintes au niveau des textes sources, contraintes qui se manifestent globalement par ce que nous avons appelé fermeture langagière.

Sur le plan formel, une phase de pré-édition est généralement requise : elle concerne notamment la définition d'un format électronique, la standardisation de la ponctuation, l'utilisation de balises du type SGML pour la définition du format et des sections logiques, la normalisation de la typographie, etc.

A l'autre bout du processus, du côté de la cible, des contraintes plus ou moins fortes peuvent peser sur les résultats. On distingue différents niveaux d'exigence vis-à-vis de la qualité recherchée :

– *pré-traduction* visant à une compréhension approximative : on peut se contenter d'une traduction mauvaise, incorrecte vis-à-vis du système et de l'idiome d'arrivée mais laissant percevoir, *grosso modo*, le sens du texte original. Ce type de traduction est utile comme *aide à la lecture*, par exemple dans des situations ponctuelles où le texte traduit n'a pas à être conservé, ou encore pour la traduction automatique de courrier électronique ou d'articles scientifiques. Tout repose sur les capacités d'interprétation du lecteur, et sur l'idée que la cohérence textuelle doit permettre de palier les erreurs de traduction locales.

– *pré-traduction* suivie d'une *post-édition*. La post-édition peut être assurée par des réviseurs ou des traducteurs professionnels, suivant les besoins. On peut considérer dans ce cas que la traduction se fait en deux phases dont la première seulement est automatique.

– *traduction de qualité* aboutissant à un produit fini. Il est évident que l'obtention d'un tel résultat impose des contraintes drastiques au niveau des textes sources : contexte figé, syntaxe réduite, terminologie et lexique fixés. La langue des textes sources doit être *artificielle*, dans le sens défini p. 184. On peut citer l'exemple de la documentation technique de Caterpillar, ou des bulletins météo traités par le système TAUM.

Pour Sager (1994 : 257), quelle que soit la qualité du résultat, la sortie du traitement automatique est toujours, elle aussi, *artificielle*, car conforme à des informations préalablement définies et enregistrées : « De façon générale, la sortie de tous les systèmes de TA est artificielle, car (…) les règles et le lexique doivent en être définis avant l'usage. »[101] On peut nuancer ces propos, dans la mesure où cette forme d'artifice n'impose pas nécessairement l'explicitation de règles, la définition d'une grammaire formelle, etc. : dans le cas d'un système de traduction se basant sur des exemples préenregistrés, ce travail d'explicitation n'a pas lieu.

La figure 11 résume les différentes phases du processus de traduction automatique (la partie en gras représente la phase automatisée).

figure 11 : *les trois étapes de la Traduction automatique*

– *Traduction semi-automatique*

Par traduction semi-automatique, on désigne les approches fondées sur le dialogue, comme le système LIDIA (H. Blanchon in Clas & Safar, 1992 : 31-47). Ce type de processus, parfois appelée *Traduction automatique fondée sur le dialogue* (TAFD),

nécessite l'intervention de l'utilisateur humain au cours du traitement automatique. Comme le note Eric Wehrli (in Clas & Safar, 1992 : 61) : « L'intérêt de l'approche interactive est de permettre à l'intervention humaine dans le processus de traduction de s'exercer non pas avant ou après la traduction d'un texte, mais pendant celle-ci, c'est-à-dire au moment où un problème surgit. » Le principe de tels systèmes est de soulager la machine des tâches les plus délicates, qui mettent en jeu des données extralinguistiques complexes à modéliser : « Ainsi, face à des problèmes délicats d'ambiguïté, le système peut afficher les diverses possibilités et laisser le soin à l'utilisateur de sélectionner la plus appropriée d'entre elles. En effet, il arrive souvent que la sélection de la possibilité la plus satisfaisante repose sur des critères faisant intervenir des connaissances non linguistiques (connaissances encyclopédiques, bon sens, etc.) » (ibid.). Dans sa description d'un système de TAFD, Eric Wehrli considère plusieurs niveaux d'interaction possibles : niveau lexicographique pour la traduction des unités lexicales inconnues, niveau des ambiguïtés syntaxiques (concernant par exemple les problèmes de rattachement comme dans la phrase « à qui avez-vous promis d'écrire ? »), niveau des ambiguïtés sémantiques ou pragmatiques (comme la portée de la négation dans « Jean ne bois pas parce qu'il est triste »).

Toute la difficulté de tels systèmes est d'identifier correctement les problèmes d'ambiguïté et de défaillance lexicale afin d'en demander une résolution à l'utilisateur. Ces systèmes ont l'avantage d'être accessibles à partir de la langue source sans connaître la langue d'arrivée, à la différence d'une post-édition requérant l'intervention d'un traducteur. Ils se basent sur l'hypothèse que les difficultés sont concentrées dans la phase d'analyse, et non pendant la génération :

> « De plus, si, comme nous l'admettons, des phrases complètement désambiguïsées peuvent être correctement traduites de façon automatique, il devrait être possible, en principe, de restreindre les interactions aux composantes d'analyse et de transfert et, par conséquent, de restreindre le dialogue à la seule langue source. » (Wehrli in Clas & Safar, 1992 : 67)

Le schéma de la figure 12 représente ce type de processus d'aide à la traduction :

---

[101] "Generally speaking, the output languages of all MT systems are artificial because they conform to the definition established in section two, namely that the rules and the lexicon of an artificial language must be established prior to its use."

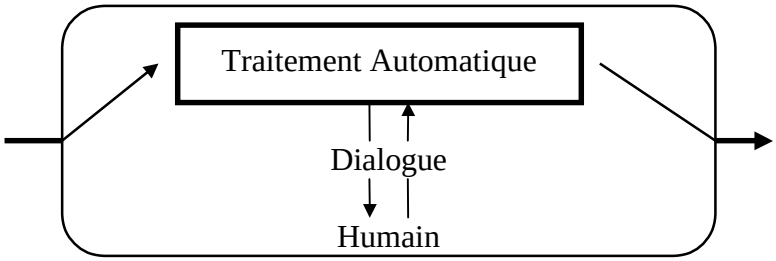


figure 12 : *Traduction fondée sur le dialogue*

– *Autres outils d'aide à la traduction*

Les *Stations de travail pour le traducteur* (STT) sont conçues pour rassembler divers outils intégrés dans un environnement cohérent et ergonomique, dans le but d'améliorer la productivité et d'accroître l'autonomie des traducteurs occasionnels ou professionnels.

Boitet (in Clas & Safar, 1992 : 11) énumère différents types d'outils liés à ce qu'il appelle la « Traduction humaine assistée par la machine » :

- l'enregistrement de traductions déjà faites pouvant « fournir rapidement la solution à de nombreux problèmes de traduction, et augmenter l'homogénéité de l'ensemble des traductions. » La constitution massive et structurée de ce type d'information aboutit à ce qu'on appelle des *Mémoires de traduction* (MT).

- les « outils destinés aux professionnels indépendants », comme les dictionnaires en ligne, les glossaires terminologiques, etc.

- les « environnements destinés à des traducteurs occasionnels » offrant « outre des dictionnaires bilingues en ligne, des outils liés à la rédaction dans la langue cible, comme des correcteurs d'orthographe, des " critiqueurs " de style, des conjugueurs, des thesaurus, etc. » Ces outils peuvent aussi s'insérer dans le cadre d'une aide à la rédaction en langue étrangère, destinée à produire directement un texte dans une langue qu'on ne maîtrise pas parfaitement.

Le processus traductionnel ainsi aidé est représenté figure 13 :

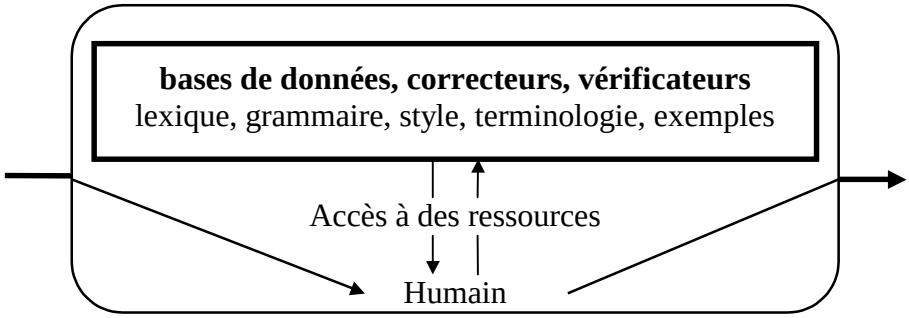


figure 13 : *accès à une Station de travail pour traducteur*

Aux différents degrés d'automatisation précédemment décrits correspondent des applications diverses. Nous empruntons la typologie énoncée par Christian Boitet (in Clas & Safar, 1992 : 5) qui distingue quatre « niches » d'application :

1. *TAO* « *du veilleur* » : il s'agit d'obtenir, à des fins de veille technologique, des traductions complètement automatiques, même de faible qualité. Ce type de traduction, visant essentiellement l'accès à l'information, connaît également un développement important sur Internet, pour la traduction de pages de la Toile ou de courriels.

2. *TAO du réviseur* : on cherche à obtenir un premier jet de traduction, de mauvaise qualité, destiné à être corrigé par l'humain en post-édition.

3. *TAO du professionnel* : on offre l'accès à des outils informatisés (dictionnaires, concordanciers, bases de données, correcteurs, etc.) à travers une interface adaptée (STT). Ce domaine est en pleine expansion, et des produits commerciaux sont distribués sur le marché (Trados, SDLX, etc.).

4. *TAO de l'auteur* : on offre à l'auteur des aides à la rédaction en langue étrangère. Aux outils mis à la disposition des traducteurs s'ajoutent les systèmes avec une approche fondée sur le dialogue.

I.2.1.3 Atouts et spécificités de l'automatisation

Il nous faut maintenant évaluer l'intérêt pratique des outils présentés. Sager (1994 : 120) exprime très clairement les données du problème :

> « La question centrale est par conséquent de définir précisément ce que l'ordinateur peut faire, et de déterminer de quelle manière il peut être naturellement utilisé, soit (i) pour faciliter le travail humain ; ou (ii) pour fournir un service que les humains ne peuvent fournir en terme de précision, de consistance et d'exactitude ; ou (iii) pour augmenter la productivité et dégager, par là même, des bénéfices à partir des investissements effectués dans l'automatisation. »[102]

– *Faciliter le travail humain*

Ce premier point est au principe même du développement des STT. Il ne peut être satisfait que si les stations de travail réunissent certaines conditions d'ergonomie et de souplesse : interface adaptée ; capacité d'enrichir les ressources terminologiques et dictionnairiques, d'augmenter et de modifier les base de données ; possibilité de fixer des paramétrages personnalisés en fonction des tâches à effectuer, etc. La philosophie de tels systèmes étant de permettre à l'utilisateur de constituer ses ressources *ad hoc*, en fonction de ses besoins spécifiques (en terminologie, en lexique, en grammaire, en informations idiomatiques, en exemples) la possibilité de paramétrer simplement des outils complexes est essentielle : comme nous le verrons par la suite (Parties II et III de ce travail) le succès de certaines techniques dépend étroitement de réglages dépendant du couple de langues, du domaine, du format des textes, etc.

Dans un commentaire sur des outils afférents à l'exploitation d'une mémoire de traduction, Eric Gaussier, David Hull & Salah Aït-Mokhtar (in Véronis, 2000 §13) notent que les résultats quantifiés de certaines techniques ne garantissent pas la réussite de leur utilisation :

> « Bien que l'on puisse estimer indépendamment la précision de l'algorithme d'alignement, il est difficile de prévoir si ce niveau de performance sera suffisant pour accroître la productivité. La valeur du système dépendra aussi de questions

[102] "The key issue is, consequently, to define what precisely the computer can do, hence how it can be used in the way people generally use machines, either (i) to ease human labour; or (ii) to provide a service that human beings cannot provide in term of accuracy, consistency or precision; or (ii) to increase productivity and thereby to recover with profits the investment made in machines the first place."

> (…) telles que la manière dont l'interface est conçue pour que la mémoire de traduction soit utilisée efficacement. »[103]

L'accès simplifié – et rapide – à des outils parfois complexes dans leur mise en œuvre, intégrés dans une architecture ergonomique adaptée au travail du traducteur, est un champ d'étude immense, et encore pratiquement vierge, tant au niveau de la recherche que du développement.

– *Accroître la rigueur de la communication*

Sager pose la question d'une amélioration qualitative apportée par l'automatisation. Dans certain cas, la TA peut-elle donner de meilleurs résultats que la traduction humaine ? On peut botter en touche en arguant que tout dépend des compétences effectives des traducteurs. Cependant, de façon générale, il est vrai que le traitement automatique présente des avantages sur le plan de la systématicité. En principe, un terme sera toujours traduit de la même manière par la machine, dans un contexte non ambigu. Nous verrons plus loin comment l'ordinateur peut être employé à des tâches de révision de traduction humaine comme la détection automatique d'erreurs de traduction (faux amis, traductions mot à mot d'idiotismes), la vérification de la cohérence terminologique, la détection d'oubli, etc.

Par ailleurs, la supériorité du traitement automatique peut se manifester, d'après Sager, dans les contraintes mêmes qu'il impose. L'obligation de rédiger dans une langue appauvrie sur le plan syntaxique et lexical peut devenir un gage d'efficacité. Les restrictions sous-langagières sont aussi un moyen de contrôler la communication et d'assurer une certaine rigueur : « L'aide informatisée à la rédaction dans un langage restreint peut aussi viser à augmenter l'efficacité de la communication en même temps que la traductibilité des documents »[104] (Sager, 1994 : 274). Un texte généré artificiellement, ou composé avec des outils d'aide à la rédaction, peut être conservé dans un format riche contenant des informations morphosyntaxiques et sémantiques intéressantes, réutilisables

[103] "While we can estimate the accuracy of the alignment algorithms independently, it is hard to know in advance whether that level of performance is sufficient to improve productivity. The value of the system will also depend on issues (…) such as whether the interface is designed in a way which allow the translation memory to be used with maximal efficiency."

au cours de traitements ultérieurs (p. ex. pour une nouvelle traduction dans une autre langue). L'automatisation des traitements textuels peut amener à un effort de rationalisation de la communication, comparable aux notations formelles dans les domaines scientifiques. Ce travail de formalisation linguistique étant limité, comme on l'a vu, aux actes de langage rétrospectifs situés dans un domaine précis.

– *Accroître la productivité*

Le troisième point concerne l'augmentation de la productivité. La question ici posée est celle du coût, avec ses deux facettes : coût en temps et coût financier. Notons qu'elle est d'une extrême complexité : les coûts de la TA impliquent des calculs à long terme, car il faut compter avec des investissements initiaux souvent importants. Il faut tenir compte des aspects suivants :

- investissement financier (en travail humain et en matériel) et durée concernant le développement initial d'un système.
- coût et durée de l'utilisation effective du système, en tenant compte des traitements automatiques et des phases de travail humain (pré-édition, post-édition).
- coût des mises à jour du système pour son adaptation et son évolution dans le temps.
- possibilité de réutiliser le système pour de nouvelles tâches et coût de cette « portabilité ».

Ces différentes phases donnent lieu à des calculs compliqués. En ce qui concerne la TA avec post-édition, Boitet (Clas & Safar, 1992 : 6) pense que celle-ci n'est « envisageable que pour de gros flux de textes homogènes et informatisés, comme des

[104] "Computer assistance in writing restricted language can at the same time be directed towards increasing both the efficacy of communication and the translatability of documents."

manuels d'utilisation ou de maintenance. »[105] L'adaptation d'un système de traduction à de nouveaux types de texte pose les mêmes problèmes : elle n'est rentable que si la quantité des textes est importante, et ne peut être envisagée pour des tâches ponctuelles : « Adapter un système de TAO du réviseur à des besoins comparativement ponctuels serait comme réoutiller une usine pour produire quelques dizaines de voitures. » (ibid.: 7)[106] Il faut en outre prendre en compte la pénibilité du travail de révision, proportionnelle à la mauvaise qualité de la traduction brute, « car les réviseurs n'acceptent pas de réviser de la trop mauvaise qualité, et préfère retraduire, ce qui, au total, est plutôt contre-productif » (ibid. : 8).

Au niveau du temps, notons que la possibilité de traitements automatiques massivement parallèles (par une segmentation appropriée des textes) autorise une rapidité de traduction hors de portée de l'humain.

Enfin, pour des systèmes complexes incluant de grandes quantités d'informations linguistiques, la mise à jour et la possibilité de réutiliser ces informations pour d'autres applications est un enjeu majeur. Les recherches en TA ont évolué dans cette direction, avec le développement d'architectures modulaires orientées vers une plus grande souplesse dans la gestion des informations linguistiques. Par ailleurs, le principe des Mémoires de traduction, cumulatif par essence, permet de contourner élégamment le problème de la mise à jour des bases de données linguistiques.

I.2.2 Bi-textes et traduction automatique

Du seul point de vue de la TA, l'idée de recourir à des bi-textes marque une évolution importante, puisqu'elle est à l'origine d'un nouveau paradigme : ce qu'on

---

[105] « Dans ces conditions, un système à 1 MF (400 KF de base et 600 KF de spécialisation au vocabulaire et au type de texte) doit pouvoir être amorti en deux ans, pour un flux de 10 000 pages par année (en comptant 10 %/an de maintenance, 60F/page de coût machine, et 100F/page de révision, contre 150F/page de traduction et 70F/page de révision pour la méthode manuelle classique, soit 60 F/page de gain pour amortir 1,2 MF »

[106] « En effet, sans compter la saisie optique ou manuelle, entraînant toujours un coût important de vérification, ni la maintenance, ni même l'achat du système de base, mais seulement sa spécialisation (600 KF) et les coûts de traduction et de révision, on arrive à 632, 680, 760 et 920 KF pour 200, 500, 1 000 et 2 000 pages, contre 44, 110, 120, 220 et 440 KF pour la méthode classique manuelle, soit environ 14.5, 6, 3.5 et 2 fois plus, respectivement [...] le point d'équilibre se situe à 9 000 ou 10 000 pages »

appelle la *Traduction automatique basée sur l'exemple* (TABE). Pour en dégager les enjeux, il est important de situer ce courant dans le contexte général de la TA.

Bourquin (1993 : 27) remarque que l'automatisation est rendue possible par la mise au jour de *régularités* :

> « Vouloir automatiser, c'est chercher à capter des régularités, les coordonner, les hiérarchiser, découvrir entre elles des solidarités sur lesquelles s'appuyer pour dégager des procédures inférencielles (...). S'agissant de la traduction on distinguera deux ordres de régularités : unilingues, interlingues. Tout le problème est de savoir où les chercher et, une fois captées, comment les calculer. »

Ces régularités s'observent sur deux dimensions transversales, à l'intérieur de chaque langue et dans le passage d'une langue à l'autre. Plus précisément, ce que révèlent ces deux dimensions, ce sont *deux formes* de régularités, deux types différents de contrainte qu'il est illusoire de vouloir réduire à un même dénominateur :

- d'une part, les régularités « unilingues » évoquées par Bourquin incluent des systèmes de *règles* phonologiques, graphémiques, morphologiques, syntaxiques qui imposent des déterminations *nécessaires*.

- d'autre part, les régularités « interlingues » ne dérivent pas de règles directement appliquées au niveau des formes linguistiques, mais découlent de correspondances établies au niveau des objets extralinguistiques désignées par ces formes : du fait de cette médiation, les régularités observables ne se ramènent pas à des prescriptions facilement systématisables sous la forme de contraintes plus ou moins obligatoires. On observe plutôt des tendances, des corrélations, des habitudes, des *régularités* émergentes sur un plan statistique.

Bien sûr, le clivage entre *règle* et *régularité* ne partage pas de manière nette l'unilingue et l'interlingue : les régularités apparaissent en grand nombre à l'intérieur d'une seule langue, et certaines formes de traduction sont basées sur l'application de règles déterministes.

Ce qui se dessine à travers cette opposition, ce sont les orientations de deux *paradigmes*, qui ont inspiré des modélisations différentes des phénomènes traductionnels : les modèles basés sur les règles et les modèles probabilistes.

I.2.2.1 La Traduction automatique basée sur les règles (TABR)

La Traduction automatique basée sur les règles est historiquement la plus ancienne. Elle s'est appuyée sur des formalismes informatiques (théorie des langages formels, théorie des automates, théorie de la compilation, etc.) et les développements de l'intelligence artificielle : représentations arborescentes, programmation logique, graphes conceptuels, etc. On distingue généralement deux approches.

I.2.2.1.1 L'approche directe

L'approche directe concerne les systèmes dits de *première génération*. Le premier de ces systèmes remonte au début des années 50, le « *Georgetown Automatic Translation* » élaboré à l'Université de Georgetown. Ces systèmes sont conçus pour un couple de langues donné, dans un sens donné. Les tâches d'analyse et de génération n'y sont pas séparées. L'unité de traduction est le mot (on y inclut aussi les expressions figées), et c'est l'environnement immédiat qui conditionne l'application des règles de remise en ordre syntaxique. Certains systèmes à approche directe sont encore en fonction aujourd'hui (p. ex. Systran).

Une telle méthode soulève des problèmes épineux :

- *la traduction est basée sur le mot à mot*, même si des règles *ad hoc* permettent de reconnaître les expressions polylexicales et de résoudre certaines ambiguïtés. Les relations structurales profondes étant ignorées, le calque syntaxique de la langue source aboutit souvent à des incohérences dans la langue cible.

- *manque de généralité* : les données ne sont pas réutilisables pour un autre couple de langues (pour traduire entre 10 langues européennes il faut donc concevoir 90 systèmes différents !). En outre, pour un même couple, les règles sont orientées dans un seul sens.

– *non-consistance des règles* : les règles n'étant pas ou peu hiérarchisées, la multiplication des règles *ah hoc* peut provoquer des conflits, « surgénérer » des constructions aberrantes, etc., ce qui rend l'enrichissement et la maintenance du système hasardeux et difficiles.

I.2.2.1.2 L'approche indirecte

Après le désenchantement consécutif au rapport ALPAC, qui mettait un sérieux frein aux ambitions de l'approche directe (et de la TA en général), on a pu constater un regain d'intérêt à partir de 1975, où la sophistication technologique, tant logicielle que matérielle, permettait l'élaboration de modélisations plus poussées. Les systèmes dits « de deuxième génération » ont cherché à pallier les principaux défauts de l'approche directe, tels que l'imbrication des grammaires et des moteurs d'inférence, la non-séparation des tâches d'analyse et de génération, etc.

Le principe directeur des systèmes de deuxième génération est la *modularité*, tant au plan des programmes que des données linguistiques. L'approche indirecte consiste à passer par une représentation intermédiaire faisant office d'interface entre langue source et langue cible, afin de ne développer, pour chaque langue, qu'un seul module d'analyse et un seul module de génération, quel que soit le couplage envisagé. On prend soin en outre, de séparer la partie algorithmique des données. C'est le même moteur d'inférence qui fonctionne pour tous les couples : seuls changent les ressources linguistiques, *id est* les dictionnaires et / ou les grammaires.

Suivant le degré de profondeur (ou d'abstraction) de cette représentation intermédiaire, on distingue deux types de systèmes indirects.

I.2.2.1.2.1 Pivot ou interlangue

L'approche par langue pivot ou interlangue se base sur une représentation indépendante de la langue source comme de la langue cible. Cette représentation est située sur le plan conceptuel. Différents formalismes ont été proposés pour ce genre de représentations. Le CETA de Grenoble (1961-1971) a d'abord exploré cette voie : la représentation adoptée était basée sur des structures syntactico-sémantiques s'inspirant de

la grammaire de dépendance de Tesnière. Aujourd'hui, les chercheurs du GETA étudient la possibilité d'interfacer les représentations d'Ariane-G5 (Boitet, 1997 ; Etienne Blanc, 2000 – ces représentations utilisent des arbres décorés rassemblant des informations des trois niveaux syntagmatique, syntaxique et logico-sémantique), avec l'interlangue développée dans le cadre du projet UML (pour *Unified Modelling Language*). Plus récemment, le système KBMT-89 (Sergueï Niremburg, 1989), développé au CMT (*Center for Machine Translation, Carnegie Mellon University*), possède un langage pivot basé sur 1 200 unités lexicales et une ontologie de 1 600 concepts (Boitet, 1993 : 117). Des modèles génériques, tels que les graphes conceptuels (Sowa, 1984) sont particulièrement adaptés (Zinglé, 1993). Notons enfin que certaines langues, présentant des caractéristiques morphologiques intéressantes, ont été proposées comme interlangue : par exemple l'aymara (parlé au Pérou ou en Bolivie) ou encore l'espéranto, pour lequel « dans pratiquement tous les cas la fin d'un mot détermine la partie du discours (…) » (C. Minnaja & L. Paccagnella, 2000).

Dans les systèmes TA à langage pivot, le module d'analyse doit extraire une représentation conceptuelle à partir du texte source, tandis que le module de génération part d'une telle représentation pour élaborer un énoncé dans la langue cible.

Dans le même esprit, le projet UNL[107], coordonné par l'*Institute of Advanced Studies* de l'Université des Nations Unis de Tokyo, est dédié à l'élaboration d'un standard universel de représentation des connaissances. Dans un entretien avec M. Lebert[108], Boitet décrit les objectifs poursuivis :

> « Il s'agit non de TAO (traduction assistée par ordinateur) habituelle, mais de communication et recherche d'information multilingue. Quatorze groupes ont commencé le travail sur douze langues (plus deux annexes) depuis début 1997. L'idée est de :
> - développer un standard, dit UNL (Universal Networking Language), qui serait le HTML du contenu linguistique,
> - pour chaque langue, développer un générateur (dit " déconvertisseur ") accessible sur un ou plusieurs serveurs, et un " enconvertisseur ". »

[107] Le site de ce projet est à l'adresse : http://www.unl.ias.unu.edu
[108] Entretien réalisé le 24/09/1998, disponible sur la Toile à l'adresse : http://ourworld.compuserve.com/homepages/mlebert/boitet.htm

Les 14 langues concernées pour le moment sont l'arabe, le chinois, le français, l'allemand, le hindi, l'indonésien, l'italien, le japonais, le letton, le mongol, le portugais, russe, l'espagnol, le russe et le thaï.

L'intérêt du recours à une représentation pivot est évident : pour traduire un texte du français vers les 10 autres langues officielles de la CE, une seule analyse du texte original suffit. En outre, les problèmes d'interprétation, d'ambiguïté, etc. sont concentrés au niveau de cette seule première phase. Lorsque l'analyse, ou l'« enconversion », est correctement effectuée, le plus dur est fait : pour les raisons de fermeture déjà évoquées, la génération est une étape théoriquement automatisable. Après analyse, un texte peut être traduit automatiquement dans autant de langues qu'on voudra. Ainsi l'intervention de l'humain (complète ou partielle) n'est requise que dans la première étape, suivant des modalités variables, dont Boitet (ibid.) donne quelques exemples :

> « L'enconversion n'est pas (si on veut de la qualité pour du tout venant) une analyse classique. C'est une méthode de fabrication de graphes UNL qui suppose une bonne part d'interaction, avec plusieurs possibilités : - analyse classique multiple suivie d'une désambiguïsation interactive en langue source, - entrée sous langage contrôlé, - encore plus séduisant (et encore pas clair, au niveau recherche pour l'instant), entrée directe via une interface graphique reliée à la base lexicale et à la base de connaissances. »

Bien entendu cette approche impose d'ignorer délibérément certaines composantes de la communication (par exemple les connotations, les aspects stylistiques, etc.) difficilement représentables dans un langage universel. La méthode est donc adaptée à des textes à dominante conceptuelle. Cependant, même si on se limite au seul niveau conceptuel, il est très difficile de construire une représentation assez générale et assez fine pour qu'il n'y ait pas altération du contenu d'un énoncé.

Enfin, même à l'intérieur d'un domaine très spécialisé, l'analyse automatique reste un objectif encore lointain. Pour le corpus de KBMT-89, limité à des manuels de PC, il a fallu malgré tout recourir au dialogue avec l'humain pour désambiguïser certaines expressions : le système a été doté d'un « *augmenteur* », posant des questions à un spécialiste ne connaissant que la langue source.

I.2.2.1.2.2 Transfert

L'autre type d'approche indirecte, basée sur le *transfert*, est dans l'immédiat moins problématique. La représentation intermédiaire ne vise pas l'universalité du conceptuel : elle reste une représentation linguistique, plus ou moins abstraite suivant le degré de profondeur choisi (généralement des structures syntactico-sémantiques). Elle nécessite, outre les modules d'analyse et de génération, un module spécialisé dans le transfert propre à un couple de langues donné. Comme pour les systèmes directs, le transfert impose une démultiplication des modules pour chaque couple impliqué. Par exemple, le système EUROTRA, concernant 9 langues européennes en 1982, devait inclure 72 modules de transfert. Il est clair que plus le nombre de langues impliquées est important, plus on a intérêt à « alléger » les modules de transfert.

Le schéma ci-dessous représente les deux types d'approche indirecte :

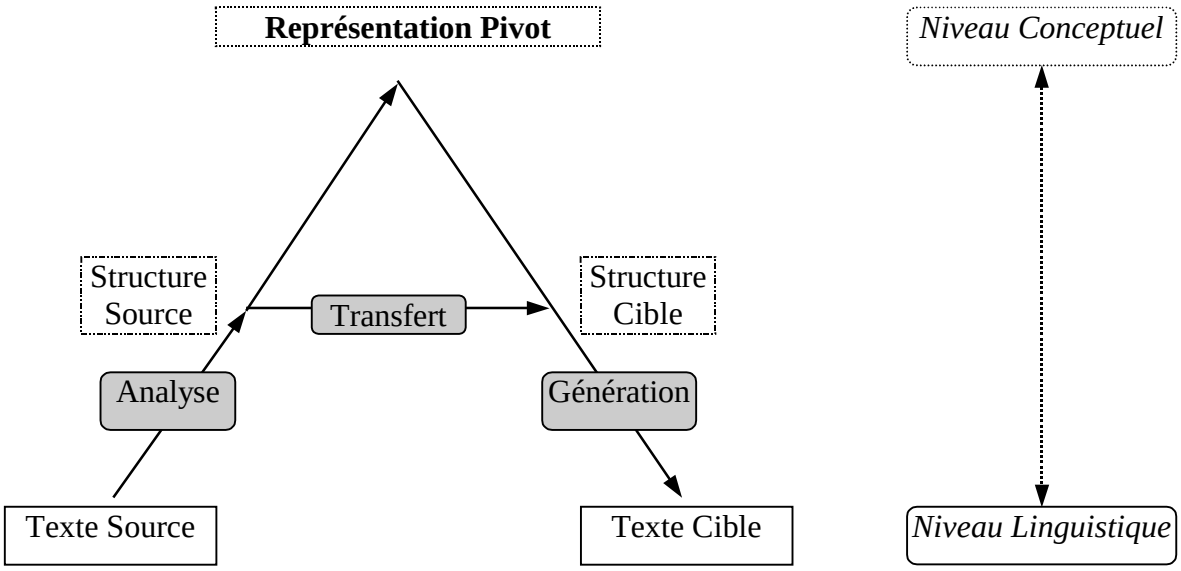


figure 14 : *approches indirectes*

La recherche d'économie, au niveau des ressources linguistiques, a orienté des chercheurs vers des modules d'analyse et de génération réversibles, suivant des algorithmes symétriques travaillant sur les mêmes grammaires et dictionnaires (L. Danlos & O. Laurens, 1991 : 110-111).

Notons que l'approche indirecte n'est pas exclusive aux systèmes de traduction assistée par ordinateur : des méthodes similaires ont été mises en œuvre afin de faciliter la traduction (humaine) d'un même texte en plusieurs langues. Pour traduire la Bible dans des langues « rares », pour lesquelles il était difficile de trouver un spécialiste maîtrisant les couples de langue impliqués, Nida (1969) propose d'effectuer une analyse profonde aboutissant à une représentation canonique formulée dans une langue tierce (en l'occurrence l'anglais). Certes, on ne voit pas bien ce qui prédisposerait telle ou telle langue naturelle à jouer le rôle de médium universel, mais l'intérêt d'une telle démarche est de permettre le partage du même travail d'explicitation et d'exégèse par des traductions différentes.

L'idée de Nida est de transformer les versets de la Bible en une série de « phrases noyaux » (« *kernel sentences* »), à la structure simplifiée[109], ne présentant pas d'ambiguïté lexicale ou syntaxique, où les tournures elliptiques sont explicitées et les références anaphoriques résolues :

> « Ces expressions restructurées sont simplement ce que de nombreux linguistes appellent des 'noyaux', c'est-à-dire les éléments structuraux de base à partir desquels la langue construit des structures de surface élaborées. En effet, une des intuitions les plus intéressantes de la 'grammaire transformationnelle' est le fait que toutes les langues se partagent une douzaine de structures de base, d'où sont issues les constructions plus élaborées par le moyen de ce qu'on appelle 'transformations' »[110] (ibid. : 39)

Nida (ibid. : 51)décompose son analyse en 5 étapes :

> « (1) identifier les propriétés structurales élémentaires de chaque mot, i.e. suivant les catégories objet, événement, qualité, et relation, (2) rendre explicite tout élément structural implicite afin de compléter les noyaux, (3) déterminer les noyaux de base sous-jacents aux structures de surface des phrases, (4) regrouper

---

[109] Nida distingue sept structures de base pour ses phrases noyaux : 1. *John ran quickly* ; 2. *John hit Bill* ;3. *John gave bill a ball* ; 4. *John is in the house* ; 5. *John is sick* ; 6. *John is a boy* ; 7. *John is my father*. Par ailleurs il organise la sémantique autour de 4 catégories principales (Nida, 1969 37-38): 1) l'objet (« *object* ») : choses ou entités 2) l'événement (« *event* ») : actions, procès, etc. 3) la qualité ou l'abstraction (« abstract ») : attributs, entités, degrés, etc.. 4) la relation, souvent exprimée par des conjonctions et des prépositions : c'est l'« expression des connections signifiantes entre les autres types de termes » ("the expressions of the meaningful connections between the other kind of terms.")

[110] "These restructured expressions are basically what many linguists call ''kernels'' ; that is to say, they are the basic structural elements out of which the language builds its elaborate surface structures. In fact, one of the most important insight coming from ''transformational grammar'' is the fact that in all languages there are half a dozen to a dozen basic structures out of which all the more elaborate formations are constructed by means of so-called 'transformations'"

> ces noyaux en ensembles reliés, (5) exprimer ces relations sous une forme qui sera optimale pour le transfert dans la langue cible. »[111]

L'exemple du tableau 4, tiré de la *Lettre aux Ephésiens* de Jean (2-8), illustre ce processus d'analyse (ibid. : 53-54). Les phrases noyaux de la deuxième colonne correspondent à l'étape 3 de l'analyse.

***Source* (*traduite en anglais*)**	***Phrases noyaux***
1-3 For by the grace	*1. God showed grace*
are ye saved;	*2. God saved you*
through faith;	*3. you believed* (*faith = event*)
4 and that not of yourselves;	*4. you did not save yourselves*
5 it is the gift of God;	*5. God gave it*
6 not of works	*6. you did not work for it*
7 lest any man should boast.	*7. no man should boast*

tableau 4 : *décomposition en phrases noyaux*

A l'issue de la cinquième étape, on obtient une formulation plus explicite, où les relations logiques entre les différentes assertions sont clairement formulées et où les problèmes d'interprétation sont censés être résolus (ibid. :54) :

> *God showed his grace to you, and in this way he saved you through your trusting in him. You yourselves did not save yourselves. Rather, God gave you this salvation. You did not earn it by what you did. Therefore no one can boast about what he has done.*

Cette forme « développée » est censée être plus adaptée au transfert. Cette démarche comporte certes une part d'arbitraire : on ne voit pas bien comment s'effectue le passage des phrases noyaux à la dernière phrase, et jusqu'où doit aller le travail d'explicitation. Mais le principe de la paraphrase explicite, comme représentation intermédiaire, est intéressant, et montre la généralité des principes de l'approche indirecte.

---

[111] "Identifying the basic structural element(s) of each word, i.e. object, event, abstract, and relational, (2) making explicit any implicit structural element which are required to complete the kernels, (3) determining the basic kernels which combine to constitute the surface structure of the sentence, (4) grouping the kernels into related sets, (5) stating these relationships in a form which will be optimal for transfer into the receptor language"

I.2.2.1.3 Problèmes soulevés par la TABR

La TABR a donné lieu à de nombreux développements mais il faut signaler quelques difficultés inhérentes à cette approche.

– *Difficultés intrinsèques aux représentations hiérarchiques*

La quasi-totalité des systèmes basés sur des règles utilise des représentations arborescentes des relations syntaxiques entre les constituants de la phrase. A ce sujet, Fuchs *et al*. (1993 : 111) énumèrent un certain nombre de difficultés liées à ce type de représentation :

1. *L'ellipse*. Le problème est de savoir s'il faut restituer les éléments manquants lors de l'analyse, et jusqu'où. Pour la phrase : *Quoiqu'aimable, il semblait fatigué*, faut-il restituer *Quoiqu'il fût aimable* ou *quoiqu'il semblât aimable* ? Sans compter les ambiguïtés qu'il faut parfois lever pour le rétablissement des formes effacées : dans *Pierre aime sa femme, et moi aussi* on peut comprendre aussi bien *moi aussi j'aime ma femme* que *moi aussi j'aime sa femme*, ce qui peut prêter à de fâcheuses confusions.

2. Les phénomènes d'adjonction tels que le « clivage » (p. ex. *c'est moi qui souligne*), *l'apposition* (*Elisabeth, ma femme, est venue*), *l'apostrophe* (*Elisabeth, viens !*), les *thématisations* (p. ex. *Le linguiste, lui, sa grammaire, il l'écrit*), *l'impersonnalisation* (p. ex. *Il faut noter que* ...). Dans tous ces cas, la structuration hiérarchique impose une tension entre les marques syntaxiques superficielles et le sémantisme.

3. Les phénomènes de *double portée*, difficilement représentables sous forme arborescente. Ainsi, dans la phrase *Il demandait encore du vin*, *encore* porte sur *demandait* autant que sur *vin*[112].

4. La *modalisation*, enfin, pose un problème de choix du gouverneur (ou tête de syntagme). Dans *je dois partir*, faut-il considérer *dois* comme un modifieur du

verbe *partir*, ou *partir* comme un complément du verbe *devoir* ? D'autant que nous avons affaire à un véritable continuum, à travers une gradation d'opérateurs modaux (au sens large) de poids plus ou moins fort : des auxiliaires (*être, avoir*), des opérateurs aspectuo-temporels (*je viens de..., je vais...*, etc.), des modaux (*devoir, pouvoir*, etc.), aux locutions verbales plus complexes (*il faut que..., j'ai le projet de ..., je suis sur le point de ...* etc.).

– *Complexité des grammaires*

D'après Pottier (1993 : 105), la faiblesse de nombreuses grammaires vient de ce qu'elles tombent dans le « piège des marques formelles ». Il critique par exemple la notion de « syntagme prépositionnel » usitée par l'école générativiste : « tout SN a une fonction casuelle, dont un des signifiants *peut* être dans certaines langues une préposition ». Ce qui importe donc c'est la fonction casuelle et non le marqueur formel. Pottier donne l'exemple de la phrase *il a réussi à son examen*, où le syntagme *à son examen* a fonction d'objet, malgré la préposition. Tandis que dans *il a réussi ce matin*, *ce matin* a bien une fonction de circonstant, sans toutefois l'apparition d'une préposition.

Les systèmes basés sur les règles doivent tenir compte de toutes ces particularités grammaticales que les grammaires classiques ignorent (dans la mesure où elles reposent sur l'intuition du locuteur). Or, cette complexité croissante compromet la consistance et la robustesse des systèmes. Comme le remarquent Fuchs *et al*. (1993 : 121) :

> « Le dilemme pour les formalismes syntaxiques est alors le suivant : soit on se limite aux constructions les plus régulières et on doit renoncer à une large couverture, soit au contraire on décrit dans la grammaire tous les cas de figure et alors, comme le démontre avec force J.P. Chanod (1992), on génère énormément de solutions "parasites" sur les phrases les plus banales (...) »

A ces problèmes de *complétude* et de *consistance* grammaticale viennent s'ajouter les difficultés liées à la *complexité algorithmique*. En intégrant un trop grand nombre d'hypothèses et de cas possibles, une phrase longue risque d'entraîner une explosion combinatoire rendant caduque toute tentative d'analyse complète.

---

[112] *encore*(*demander*(*vin*)) *vs* *encore*(*demander*) & *encore*(*vin*) *vs* *demander*(*encore*(*vin*))

I.2.2.2 La Traduction automatique basée sur l'exemple (TABE)

L'approche basée sur l'exemple, bien que plus simple dans ses principes que l'approche à base de règle, est sensiblement plus récente. Il a fallu attendre l'explosion du marché de la micro-informatique et la banalisation des textes sous format numérique, pour penser à tirer parti de la masse des traductions déjà faites. Le principe directeur de la TABE repose sur un constat très simple, ici énoncé par Isabelle (1992 : 726) : « La masse des traductions produites chaque année contient infiniment plus de solutions à plus de problèmes que tous les outils de référence existants et imaginables ! »

On attribue à Nagao la première inspiration dans ce sens. Il expose, en 1984, le concept de « traitement par analogie » :

> « L'homme ne traduit pas une phrase simple en faisant une analyse linguistique profonde. L'homme traduit plutôt en décomposant d'abord correctement une phrase donnée en fragments, en traduisant ensuite ces fragments dans l'autre langue et en assemblant finalement correctement les traductions des fragments en une longue phrase. La traduction de chaque fragment se fera par le principe de traduction par analogie avec des exemples corrects comme référence. (...) » (cité par Somers, 1993a : 153)

Ces fragments, d'un niveau inférieur à la phase, évoquent ce que nous appelons des *unités de traduction* (nous abrégerons désormais par UT). De nombreux travaux (E. Sumita & Y. Tsutsumi, 1988 ; S. Sato & M. Nagao, 1990 ; P. Brown *et al*. 1990 ; B. Collins *et al*., 1995) se sont inspirés de ce concept de traduction par analogie. Examinons rapidement les grandes lignes d'une telle approche.

– *Processus*

On peut critiquer les aspects psychologiques de la description de Nagao, mais elle a le mérite d'esquisser, dans les grandes lignes, les étapes d'un processus de TABE :

- segmentation de la phrase en unités de traduction.
- recherche des fragments *analogues* dans la Mémoire de traduction (MT). Concrètement, cela implique le développement d'une mesure de similarité, permettant de comparer des fragments proches mais non identiques (cf. S.

Niremburg *et al*., 1993). Pour chaque fragment trouvé, la mémoire fournit les équivalents traductionnels.

– recomposition des fragments traduits pour obtenir la traduction de la phrase en entier (cf. Sato, 1995).

Ce modèle soulève cependant des problèmes épineux :

– selon quels critères définir l'analogie : syntaxiques, sémantiques ?
– l'exhibition d'une phrase analogue peut éventuellement être indicative pour un traducteur humain, comme modèle de traduction, mais comment s'en servir pour obtenir automatiquement une traduction exacte du fragment original ?
– que faire lorsqu'un fragment n'a pas d'identique ni d'analogue en MT ? Et que faire si celle-ci en donne plusieurs traductions possibles ?

Par ailleurs, les étapes de fragmentation et de recomposition restent problématiques : suffit-il de découper, traduire, puis recoller pour obtenir une phrase correcte en langue d'arrivée ? Un bref examen nous révèle que cette idée est bien naïve. Les unités de traduction, telles que nous les avons définies, ne peuvent fournir des fragments de phrase suffisants pour la recomposition : d'une part une UT peut être formée d'unités discontinues, d'autre part la recombinaison des UT dans la phrase cible impose de se conformer à des règles de composition morphosyntaxique très différentes des règles de composition de la phrase originale. Il apparaît donc que les UT ne peuvent être considérées comme des fragments qu'il suffirait de traduire et coller bout à bout, dans l'ordre des unités source, pour obtenir une phrase compréhensible dans la langue d'arrivée. Les seuls fragments de phrase qui pourraient avoir une telle autonomie sont les phrases elles-mêmes, à moins de supposer que les grammaires des langues source et cible sont pratiquement identiques. En outre, l'autonomie d'un fragment (c'est-à-dire sa vocation à former une unité structurale indépendante, juxtaposable à d'autres fragments du même type) est une propriété dépendante d'un contexte spécifique, et non une propriété intrinsèque à ce fragment. Par exemple, le fragment *la TAO aura du succès* peut certes apparaître comme une phrase autonome, vis-à-vis de la traduction ; mais il peut aussi apparaître comme subordonnée dans une phrase plus grande : *il sera riche quand la TAO aura du succès*.

Dans ce dernier contexte, le même fragment, qui n'a plus exactement la même signification, peut être traduit différemment, par exemple en anglais, où les règles de concordance temporelle sont différentes. De manière générale, même le niveau de la phrase ne garantit pas l'autonomie vis-à-vis de la traduction. Gaussier, Hull & Aït-Mokhtar (in Véronis, 2000 §13) critiquent l'hypothèse de l'autonomie phrastique : « Il y a un certain danger à supposer que la phrase est l'unité de traduction. Dans de nombreuses situations (sinon la majorité), le contexte déterminant est absent de la phrase à traduire et peut être trouvé ailleurs dans le document. » [113]

Tout le problème de la démarche proposée par Nagao se résume à ce dilemme : si l'on retient des fragments assez grands pour pouvoir être traduits de manière autonome, la chance d'avoir à réutiliser ces fragments est très faible ; si au contraire, on recherche des fragments plus petits, et donc plus fréquents, il devient impossible de les traduire de manière autonome.

Pour sortir de cette contradiction, certains auteurs ont proposé des solutions originales : par exemple, Christos Malavazos *et al*. (2000) ont développé une technique d'extraction de « modèles » de traduction (« *translation templates* »), c'est-à-dire des couples de phrases comportant des parties constantes et des parties variables. Lors d'un processus d'apprentissage, la comparaison deux à deux de couples de phrases alignées permet de confronter les parties communes et les éléments qui commutent de chaque côté, et d'établir des équivalences : d'une part les parties communes (qui forment en quelque sorte des phrases à trou) sont enregistrées comme modèles, avec les paradigmes correspondants à chaque variable, et d'autre part des équivalences sont enregistrées au niveau des unités de traduction qui commutent au niveau de ces variables. Les auteurs donnent l'exemple suivant (les parties variables sont en italiques, cf. Malavazos *et al*., 2000 :1.3) :

angl.: Style Manager *help* menu
grec : Καταλογος *βοηθειας* διαχειριτη υφους

[113] "There is a certain danger with assuming that the sentence is the unit of translation. Many (if not most) situations, important context is missing from the sentence being translated and can been found elsewhere in the document."

+
angl.: Style Manager *file* menu
grec : Καταλογος *αρχειων* διαχειριτη υφους

⇒ modèle : Style Manager X menu ↔ Καταλογος X' διαχειριτη υφους
X = *help* ↔ X'= *βοηθειας*
X = *file* ↔ X'= *αρχειων*

La prise en compte des variations locales permet ainsi d'enregistrer des exemples plus généraux, susceptibles d'apparaître à nouveau (remarquons que pour aboutir à une généralisation, le modèle doit déjà avoir au moins deux occurrences dans le corpus d'apprentissage). En outre, les équivalences entre « unités de traduction »[114] sont toujours liées à leur contexte d'apparition, ce qui leur donne une plus grande pertinence. Lors du processus d'apprentissage, ces équivalences sont réutilisées dans l'extraction des modèles.

Après l'extraction des modèles, les auteurs proposent une méthode de TABE suivant un modèle descendant, cherchant la meilleure solution d'abord :

1. Recherche en mémoire de la phrase à traduire, telle quelle (« *Full matching* »).
2. Si elle n'est pas trouvée, recherche d'un modèle s'ajustant à la phrase (« *Template Matching* »). Les unités correspondant aux variables sont alors traduites séparément (« *Local Matching* »).
3. Si aucun modèle ne convient, recherche de la phrase la plus proche (« *Fuzzy Matching* »). Ensuite, les unités divergentes sont traduites séparément (« *Local Matching* »).

Cette méthode nous paraît très intéressante, car elle est basée sur le principe de commutation, que nous examinons plus loin (cf. chap. III.1.3.3). Mais dans l'évaluation décrite, elle n'assure qu'une faible couverture : là encore, ce genre d'approche ne peut fonctionner qu'avec des phrases redondantes par rapport à la mémoire de traduction.

[114] c'est le terme employé par les auteurs.

– *Conditions de mise en œuvre*

La répétition est en effet le problème majeur de la TABE : on ne peut traduire que ce qui a déjà été traduit, ce qui est redondant avec les exemples présents en mémoire. Le succès de l'opération de traduction dépend donc entièrement des caractéristiques de la MT, et de son homogénéité avec le texte à traduire. Pour qu'un système de TABE soit fonctionnel, la MT et les textes à traduire doivent être en relation de fermeture mutuelle. Autrement dit, les textes à traduire doivent comporter le moins de nouveauté possible par rapport aux textes constituant la base d'exemples. Le cas se présente, par exemple, lors de la retraduction d'une documentation technique après mise à jour. Gaussier, Hull & Aït-Mokhtar (in Véronis, 2000 §13) citent une étude réalisée par la société Xerox : « Une étude de cas a montré que le système de MT renvoyait des correspondances exactes pour 70 % des phrases du manuel d'une automobile, en utilisant une mémoire de traduction issue du manuel d'un tout autre modèle du même constructeur. »[115]

La redondance des traductions à venir par rapport aux exemples codés en mémoire est la condition principale de la viabilité de ce type d'approche.

I.2.2.2.1 Exemple d'architecture

Bien entendu, la taille de la MT constitue un facteur de réussite : plus elle contient d'exemples, et plus son spectre de réponse est large. Les informations stockées en mémoire étant par nature cumulatives, on peut tirer parti de cette caractéristique pour concevoir un fonctionnement cyclique où les nouvelles traductions viendraient enrichir la mémoire de nouveaux exemples, et en augmenter progressivement la couverture.

Nous avons schématisé figure 15, un exemple d'architecture de TABE incluant une phase humaine de post-édition, où le principe d'augmentation progressive est mis en œuvre dans le cycle de traduction. Le processus comporte 5 étapes :

1. le texte à traduire est segmenté,
2. des segments identiques (ou ressemblants) sont recherchés en mémoire,

[115] "One case study within Xerox found that the TM system returned exact matches for 70° % of the sentences in an automobile manual, using a translation memory derived from the manual of an entirely different model from the same manufacturer."

3. les traductions des segments sont recomposées,
4. le résultat est révisé par un post-éditeur,
5. le texte à traduire et la traduction révisée viennent alimenter la mémoire de nouveaux exemples.

Les étapes 1 et 2 peuvent être simultanées, si la segmentation est conçue comme une opération itérative opérant sur des fragments de plus en plus petits tant qu'aucun exemple n'a été trouvé. Dans l'opération 5, il est important que le réviseur puisse fournir une version correctement segmentée de la traduction correcte.

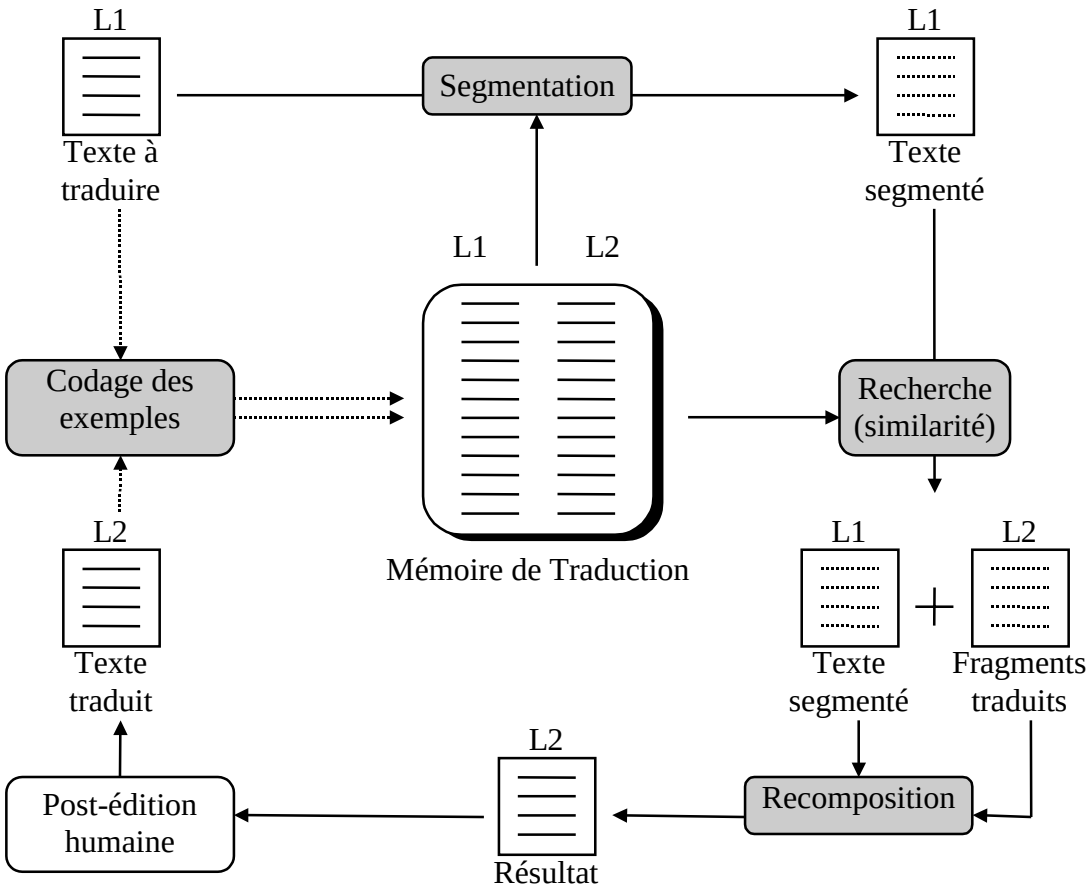


figure 15 : *architecture d'un système à augmentation progressive*

Notons qu'une telle architecture est symétrique vis-à-vis du couple de langues : la traduction peut être effectuée dans les deux sens.

I.2.2.2.2 Constitution de la mémoire de traduction

On a vu le rôle central que joue la MT dans ce type de démarche. Tout repose sur la richesse, la pertinence et la structuration des informations qu'elle recèle. La constitution d'une mémoire de traduction s'articule autour des étapes suivantes :

- *rassemblement et sélection du corpus* de textes traduits qui serviront d'exemples.
- *alignement* des textes bilingues parallèles. Il s'agit d'apparier les zones correspondantes entre les deux versions du corpus. L'alignement peut être réalisé entre des points (coordonnées textuelles), des syntagmes, des phrases, des paragraphes, etc. Il nécessite en général une opération de *segmentation* basée sur des marqueurs de surface (balises, ponctuation, marques typographiques, etc.). Pour la constitution massive de corpus alignés, il est crucial de recourir à des méthodes d'alignement automatique.
- *mémorisation des exemples*. Les exemples peuvent éventuellement être enregistrés sous des formats riches, comportant des informations lexicales, morphologiques, syntaxiques ou sémantique. Là encore des méthodes automatiques peuvent être mises en œuvre avec profit (comme l'étiquetage morphosyntaxique, etc.). Etant donné le volume d'information attendu, le mode d'indexation a son importance pour autoriser un accès efficace aux informations pertinentes.
- *extraction d'informations traductionnelles*. A partir des exemples codés en mémoire, il est possible d'extraire automatiquement des informations complémentaires, utiles au fonctionnement de la MT. Par exemple, il peut être intéressant d'identifier des unités de traduction de part et d'autre, et d'établir des correspondances entre ces unités : ce type d'extraction permet d'assurer la génération et la mise à jour de glossaires bilingues intéressant les traducteurs comme les terminologues. En outre, il est possible d'élaborer des « modèles » de traduction, comportant des parties variables. Enfin, l'analyse des exemples de traduction peut également fournir des correspondances au niveau syntaxique, par

l'extraction d'équivalences (plus ou moins fréquentes, plus ou moins générales, plus ou moins complètes) entre des schémas lexico-syntaxiques élaborés.

C'est un champ d'étude très vaste qui est en train de s'ouvrir avec le développement des techniques dédiées à ces différentes tâches. Dans les deuxième et troisième parties de ce travail, nous nous concentrerons sur la mise en œuvre et l'évaluation de méthodes statistiques simples appliquées aux deux dernières étapes, c'est-à-dire l'alignement de segments équivalents et l'extraction de correspondances entre unités de traduction.

I.2.2.3 TABE vs TABR : point de vue heuristique

D'un point de vue général, les deux approches, basée sur les règles ou basée sur l'exemple, procèdent d'heuristiques différentes.

– *Niveau de description*

Par rapport aux formalismes logiques de la TABE, la démarche basée sur l'exemple est avant tout empirique. Les grammaires formelles sont construites à partir de règles entretenant un rapport dialectique avec la réalité linguistique : leur établissement s'inspire de données empiriques mais leur mise en œuvre est *a priori*. Elles constituent un modèle artificiel des transformations traductionnelles. A l'inverse, l'extraction d'informations basées sur l'exemple s'appuie sur l'observation *a posteriori* de phénomènes concernant le résultat de traductions humaines. Ce changement de perspective a des conséquences importantes : tandis que pour la TABR la *qualité* de la traduction est un objectif, pour la TABE c'est d'abord une donnée, qui concerne les caractéristiques linguistiques et formelles des traductions qui alimentent le système.

En outre l'activité traduisante étant gouvernée par un mixte de liberté et de contraintes, la stratification des choix de traduction n'autorise pas, dans la réalité, le transfert systématique d'une forme linguistique à une autre : de fait, tandis que la TABR recherche l'établissement des *meilleures* règles, le point de vue basé sur l'exemple vise à capter les régularités les *plus significatives*, pour ensuite essayer de les généraliser. Quand la TABR se situe de plain-pied dans le paradigme du *transcodage*, la TABE part de la

traduction au sens large, pour déterminer, de manière inductive, les conditions d'un transcodage. Ce passage de la traduction au transcodage est sans garantie : pour que la réduction soit possible, il faut atteindre un seuil critique au niveau de la masse des données empiriques qui alimentent le système, et l'on ne peut déterminer ce seuil *a priori*.

– *Mise à jour*

Dans le traitement des informations linguistiques, il faut insister sur une différence fondamentale concernant la mise à jour des systèmes : tandis que les systèmes de TABR sont relativement rigides et supportent mal l'insertion de nouvelles règles (Somers, 1993a), le principe cumulatif d'une MT autorise, et recommande, l'adjonction d'exemples nouveaux.

– *Robustesse*

La nature du fonctionnement analogique implique une certaine résistance au « bruit » généré par les incorrections textuelles (coquilles, fautes de frappe, fautes d'orthographes, incorrection grammaticale, etc.), dans la mesure où le critère de *similarité* dessine par définition un continuum : la non-reconnaissance d'une seule unité n'empêchera pas l'identification de l'exemple le plus proche (cette propriété est désignée en anglais par l'expression « *gracious degradation* ».

– *Séparation analyse / génération vs symétrie*

La structure même de la MT autorise un usage symétrique des exemples qui y sont codés : le sens de traduction est indifférent. Certains systèmes formels sont basés sur des règles réversibles, mais la plupart du temps les modules d'analyse et de génération sont séparés et nécessitent la réécriture de grammaires différentes.

– *Fermeture linguistique vs Fermeture phrastique*

Nous avons introduit le concept de *fermeture linguistique* pour désigner la contrainte pesant sur des usages linguistiques conformes à des systèmes morphosyntaxiques et lexicaux fixés à l'avance : ce type de fermeture définit les limites du champ d'application

des systèmes basés sur les règles. Pour les approches à base d'exemple la contrainte de fermeture est beaucoup plus forte, car le succès dépend de l'existence préalable, en mémoire, de phrases entières ou du moins de fragments autonomes pour la traduction. Cette forme de contrainte, que nous désignerons par le terme de *fermeture phrastique*, a pour conséquence de limiter sévèrement le champ d'application des systèmes de TABE : un haut niveau de répétition et de stéréotypie est requis.

– *Processus séquentiel vs Processus parallèle*

Un des points forts des approches basées sur l'exemple concerne la calculabilité. Kitano (cité par Somers, 1993a : 161) remarque que la recherche d'exemples en mémoire est plus propice à un traitement parallèle :

> « Dans le modèle basé sur la mémoire, l'application séquentielle des règles [comme dans les analyses syntaxiques] est éliminée et remplacée par un processus massivement parallèle de recherche dans la mémoire. De plus, la structure syntaxique et l'interprétation sont préindexées pour éviter ou minimiser suffisamment l'opération coûteuse d'unification. »

De cette manière, la complexité temporelle qui pénalise parfois les systèmes à base de règles peut être convertie en complexité spatiale (taille de la MT), et répondre avantageusement aux problèmes d'explosion combinatoire.

I.2.2.4 Les systèmes mixtes

Il est important de noter que les approches à base de règles et à base d'exemple ne sont pas incompatibles. Au contraire les deux perspectives sont complémentaires, comme le montrent certaines expériences unissant TABE et TABR.

Par exemple, E. Sumita & H. Iida (1992) ont développé un système original intégrant règles et exemples. Ils partent du constat que, pour certains problèmes spécifiques, l'usage de règles transformationnelles est problématique dans la mesure où la traduction ne se fait pas de manière compositionnelle à partir des mots de la langue cible. Dans leur système, l'entrée à traduire est d'abord analysée suivant les méthodes classiques d'analyse syntaxique. Si un cas est reconnu problématique (par ex. un cas d'ambiguïté syntaxique) et si l'estimation du coût des calculs est trop élevée, le système recourt à une base d'exemple. La recherche d'un exemple dans la MT se fait au moyen d'une mesure de similarité basée

sur un thesaurus (où les distances sémantiques sont représentées par le nombre de paliers rencontrés dans la relation d'hypo-/hyperonymie).

Dans une optique similaire, O. Furuse & H. Iida (1992), ont développé une méthode basée sur l'exemple pour un système de TA utilisant l'analyse syntaxique. Leur système est totalement hybride dans la mesure où les deux approches sont parfaitement intégrées. Le stockage des exemples se fait suivant trois différents niveaux d'abstraction. Pour chaque exemple on mémorise a/ la chaîne littérale ; b/ un modèle général avec des variables ; c/ des règles syntaxiques correspondant à différentes stratégies de traduction. Cette approche peut apparaître comme un prolongement de l'extraction des modèles de traduction, proposée par Malavazos *et al*. (2000). Voici une illustration de chacun de ces niveaux :

Source en japonais	***Traductions possibles en anglais***
(*a*) *sochira ni okuru*	*will he send it to you ?*
(*b*) *X o onegai shimasu*	*may I speak to X*' (*X=jimukyoku* / *office*) *please give me the X*' (*X=bangoo* / *number*)
(*c*) *N1 N2 N3*	*N3' of N1 '* *N1= kaigi* / *meeting,* *N2= Kaisa* / *opening* *N3= kilan* / *time* *N2' N3' for N1'* *N1=sanka* / *participation* *N2=mooshikomi* / *application* *N3=yooshi* / *form*

tableau 5 : *trois niveaux d'exemple dans un modèle mixte*

Ces connaissances sont utilisées de façon hiérarchique et coopérative : dans la mesure du possible, le système essaie de traduire sur la base des exemples ; en cas d'échec, c'est l'analyse qui prend le relais. Ici, l'analyse intervient lorsque la recherche analogique à échoué, à l'inverse du modèle de Sumita & Iida. Si l'entrée n'existe pas telle quelle (niveau a), on cherche un modèle s'approchant (niveau b), et si aucun modèle n'est

satisfaisant on lance les transformations au niveau de patrons syntaxiques abstraits (niveau c). On privilégie donc l'information issue des exemples par rapport au traitement analytique.

Mais les règles du niveau c ne sont pas purement syntaxiques. Ces règles sont choisies par substitution des mots sémantiquement similaires. Par exemple : *shukuhaku mooshikomi yooshi* donne *application form for accommodation.* Cette approche a le mérite de subordonner le choix de la structure syntaxique au contenu sémantique.

Le système décrit par Watanabe (1992) s'inspire des mêmes idées : les exemples de traduction sont stockés sous forme de graphes acycliques étiquetés (ou arbres de dépendances). Les règles de transfert consistent en la donnée de paires d'arbres (les exemples) accompagnés de règles de correspondance entre les nœuds. La recherche en mémoire de traduction se fait par un calcul de similarité permettant de faire correspondre une entrée avec des exemples existants. La similarité est issue de la composition d'une mesure de distance structurale et d'une mesure de distance lexicale. La sortie est alors obtenue par unification des structures correspondantes.

Sous certains rapports, les MT de ces systèmes sont aux systèmes de règles ce que les dictionnaires de formes fléchies sont aux systèmes de règles de flexion : dans les deux cas, on convertit la complexité en temps (de calcul) en complexité en espace (mémoire). En effet dans les dictionnaires de formes fléchies la génération a déjà été faite, et l'on se contente d'en stocker le produit indexé par chaque lemme. Dans les systèmes mixtes le transfert des structures a déjà été effectué, et celles-ci sont indexées par similarité. Les problèmes algorithmiques sont différents : on ne se soucie plus de l'application récursive de règles de transformation, mais les méthodes de recherche et d'indexation deviennent centrales. En outre, le problème de la *consistance* des règles est remplacé par celui de la *complétude* de la mémoire de traduction – les exemples structurés (arbres étiquetés) devant être choisis en vue d'assurer la meilleure couverture des phrases ultérieurement soumises à la traduction.

D'un point de vue heuristique, il faut retenir l'articulation modulaire et hiérarchique de ces méthodes mixtes : lorsqu'une tâche se révèle trop complexe pour un système

d'inspiration logique, on débraye en s'appuyant sur des principes analogiques ; inversement, si l'analogie ne donne rien, les règles peuvent peut-être fournir des solutions de repli. En fonction du problème, on utilise en priorité la méthode la plus sûre et la plus simple – puis on raffine, par l'adjonction de techniques complémentaires.

I.2.3 Bi-textes, aide à la traduction et applications dérivées

Si les mémoires de traductions fournissent une alternative intéressante aux approches classiques de la TA, les applications les plus nombreuses (et pour le moment les plus réalistes) des corpus bi-textuels se situent dans le champ élargi de l'aide à la traduction. Les bi-textes renferment une mine d'informations intéressantes pour la constitution d'outils simples, conformes à ce que Bar-Hillel (1964 : 183) appelait « un usage modeste et judicieux des instruments d'aide automatiques »[116].

Entres autres usages « modestes et judicieux », les bi-textes peuvent avoir des applications concrètes dans des domaines divers :

– *Lexicologie, lexicographie et terminologie bilingues*

Nous verrons (cf. partie III) qu'il est possible, à partir d'un bi-texte, d'extraire automatiquement des correspondances sur le plan lexical : ces données brutes, même si elles contiennent un bruit certain, peuvent fournir des informations de premier ordre pour les lexicologues, les lexicographes, ou encore les terminologues.

Avec des projets d'envergure comme la rédaction du *Collins Cobuild Dictionary*[117], les corpus électroniques sont promis à devenir une composante essentielle dans la rédaction des dictionnaires : ils permettent de donner une assise empirique aux définitions en liant le sens à l'usage, l'intuition du lexicographe n'étant plus, et ne devant plus être, le

[116] "a judicious and modest use of mechanical aids".

[117] Depuis le début des années 80, le projet Cobuild (Sinclair, 1987) vise à la constitution d'un corpus de grande dimension, la *Bank of English*, comportant à l'heure actuelle quelques 320 millions de mots de textes anglais, incluant des manuels, des romans, des journaux, des guides, des magazines et 20 millions de mots de discours oral retranscrit (cf. le site officiel : http://titania.cobuild.collins.co.uk/about.html). Dans une perspective de linguistique empirique, ce corpus est dédié à fournir les informations pour l'étude de différents aspects de l'usage : sens des mots, grammaire, pragmatique, idiome, etc.. La *Bank of English* comporte un étiquetage morphosyntaxique, et une partie a été automatiquement parsée avec la collaboration de Fred

seul guide. En outre, la publication des dictionnaires sous la forme de CD-ROM, permet d'enrichir les définitions de nombreux exemples, à l'intérieur desquels il est facile de naviguer. Ainsi, le dictionnaire bilingue *Oxford-Hachette English-French Dictionary* contient de nombreuses illustrations, avec un co-texte minimal, des équivalences proposées. Ce dictionnaire a été rédigé sur la base d'un corpus anglais-français contenant plus de 10 millions de mots chacun (Grundy, 1996 ; Knowles, 1996) ; particulièrement riche en exemples, il n'a cependant pas été construit à partir de textes alignés : on peut imaginer que cela aurait pu apporter un surcroît d'information, chaque exemple pouvant donner accès à sa contrepartie, avec le co-texte, dans l'autre langue.

En ce qui concerne l'étude de la traduction des termes, on utilise les mêmes techniques que pour l'extraction des correspondances lexicales, couplées à des méthodes d'identification de composés terminologiques, sur la base de patrons syntaxiques (Béatrice Daille *et al*., 1994 ; Gaussier, Hull & Aït-Mokhtar in Véronis, 2000 § 13).

– *Le bi-texte comme Mémoire d'entreprise*

Isabelle (1992) note que le volume annuel des traductions effectuées au Canada dépasse le demi-milliard de mots : si les traducteurs avaient accès à cette masse gigantesque de traduction, ils y trouveraient un véritable gisement d'exemples illustrant des problèmes concrets.

Ainsi, dans la mesure où la plupart des outils de référence contiennent des informations d'ordre général, il peut être plus intéressant pour un traducteur de chercher ses informations dans un corpus de traduction adapté à son sujet que d'utiliser un dictionnaire standard. La possibilité d'extraire des « concordanciers à usages divers » permettant d'extraire des occurrences avec leur co-texte immédiat est au cœur d'une recherche efficace. Eliot Macklovitch (1992) montre comment, en partant d'une base de données contenant des extraits du *Journal des débats* de la Chambre des communes du Canada, un tel concordancier peut donner tout un éventail de solutions pour la traduction d'une expression argotique telle que *to be out of lunch*, sans véritable équivalent en français. Comme le suggère Macklovitch (1993a : 6), « en alignant automatiquement les textes sources et leur traduction, nous pouvons transformer les archives d'un service de

Karlsson de l'université d'Helsinki.

traduction en une mémoire d'entreprise interactive qui, à la différence de nos collègues humains, n'oublie jamais. » Le bi-texte peut devenir en quelque sorte un format de base de données traductionnelle permettant, à la différence des dictionnaires, de stocker des informations concrètes en rapport avec la pratique réelle du traducteur.

– *Vérification des traductions*

D'autre part, l'alignement d'un texte et de sa traduction permet d'appliquer des outils de vérification automatique signalant au réviseur d'éventuelles anomalies. Isabelle (1992 : 10) remarque que les meilleurs traducteurs ne sont pas à l'abri d'erreurs élémentaires comme l'emploi de faux amis tels que (angl.) *library* pour traduire (fr.) *librairie*, (angl.) *physician* pour traduire (fr.) *physicien* : « Même quand les traductions sont faites par les meilleurs traducteurs du Canada, il apparaît que la pression due à des délais trop courts rend difficile le contrôle des interférences linguistiques. »[118]

Pour éviter ce genre d'interférences le système Transcheck (Isabelle *et al*., 1993) détecte la présence de faux amis, d'emprunts et de calques jugés illicites. Les portions de texte « oubliées » par le traducteur sont automatiquement repérées par le système d'alignement. Pour l'avenir, Macklovitch prévoit d'autres extensions, comme la vérification de la traduction des expressions numériques, particulièrement fastidieuse pour le réviseur. Notons que des formalismes simples, comme les automates à états finis, donnent de bons résultats pour le repérage et la traduction locale de structures représentant des sous-systèmes relativement fermés, comme l'expression des nombres, des sommes d'argent, des dates, des noms propres, des noms d'institution, de titres officiels, etc. (C. Fairon & J. Senellart, 1999). Des modules de transduction spécialisés pourraient ainsi prendre en charge la traduction ou la vérification d'informations importantes, qu'elles soient disséminées dans le texte ou regroupées à l'intérieur de listes ou de tableaux.

Dans le même esprit, Isabelle (1992 : 729) propose d'automatiser la vérification de la cohérence terminologique :

[118] "Even though these translations are the work of some of the best translators in Canada, it appears that the time pressure under which they are produced makes linguistic interference harder to control."

> « On pourrait encore envisager la vérification automatique d'autres propriétés. Celle de la cohérence terminologique constitue un candidat possible. On imagine en ce cas un mécanisme capable d'explorer un bi-texte en y recherchant certaines classes de situations où un terme du texte de départ se voit associer plusieurs équivalents différents dans le texte d'arrivée. »

Macklovitch (1995a) a mis en œuvre ce genre de vérification au sein du système *Transcheck*. La tâche s'est avérée difficile, la cohérence terminologique n'étant pas incompatible avec une certaine variabilité due aux emplois anaphoriques, aux formes elliptiques et contractées, à la variabilité morphologique et syntaxique des termes, aux possibilités de paraphrases, etc. Ainsi, le système produit du bruit en signalant comme douteuses des traductions tout à fait licites : l'auteur montre cependant qu'un tel outil de vérification peut se révéler utile dans les domaines où la technicité impose une cohérence terminologique stricte, le bruit (environ une incohérence signalée sur trois correspond à une traduction correcte) n'étant pas assez important pour compenser le gain de temps apporté par l'automatisation. Il semble en outre que dans une version plus élaborée intégrant une analyse morphosyntaxique plus fine, le bruit pourrait être considérablement diminué.

Dans l'architecture intégrée d'une station de travail pour le traducteur, la constitution de bi-texte apparaît ainsi comme une pièce maîtresse, autour de laquelle s'articulent des outils variés : concordancier, vérificateur de traduction, base de données terminologiques, etc.

– *Dispositif de dictée pour les traductions*

La reconnaissance de la parole a accompli des progrès importants ces dernières années. Or la reconnaissance d'une traduction prise sous la dictée, la machine ayant connaissance du texte source, est un sous-problème plus facile à résoudre, le texte original tenant lieu de source d'information supplémentaire permettant de trouver des solutions aux ambiguïtés. P. Brown *et al*. (1992b :10) ont montré que l'incertitude au niveau des mots pouvait chuter de 63,61 à 17,2 lorsque le modèle de la langue cible est doublé d'un modèle (probabiliste) de traduction : « A la lumière de ces résultats, il est raisonnable d'espérer mettre en œuvre, avec une grande précision, la reconnaissance d'un discours fluide dont le

texte est contraint par un rapport de traduction avec une séquence déjà connue »[119]. C'est grâce au développement des corpus bi-textuels qu'on pourra extraire les jeux de paramètres nécessaires aux modèles probabilistes qui interviennent dans le fonctionnement d'une « *Dictation Machine for Translators* » (Isabelle, 1992).

– *Recherche d'information multilingue*

Les corpus alignés peuvent aussi se montrer utiles en recherche d'information multilingue (en angl. abrégé par CLIR), car ils constituent des réservoirs d'exemples de traduction de termes, à partir desquels on peut effectuer une traduction grossière des requêtes. Comme l'ont montré R. Brown, J. Carbonnel & Y. Yang (in Véronis, 2000 § 14), même légèrement bruyants, ces exemples de traduction permettent d'effectuer des recherches d'informations multilingues aussi performantes que dans le cas monolingue.

[119] "It is reasonable in view of these results to hope that high accuracy recognition of fluent speech is possible with present day speech technology when the text is constrained to be the translation of a known source language sequence."

I.3 Conclusion de la première partie

On constate que les applications potentielles des corpus bi-textuels sont nombreuses. Qu'on se situe dans une perspective de traduction automatique *stricto sensu* ou d'aide à la traduction dans le sens général, la position centrale de ces corpus, structurés et élaborés sous la forme de *Mémoires de traduction*, aboutit à un renversement de paradigme : on ne peut plus, pour des raisons d'implémentation ou de simplicité méthodologique, réduire la traduction au seul transcodage. Les bi-textes sont avant tout les produits d'une activité communicative complexe qui dépasse de loin les seules spécifications des codes linguistiques. L'étude détaillée de la traduction dans ses différents aspects pragmatiques, conceptuels, stylistiques et linguistiques nous a permis de montrer que la relation d'équivalence sous-jacente entre un texte et sa traduction n'est jamais *a priori* réductible à un réseau d'équivalence défini entre deux codes : le sens du message source, de même que le sens du message cible, se situe au-delà du langage, et même au-delà de l'ensemble des codifications culturelles qui structurent des communautés d'individus – il se situe dans la singularité des situations de communication.

Dans la démarche des méthodes classiques de traduction automatique, les *codes* préexistent aux *messages* : toute la difficulté consiste alors à recevoir *correctement* un message, dans une phase d'analyse, pour en recréer *correctement* un nouveau, dans une phase de génération, la notion de *correction* étant définie par la compatibilité des messages avec les codes linguistiques. La force, et la faiblesse, des systèmes basés sur les règles tient dans leur réductionnisme : en tant que modèle mathématisé des réalités langagières, leur fonctionnement est relativement contrôlable ; en tant que simplification de ces mêmes réalités, ils en intègrent difficilement la complexité, puisque la définition linguistique du transcodage met en jeu, nous l'avons vu avec la notion d'idiome, une succession de filtres difficilement systématisable.

Le paradigme bi-textuel renverse la donne : la pierre angulaire est désormais le *message*, ou du moins ce qu'il en reste, son résidu textuel. De ce message il faut par suite faire émerger des codes, puisque l'objectif poursuivi est de partir du singulier pour

atteindre le général, de réutiliser *ailleurs* – pour des messages nouveaux et des situations nouvelles, différentes ou homologues – le résultat de traductions déjà faites. Ce changement de perspective a des conséquences profondes :

- d'une part, il impose d'adopter, au moins dans un premier temps, un point de vue résolument empirique, en abandonnant toute tentation normative. Si l'on veut tirer profit de ces milliards de mots traduits chaque année, évoqués par Isabelle, on ne peut se contenter d'énoncer ce que la traduction *devrait* être : il faut partir de ce qu'elle *est*. Cela ne signifie pas qu'il faille renoncer à toute tentative de théorisation : c'est pourquoi dans la première partie de ce travail nous avons cherché à établir, dans les grandes lignes, une esquisse de théorie de la pratique traductionnelle.

- d'autre part, les codes susceptibles d'émerger d'une masse de traduction n'auront sans doute pas la même forme que les codes systématisés au moyen de base de règles. Car la relation entre un texte et sa traduction n'a rien de mécanique : si des contraintes se manifestent, c'est toujours à travers la liberté des choix de traductions ; si des généralités se font jour, c'est toujours dans et par le chaos d'une multitude de conditions singulières ; si des régularités sont observables, c'est à travers l'opacité de phénomènes trop éloignés du texte et des codes linguistiques pour ne pas être considérés comme contingents. Bref, l'étude locale des bi-textes ne donnera rien de semblable à des systèmes d'équivalence au niveau du lexique et de la grammaire ; en revanche, des équivalences peuvent apparaître au niveau la masse des phénomènes, l'éloignement procuré par le nombre permettant d'atténuer le brouillage des singularités locales. Du fait de ce changement d'échelle, rien ne permet de déterminer *a priori* la nature des régularités observées, et leur niveau de granularité : obtiendra-t-on des systèmes de transfert au niveau d'unités lexicales, de catégories grammaticales, de structures syntaxiques ? Seul l'observation statistique des corpus peut permettre de répondre à ces questions.

L'utilisation des corpus bi-textuels pour l'aide à la traduction dépend beaucoup de ce dernier point : l'exemple peut être *intéressant*, en tant que cas d'espèce, pour ce qu'il a de particulier ; mais il est surtout *utile* pour ce qu'il contient de généralisable.

Approche inductive et recours aux instruments statistiques sont les deux principales options méthodologiques qui gouvernent notre étude sur l'utilisation des bi-textes. Dans le travail expérimental que nous présentons dans les chapitres suivants, nous avons en outre délibérément écarté un certain nombre de techniques qui mettent en jeu des informations linguistiques, lexicales, morphologiques ou sémantiques. Les techniques étudiées se concentrent sur des régularités purement formelles, superficielles, observables sans modélisation préalable des codes linguistiques. Cet effort d'austérité n'implique pas que ces données n'aient pas leur place au sein des systèmes d'aide à la traduction basés sur l'exploitation des bi-textes. Au contraire, nous essaierons de montrer que les méthodes aveugles, sans *a priori* – nous dirons « *alinguistiques* » – ont tout à gagner d'un couplage avec des informations descriptives, élaborées, concernant les langues mises en jeu. Par ce travail de réduction, nous chercherons à explorer l'étendue et les limites des propriétés formelles – supposées – qui lient un texte et sa traduction. Dès lors que l'on considère l'« objet » bi-texte de manière isolée, coupé de son contexte interprétatif, force est de constater que ce n'est qu'un résidu, la trace objectivée d'une situation de communication. En refusant de prendre en compte les codes linguistiques, on ne fait qu'aggraver un peu plus cette coupure, qui de toute manière était inévitable : la nature incomplète, tronquée, de l'objet ne s'en trouve pas modifiée.

Puisque l'équivalence traductionnelle, au sens plein, n'est pas directement observable, quelles sont donc ces propriétés superficielles d'où émergeront peut-être les régularités espérées ? Nous en avons retenue une, selon nous essentielle, qui constituera le fil conducteur de toutes nos expérimentations : la *correspondance*, c'est-à-dire l'occurrence, de part et d'autre du bi-texte, de phénomènes assimilables les uns aux autres par l'intermédiaire d'une relation d'équivalence *supposée*.

Nous articulerons le concept de correspondances en deux moments : d'abord, nous étudierons la correspondance d'unités formelles, les *phrases*, définies de manière opératoire à partir d'une syntaxe simplifiée (essentiellement basée sur le repérage d'indices

typographiques). Ensuite nous chercherons les correspondances au niveau d'unités lexicales susceptibles de recouvrir un même contenu, sur le plan de la désignation.

Par la mise en œuvre et l'évaluation des techniques statistiques destinées à extraire ce type de correspondance, nous espérons montrer ce que le recours aux Mémoires de traduction peut apporter à l'aide à la traduction. Pour une telle évaluation, le choix des corpus est bien sûr déterminant : la difficulté est de trouver un corpus bilingue assez grand pour faire apparaître certaines régularités, tout en sachant que le volume des informations à traiter « alourdit » considérablement la mise en œuvre d'un protocole expérimental intégrant de nombreuses techniques et des paramètres variables. Comme le notent R. Bindi *et al*. (1994 : 29), la constitution du corpus est déterminante : « des corpus linguistiques représentatifs soigneusement sélectionnés, stratifiés et classés, sont essentiels à la création de méthodes et d'outils pour l'évaluation des techniques du TAL, de ses approches, de ses composants, de ses systèmes et enfin de ses performances. » Dans la mesure où les ressources matérielles, temporelles et financières dont nous disposons sont peu commensurables avec l'ampleur de la tâche, nous chercherons à appliquer des heuristiques débouchant sur des raccourcis. Il nous faudra décrire au mieux les caractéristiques formelles des corpus employés, afin de les corréler avec certains résultats : c'est, nous le verrons, un des enjeux majeurs de la recherche à venir dans le domaine de l'exploitation des bi-textes.

Partie II

Constitution de corpus bi-textuels : les techniques d'alignement

« The question of granularity of translation alignment brings up some fundamental issues that are currently at the center of translation theory but which receive little attention from computationalists. The theoretical question concerns the extent of the translator's responsibility to particular words and phrases on one hand, and to the overall function of the text on the other. »

Martin Kay, Préface à *Parallel Text Processing* Jean Véronis, 2000

II Constitution de corpus bi-textuels : les techniques d'alignement

On parle généralement de textes *parallèles* pour désigner des textes en relation de traduction mutuelle[120]. Notons que le terme est parfois utilisé dans le cas de textes rédigés dans des langues différentes et comparables sur le plan du sujet traité ou du contenu. Mais nous n'utiliserons dans la suite de ce travail que l'acception réservée à des textes en relation de traduction.

Comme le note Véronis (2000 :1), l'existence de corpus parallèle n'est pas un phénomène récent : déjà, la pierre de Rosette, datant de 196 av. J.-C., représentait deux versions d'un même texte, en grec et en égyptien, retranscrit selon trois systèmes d'écriture. Toute traduction implique la création de deux textes parallèles, même s'ils ne coexistent pas matériellement sur un même support ; l'existence de textes parallèles remonte par conséquent aux premières traductions.

Parmi ces textes parallèles, on trouve des corpus d'un genre particulier, dont les portions correspondantes sont mises côte à côte : nous dirons que ces corpus sont *alignés*. Les corpus parallèles alignés sont divers, et courants : traités, textes législatifs, traductions d'œuvres poétiques en face du texte original, éditions bilingues à vocation didactique, etc. Certains textes alignés sont très anciens : les traductions de la Bible, depuis la *Vulgate* de saint Jérôme, du fait de leur système de division en livres et en versets numérotés, étaient virtuellement alignées avec les textes originaux.

Pour résumer l'expression un peu longue de « textes bilingues parallèles alignés », nous employons le terme de *bi-texte*, dont l'usage s'est peu à peu imposé depuis sa création par B. Harris (1988). Par *corpus bi-textuel*, nous désignerons une collection de textes parallèles alignés.

Comme nous l'avons vu, les bi-textes sous format électronique peuvent avoir des applications nombreuses. Dans le champ spécifique de l'aide à la traduction, ils permettent d'aborder les problèmes de traduction sous un angle original, dans une approche « à base mémorielle », pour reprendre les termes d'Isabelle (1992 : 727) :

> « Le bi-texte constitue de ce fait l'amorce d'une approche à base mémorielle : au lieu de recréer à chaque fois une solution à un problème de traduction particulier, on se donne la possibilité de rappeler les solutions déjà trouvées. »

Le succès de ce type d'approche dépend avant tout de la quantité de bi-textes disponibles pour alimenter la mémoire de traduction. C'est pourquoi la possibilité de constituer automatiquement des bi-textes de grande taille et de bonne qualité est un enjeu de premier plan.

Or la tâche n'est pas si triviale qu'il y paraît au premier abord. Certes, la plupart du temps, la traduction conserve les découpages textuels : de la source à la cible, les chapitres, les sections, les paragraphes, voire les phrases, se correspondent *généralement* de façon biunivoque. Mais *généralement* ne signifie pas *systématiquement*, et il suffit d'un seul accroc dans ce réseau de correspondance pour corrompre la biunivocité de tout l'ensemble : par exemple, la simple omission d'une phrase traduite au début du texte peut engendrer un décalage sur tout le reste des appariements. Et dans la réalité, on observe que la biunivocité n'est pas aussi fréquente qu'on pouvait s'y attendre : il n'est pas rare qu'une seule phrase soit traduite par zéro, deux, ou plus de deux phrases cibles. C'est pourquoi il est nécessaire d'élaborer des stratégies sophistiquées afin de « recoller les morceaux » et d'aboutir à des alignements corrects.

Les premiers essais d'alignement automatique datent de 1987 : Kay & Röscheisen (1988, 1993) implémentent une méthode basée sur la distribution des mots, en n'utilisant aucune source d'information complémentaire en dehors des deux textes à aligner. Les auteurs montrent qu'en observant des cooccurrences de mots à l'intérieur de zones probablement correspondantes (le début et la fin des textes, ainsi que les zones se situant au même niveau, dans chacun des textes) il est possible d'extraire des correspondances lexicales, qui peuvent servir ensuite de « points d'ancrage » pour aligner les phrases. Le

[120] Nous ne prêterons pas attention au sens de la traduction.

grand mérite de ces premières recherches est de montrer qu'il est possible d'aligner sans passer par le sens, en se basant sur des propriétés purement formelles. Dans le même esprit, un autre type de technique se fait jour : les travaux parallèles de P. Brown, J. Lai & R. Mercer (1991) et W.A. Gale & K.W. Church (1991, 1993) obtiennent de bons résultats en se basant sur l'observation des longueurs de phrase. Par ailleurs, les systèmes étudiés intègrent une modélisation des probabilités empiriques des différents types de regroupements (ou *transitions*, du type 1-1, 1-0, 0-1, 1-2, 2-1, etc.).

A la suite de ces travaux fondateurs deux principales directions sont ouvertes : 1/ l'utilisation d'ancrages lexicaux et 2/ le développement des systèmes probabilistes modélisant les variations des longueurs de phrase, formeront le noyau dur de toutes les techniques mises en œuvre par la suite.

Etonnamment, les indices superficiels semblent très efficaces. La « philosophie » qui guide ces recherches s'appuie sur un constat de bon sens : bien souvent, un humain peut aligner deux textes sans connaître les deux langues impliquées, uniquement en se basant sur des indices formels tels que découpages en sections, longueurs des phrases, récurrences de certains couples d'unités, graphies ressemblantes, traduction des nombres et des noms propres, etc. Ce sont ces mêmes indices qui alimenteront la plupart des algorithmes d'alignement.

Ainsi, pour tirer parti des ancrages lexicaux, Church (1992) propose de se baser non sur les distributions, mais sur les ressemblances superficielles caractérisant les chaînes de caractères des mots apparentés : l'observation de 4-grammes est appliquée au repérage de mots apparentés (ou *cognats*), et les couples d'unités lexicales en correspondance supposée forment des nuages de points, qu'on peut ensuite filtrer pour extraire l'alignement des zones de forte densité. La comparaison des chaînes de caractères aboutit au développement de systèmes mixtes, intégrant à la fois des modèles probabilistes relatifs aux longueurs et des modèles orientés vers les ressemblances superficielles (M. Simard, G. Foster & P. Isabelle, 1992 ; A. Mc Enery & M. P. Oakes, 1995). A la suite de Kay & Röscheisen, F. Débili & E. Sammouda (1992) montrent qu'il n'y a pas de cercle vicieux dans le fait d'utiliser successivement l'alignement des mots pour aligner les phrases, et l'alignement des phrases pour aligner les mots : le processus converge vers un alignement de plus en plus précis, chaque étape apportant de nouvelles informations. Par ailleurs, le repérage des

cognats se raffine petit à petit : Mc Enery & Oakes (1995) en proposent une caractérisation améliorée en faisant intervenir le coefficient Dice[121] dans la comparaison de deux chaînes.

Pour l'identification des ancrages lexicaux, l'étude des distributions lexicales donne également de bons résultats. P. Fung & K.W. Church (1994) proposent une méthode simple basée sur un pré-découpage grossier des deux textes en zones d'égales importances : les occurrences et cooccurrences des unités dans les zones correspondantes permettent alors d'établir de manière fiable une liste d'unités équivalentes pouvant servir d'amorçage à un processus itératif du type de celui décrit par Débili & Sammouda (1992). S. Chen (1996) élabore aussi une méthode d'alignement en se basant sur l'appariement des mots, en s'inspirant du modèle de traduction basé sur l'exemple développé par P. Brown *et al*. (1993).

Enfin, M. Davis, T. Dunning & B. Ogden (1995) montrent comment combiner différents types d'indices pour les intégrer dans un même cadre algorithmique. Avec une approche similaire, des résultats très satisfaisants sont obtenus par P. Langlais & M. El-Beze (1997) : divers indices, basés sur les longueurs de phrases, les chaînes identiques (transfuges), les cognats, les probabilités de transitions, sont pondérés de façon à optimiser les performances. I. D. Melamed (1997) combine aussi plusieurs indices, en utilisant des heuristiques adaptées pour réduire l'espace de recherche et minimiser les chances d'erreur.

On constate que les techniques sont nombreuses : les résultats obtenus à l'issue de la deuxième campagne du projet ARCADE (Véronis, 1997) montrent en outre qu'elles sont parvenues à maturité, certains auteurs considérant le problème comme étant pratiquement résolu. Mais une vision trop optimiste risque de masquer la véritable nature du problème : la difficulté de la tâche ne peut être évaluée dans l'abstrait, car elle dépend étroitement du type de traduction mise en jeu, et des techniques éprouvées peuvent se révéler calamiteuses sur un corpus spécifique. Pour affirmer le problème résolu, il faudrait avoir réglé la question de l'équivalence traductionnelle en général, et nous en sommes encore loin : entre les résultats récents et cet objectif théorique encore lointain, nous sommes convaincu qu'il reste une importante marge de progression.

---

[121] Pour deux chaînes de longueur n_1 et n_2, comportant n_{12} caractères communs, on a : $Dice = 2*n_{12}/(n_1+n_2)$

C'est pourquoi la deuxième partie de notre étude est dédiée à l'exposition et à l'évaluation des principales méthodes développées pour l'alignement. Par un travail expérimental, nous chercherons à comparer les différentes techniques, afin d'en marquer les limites et d'en montrer les possibilités. Nous essaierons, le cas échéant, d'apporter des ajustements aux méthodes présentées : nous espérons ainsi mettre à jour les directions de recherche les plus prometteuses.

Mais avant toute chose, la notion même d'alignement requiert des éclaircissements : formellement, certains alignements mettent en correspondance des segments textuels, d'autres des jeux de coordonnées ; certaines techniques opèrent au niveau des mots, d'autres au niveau des phrases, ou des paragraphes ; certains algorithmes exigent la vérification de conditions préalables, comme le préalignement des paragraphes des deux textes impliqués, ou le strict parallélisme de l'ordre des séquences textuelles – d'autres non. Une brève revue de l'état de l'art nous révèle que l'alignement recouvre des notions variables : avant d'aborder les techniques, il est nécessaire d'harmoniser ces conceptions divergentes en exposant quelques prolégomènes.

II.1 Le concept d'alignement

Isabelle (1992 : 724-725) donne une définition formelle du concept de bi-texte : « un bi-texte est un quadruplet < T_1, T_2, *Fs*, *C*> dans lequel T_1 et T_2 sont deux textes, *Fs* est une fonction qui décompose ces textes en des ensembles de segments, et *C* est un ensemble de correspondances entre *Fs*(T_1) et *Fs*(T_2) . »

Pour notre part, nous préférerons écrire *A* plutôt que *C*, car nous réservons le terme de *correspondance* à une autre forme d'appariement (cf. *Partie III*). *A* est donc un sous-ensemble du produit cartésien *Fs*(T_1) x *Fs*(T_2).

Les couples de *A* seront appelés *binômes traductionnels*. Nous emploierons le terme générique de *segments* pour désigner les deux portions textuelles appariées dans ces

binômes. En ce qui concerne la fonction de segmentation, on parlera de *granularité*[122], pour désigner la finesse des segments produits : paragraphes, phrases ou syntagmes.

La définition précédente est assez vague pour inclure différents types formels d'alignement :

- alignement *complet* vs *partiel*, suivant que toutes les unités des deux textes apparaissent – ou non – au moins une fois dans un couple de C ;

- granularité *fixe* vs *variable*, suivant que le bi-texte contient des appariements entre segments de même rang (p. ex. des phrases) ou de rangs différents (p. ex. des phrases et des mots) ;

- segmentation *de type partition* vs *hiérarchique vs complexe* : si les segments produits par *Fs* ne se recouvrent pas, ils forment une *partition* des textes ; s'ils se recouvrent suivant des rapports d'inclusion, la segmentation est dite *hiérarchique* ; s'il existe des recouvrements sans inclusion, la segmentation est dite *complexe*. Une segmentation hiérarchique peut découler des différents paliers formels du texte : section, paragraphe, phrase, syntagme, mot. Pour Isabelle (1992 : 725), l'analyse syntaxique peut être au principe d'une telle segmentation : « Une façon très naturelle d'obtenir un tel système hiérarchique consiste à assimiler la fonction de segmentation *Fs* à une fonction d'analyse syntaxique (…). Une telle décomposition permet la formulation de correspondances hiérarchisées entre les deux textes (…) ».

Nous pouvons maintenant donner une définition générale de l'alignement : nous désignons par *alignement* l'opération consistant à faire correspondre, au sein de deux textes parallèles, les segments textuels qui sont en relation d'équivalence traductionnelle.

[122] calque de l'anglais *granularity*, antonyme des termes *résolution* ou *définition* – une granularité importante correspondant à une faible résolution.

Dans la suite de notre travail, nous négligerons la donnée du sens initial de la traduction : par commodité, nous conserverons les termes de *source* et *cible* à seule fin de différencier formellement les deux côtés du bi-texte. Nous emploierons le terme de *version* pour désigner chacun de ces côtés, et nous noterons T et T' les textes des *versions* source et cible, L et L' les langues source et cible[123].

Il est temps de préciser, sur un plan formel, ce qu'on entend par texte *parallèle*. Le parallélisme implique la notion de *compositionnalité*, explicitée par Isabelle (1992 : 724) « (...) les traductions obéissent à un principe dit de compositionnalité : la traduction d'un segment complexe est généralement une fonction de la traduction de ses parties, et ce, jusqu'au niveau d'un ensemble d'unités élémentaires ». C'est donc la compositionnalité qui permet de ramener le principe d'équivalence traductionnelle, défini globalement entre les deux textes, au niveau d'unités plus petites que ces textes.

La compositionnalité de la traduction peut s'observer du point de vue du processus traductionnel : le plus souvent, le traducteur humain traduit segment après segment, même s'il garde à l'esprit l'ensemble du texte (il est rare qu'on lise un texte de bout en bout pour le traduire ensuite d'une traite, en se basant sur la mémoire, sauf en interprétariat, puisque c'est le principe de la traduction consécutive). Mais nous ne nous situons pas dans la perspective du processus : nous partons du résultat (T et T'), et il nous faut examiner quelles sont les conditions formelles de parallélisme nécessaires à la mise en œuvre de l'alignement.

Nous définirons le parallélisme comme la conjonction de deux propriétés :

– *quasi-bijection* : tout d'abord, la compositionnalité traductionnelle ne se résume pas à la compositionnalité des unités de T et T' séparément. Elle n'implique pas seulement que T et T' soient décomposables chacun de leur côté, mais requiert l'existence d'une décomposition de T *congruente* à une décomposition de T'. Cette congruence peut être formulée dans l'hypothèse de *quasi-bijection* : chaque

[123] En ce qui concerne notre corpus, nous désignerons arbitrairement l'anglais comme la langue source et le français comme la langue cible.

segment de T est supposé être en relation bijective avec un segment de T'. On tolère cependant qu'il y ait omission ou insertion d'un segment d'un côté ou de l'autre, de façon marginale (d'où le « quasi »).

– *quasi-monotonie* : lorsque l'ordre des segments de T est supposé respecter, peu ou prou, l'ordre des segments équivalents de T'.

En résumé :

parallélisme : quasi-bijection + quasi-monotonie

Ces notions sont assimilables aux conditions de « quasi-bijection » et « quasi-synchronisation » introduites par J.-M. Langé & E. Gaussier (1995 : 71).

A strictement parler, la compositionnalité n'implique que la condition de quasi-bijection, et il est tout à fait envisageable de mettre en correspondance des segments non monotones, comme l'ont montré certains travaux. Par exemple, C. Fluhr, F. Bisson & F. Elkateb (in Véronis, 2000 §9) proposent une technique d'alignement inspirée des méthodes de recherche d'information interlinguistique (en anglais *Cross Language Information Retrieval*, CLIR), s'appuyant sur le contenu des phrases, dont on traduit les mots au moyen d'un dictionnaire de transfert. Avec cette technique, pour un segment source, la recherche d'un équivalent traductionnel peut couvrir la totalité du texte cible, sans aucune contrainte de position.

Mais la majorité des systèmes d'alignement se basent sur la monotonie des séquences textuelles, et, dans la pratique, la plupart des traductions respectent les critères du parallélisme. Ainsi, par défaut, nous étudierons l'alignement dans le cas général de la monotonie, et nous traiterons les alignements non-monotones comme des cas particuliers.

Par ailleurs, nous verrons que certains systèmes d'alignement font l'économie de la segmentation, et établissent directement des liens biunivoques séquentiels entre des unités lexicales de T et de T' : les *points d'ancrage* (en anglais « *anchor points* »). Il serait donc possible d'élargir le concept de parallélisme à des textes où l'on ne peut appliquer l'hypothèse de compositionnalité traductionnelle (selon laquelle la traduction du tout est

une fonction de la traduction de ses parties). En ce qui nous concerne, nous resterons dans le cadre restreint où le parallélisme présuppose la possibilité de décomposer. Notons par ailleurs que dans les systèmes d'alignement se basant sur des points d'ancrage, la segmentation ne précède pas la mise en évidence du parallélisme, mais elle en dérive. En effet, la donnée d'une série de points d'ancrage permet de segmenter le texte à partir des zones encadrées par ces points. Par suite, le parallélisme des points d'ancrage n'a de sens que si l'on considère ces zones comme équivalentes. On retrouve alors l'hypothèse de compositionnalité, condition inhérente à l'alignement.

II.2 L'alignement phrastique

L'alignement phrastique est une des formes les plus courantes de l'alignement. Nous nous proposons ici de faire un bilan de l'état de l'art et de développer quelques idées quant aux améliorations possibles.

II.2.1 Segmentation

Définir rigoureusement la phrase comme unité linguistique est un problème à part entière. Plusieurs niveaux s'entrecroisent sans permettre d'établir des caractérisations convergentes et consistantes : la prosodie, la ponctuation, la syntaxe, la sémantique et le niveau des propositions logiques.

D'un point de vue syntaxique, on peut difficilement s'en tenir à une définition canonique du type : Phrase = SN + SV. Tout d'abord, parce que cela n'inclut pas les syntagmes non verbaux (du type : « Bien sûr », ou, « Pourquoi pas ? » ou simplement « non »), très courants dans certains genres textuels (notamment les transcriptions de l'oral). Ensuite parce que ce genre de définition a du mal à rendre compte des phrases complexes possédant plusieurs noyaux verbaux, qu'il s'agisse de propositions coordonnées ou subordonnées (énumérations, listes, propositions enchâssées, etc.).

Dans l'exemple suivant[124], représentatif d'un certain type de discours juridique, on voit comment deux types d'énumération s'enchâssent à l'intérieur d'une unité plus large, syntaxiquement complexe, qu'on hésite à appeler « phrase » :

> *Le Parlement européen ,*
> *- vu la proposition de la Commission au Conseil* (*COM*(*96*)*0049 - 96*/*0039*(*CNS*)),
> *- consulté par le Conseil conformément à l'article 43 du traité CE* (*C4-0156*/*96*)*2*,
> *- vu l'article 58 de son règlement*,
> *- vu le rapport de la commission de l'agriculture et du développement rural* (*A4-0264*/*96*),
> *1. approuve, sous réserve des modifications qu'il y a apportées, la proposition de la Commission*;
> *2. invite le Conseil, au cas où il entendrait s'écarter du texte approuvé par le Parlement, à en informer celui-ci*; (…)

Cet exemple illustre en outre la richesse des procédés typographiques et des marques de ponctuation mis en œuvre dans certains textes : passage à la ligne, virgule et tiret pour la première énumération, point virgule et numérotation pour la seconde. Dans les textes fortement structurés par les énumérations, la logique de la ponctuation peut être à la fois complexe et fluctuante.

Dans l'exemple suivant, tiré du même texte que précédemment, le premier « : » est suivi par une majuscule, mais pas les autres ; en outre, l'alinéa a) se termine par un point virgule tandis que le b) se termine par un point, car marquant la fin du 1) : l'architectonique du texte est inscrite dans sa ponctuation.

> *Les répercussions de ces accords doivent être prises en considération et la proposition poursuit donc les objectifs suivants*:
> *1. La protection agricole de l'UE doit être adaptée à la suite de la tarification du GATT et certains aménagements techniques doivent être apportés au règlement 3448/93 pour éliminer les dispositions qui ne sont plus applicables*:
> *a*) *plus particulièrement, les modifications au système tarifaire englobent la suppression des dispositions concernant la méthode de calcul des éléments mobiles qui ont été rejetés à la suite de la tarification et des adaptations aux droits additionnels sur les sucres et sur les farines, lesquels ne sont plus calculés, mais également tarifés dans le même contexte*;

[124] Extrait d'un rapport du parlement européen (réf. A4-0264) disponible à l'adresse : http://www.europarl.eu.int

> *b*) *des modifications doivent être apportées aux dispositions relatives à la protection agricole*. *Certaines marchandises relevant du règlement 3448/93 peuvent être soumises à des droits additionnels, tels qu'ils sont fixés par l'article 5 de l'Accord sur l'agriculture* (*clause de sauvegarde spéciale permettant aux parties contractantes d'interdire des importations à des prix inférieurs aux niveaux normaux*).
> *2. Par ailleurs, certaines règles de gestion doivent être introduites pour veiller au respect des engagements contractés en matière de restitutions à l'exportation. En ce qui concerne les restitutions, la Communauté s'est engagée à ne pas accorder de restitutions au-delà d'un montant maximum par exercice financier, pour les produits agricoles exportés sous forme de marchandises ne relevant pas de l'annexe II.*

Faut-il compter ici deux, trois, cinq phrases, ou bien plus ?

Ainsi, si l'on admet une définition extensive de la phrase basée sur le noyau verbal régissant, on peut aboutir à des phrases atteintes de gigantisme, formant un texte entier s'étalant sur plusieurs pages. Mais si l'on se base seulement sur certains indices typographiques, tels que les sauts à la ligne, n'importe quel syntagme, dûment énuméré, pourra revendiquer son statut de phrase.

Ces phénomènes aboutissent parfois à un renversement de l'ordre traditionnel de l'analyse grammaticale : dans certains cas, ce n'est plus la succession des phrases qui forme le texte, mais la structuration du texte qui conditionne la formation des phrases. Dans l'exemple précédent, la double énumération correspond à une structure macrosémantique normalisée, qu'on pourrait résumer ainsi : - énumération des *motifs* - énumération des *résolutions* législatives.

Il en résulte que le plan sémantique est insuffisant, lui aussi, à caractériser l'unité de la phrase. Comme l'affirme Rastier (1994 : 115-116), les relations sémantiques ne cessent de transgresser les limites de la phrase. Ainsi « à l'autonomie syntaxique qui refléterait la complétude et l'autosuffisance de la prédication, on doit opposer les relations sémantiques qui rattachent toute phrase à son contexte linguistique et situationnel. Si bien que le découpage d'un texte en phrases n'est pas si simple et la recherche d'un point n'y suffit pas. »

Sans chercher à épuiser ce sujet, nous nous tiendrons, dans le cadre de nos expérimentations, à une définition opératoire de la phrase : une phrase est une unité

textuelle minimale dont les frontières sont marquées par des indices typographiques de début et de fin.

La syntaxe de ces indices est lâche, souvent ambiguë, mais l'intuition (reposant sur la compréhension du texte) permet en général de trancher de manière assez sûre.

Ces indices peuvent être, par exemple : un point, un point d'exclamation, un point d'interrogation, un point virgule, deux points, un tiret, une position en début de ligne, une marque de paragraphe, une marque d'alinéa, les limites d'une case de tableau, etc.

Les éléments de liste, les titres, les légendes de graphique seront aussi considérés comme des phrases, en tant qu'ils ont une certaine autonomie formelle.

Cette définition accorde aux marques typographiques un rôle dans la mécanique textuelle, dans le cadre d'une sémiologie élargie, et les unités dégagées deviennent indépendantes des données purement linguistiques. Par exemple les trois phrases suivantes :

> *Il faut en outre*:
> *- introduire un système obligatoire d'indication de la qualité et de l'origine de la viande et des produits de viande, et*
> *- intensifier sans relâche les mesures d'information et de publicité pour les produits de viande bovine.*

peuvent être réécrites en une seule phrase, sans que rien n'ait changé sur le plan strictement linguistique :

> *Il faut en outre introduire un système obligatoire d'indication de la qualité et de l'origine de la viande et des produits de viande, et intensifier sans relâche les mesures d'information et de publicité pour les produits de viande bovine.*

Nous adopterons donc un point de vue très proche de celui décrit par Michel Simard (1997 : 491) pour la constitution du corpus BAF[125], lorsqu'il énumère les critères de segmentation suivants : « Une phrase est une séquence de mots syntaxiquement autonomes, terminée par un point. (…) Les titres sont des phrases. (…) Les marqueurs

[125] cf. la description du § II.3.1

d'énumération sont des phrases [par exemple III, N.B. etc.] (…) Les items d'énumération sont des phrases. (…) Chaque cellule d'un tableau est une phrase. »[126]

II.2.1.1 Règles syntaxiques du découpage en phrase

Pour les expérimentations présentées plus loin, nous avons formalisé un certain nombre de règles syntaxiques simples permettant la segmentation. Nous nous sommes intentionnellement limité aux cas de figure les plus simples et les plus fréquents.

Nous avons tenu compte de quatre marqueurs : le point, le point virgule, les deux points et le saut de ligne.

De ces marqueurs, un seul est ambigu : le point, lorsqu'il marque une abréviation.

Nous avons ignoré les points dans les cas suivants :

- point non suivi d'un espace et d'une majuscule, ou
- point suivant une lettre majuscule isolée (sigle), une lettre minuscule isolée (pour les cas du type *i.e.*, *e.g.*, *n.b.*, *q.e.d.*, *c.à.d.*, *n.*), ou une abréviation standard du type : *etc.*, cf., *pp.*, *Mr.*, *Mrs*, *MM.*, etc.

Ces traitements d'exception dépendent du corpus. Les textes italiens, par exemple, sont riches en acronymes du type *A.CO.TRAL.*, pouvant demander un traitement spécifique. Pour certains textes, il peut être nécessaire de fournir un dictionnaire d'abréviations *ad hoc*.

Sous la forme d'expression régulière, notre grammaire s'écrit :

Phrases = (*Mots Séparateurs* (*Séparateurs*)*)* *DerniersMots point espace Maj* |
(*Mots Séparateurs* (*Séparateurs*)*)* *SymbolesTerminaux*

Séparateurs = {*espace, virgule, point d'exclamation, point d'interrogation, tiret, guillemets, parenthèses*}
Mots = toute suite de caractères sans séparateur
DerniersMots = Mots à l'exception des abréviations (*etc, cf, pp, MM*) *et des mots d'une lettre.*

[126] "A sentence is a syntactically autonomous sequence of words, terminated by a full-stop punctuation. (…) Titles are sentences. (…) Enumerators are sentences (…) Items of an enumeration are sentences. (…) Each cell in a table is a sentence."

Maj = symbole additionnel indiquant la position d'une majuscule à l'initiale du mot suivant
SymbolesTerminaux = deux points, point virgule, marque de paragraphe

Dans le cas de notre corpus, nous avons estimé la proportion de segmentations erronées dues à des d'abréviations non prévues (comme *Ph.*, *doc.*, *cat.*, *prép.*, *alim.*) à moins de 0,5 %.[127]

II.2.2 Notations et mesures formelles

Il existe plusieurs manières de représenter un alignement. Nous donnons ici un aperçu des conventions les plus courantes.

II.2.2.1 Représentation bidimensionnelle

Dans chaque texte, les unités textuelles (paragraphes, phrases, mots ou caractères) peuvent être représentées par un entier exprimant leur position.

Les deux textes parallèles constituent alors un espace à deux dimensions, où la correspondance de deux unités est représentée par un point. Dans cet espace, l'alignement du bi-texte peut être décrit par la suite des points marquant les coordonnées des unités appariées (cf. l'exemple de la figure 16). Aux deux extrémités, les points correspondant au début et à la fin des textes seront appelés *origine* et *fin* de l'espace ; la droite théorique passant par ces deux points sera appelée *diagonale* de l'espace.

[127] Pour des études sur la désambiguïsation de ces problèmes de ponctuation cf. les travaux de Gregory Grenfenstette & Pasi Tapanainen (1994), et de David Palmer (1994).

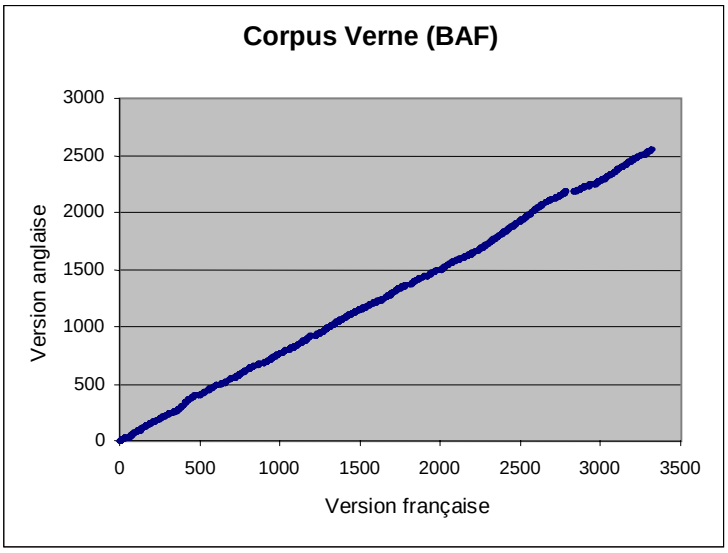


figure 16 : *représentation bidimensionnelle de l'alignement du corpus Verne* (*extrait du corpus BAF*)

II.2.2.2 Notation ensembliste

Soient T et T' deux textes parallèles segmentés en phrases. On notera :

$T = P_1 P_2 \ldots P_n$

$T' = P'_1 P'_2 \ldots P'_m$

Soit A un alignement de T et T'. A est un sous-ensemble de T x T', c'est-à-dire un ensemble de k binômes :

$$A = \{(P_{i1};P'_{j1}),(P_{i2};P'_{j2}),\ldots,(P_{ik};P'_{jk})\} \quad \text{avec } 0 \leq i_1 \ldots i_k \leq n \text{ et } 0 \leq j_1 \ldots j_k \leq m$$

respectant les conditions de quasi-monotonie et de quasi-bijection.

Pour les phrases n'ayant pas de correspondant, suite à une insertion ou une omission, deux conventions sont possibles : soit on introduit deux phrases vides P_0 et P'_0, avec lesquelles ces phrases seront alignées (et donc apparaîtront dans A), soit on ne les fait pas figurer dans A. On verra plus loin comment ces deux possibilités peuvent affecter le calcul des indices d'évaluation d'un alignement.

Dans les cas où une même phrase apparaît dans plusieurs de ces couples, on peut procéder à des regroupements et aboutir à une notation élargie.

Par exemple, pour $T = P_1\ P_2\ P_3\ P_4\ P_5\ P_6$ et $T' = P'_1\ P'_2\ P'_3\ P'_4\ P'_5\ P'_6\ P'_7$, l'alignement :

$$A = \{(P_1;P'_1),(P_1;P'_2),(P_2;P'_0),(P_3;P'_3),(P_4;P'_4),(P_5;P'_4),(P_6;P'_5),(P_6;P'_6),(P_6;P'_7)\}$$

s'écrit de manière condensée :

$A = \{(P_1;P'_1P'_2)(P_2;)(P_3;P'_3)(P_4P_5;P'_4)(P_6;P'_5P'_6P'_7)\}$ ou encore,
$A = \{(P_1;P'_{1..2})(P_2;)(P_3;P'_3)(P_{4..5};P'_4)(P_6;P'_{5...7})\}$

Dans le cas général, on notera le i^{ème} binôme : $B_i=(Pn_i\ldots m_i;P'p_i\ldots q_i)$.

Les binômes de la première notation, sous forme de couple, sont en fait des binômes fragmentaires : la présence du couple $(P_5;P'_4)$ dans cette notation n'indique pas que P_5 est exactement aligné avec P'_4, mais qu'ils font partie d'un binôme plus large $(P_4P_5;P'_4)$, à l'intérieur duquel ils sont alignés.

Par ailleurs, ce formalisme autorise les correspondances croisées et discontinues, mais ne permet pas de représenter des correspondances hiérarchisées (par exemple un alignement où l'appariement global de deux paragraphes coexisterait avec les appariements plus fins des phrases qui les composent).

II.2.2.3 Notation sous forme de chemin

Isabelle & Simard (1996 : 4), se sont intéressés à une catégorie d'alignement vérifiant les conditions suivantes :

- absence de correspondances croisées :

$$\forall\ (i,i',j,j')\ [(\{i,j\}\in A)\wedge(\{i',j'\}\in A)\wedge(i'<i) \Rightarrow (j'<j)]$$

- absence de correspondances à chevauchement partiel :

$$\forall\ (i,j,j')\ [(\{i,j\}\in A)\wedge(\{i,j'\}\in A) \Rightarrow \forall\ i'\ [(\{i',j\}\in A) \Rightarrow (\{i',j'\}\in A)]]$$

– absence de correspondances discontinues :

$$\forall (i,j,j') [(\{i,j\}\in A)\wedge(\{i,j'\}\in A) \Rightarrow \forall j'' [(j''>j)\wedge(j>j'\}\Rightarrow (\{i,j''\}\in A)]]$$

Ces trois conditions résument et formalisent la propriété de *monotonie* introduite précédemment. Dans le cas d'un alignement monotone et complet[128], c'est-à-dire attribuant des correspondants (même vides) à toutes les phrases de chaque texte, on peut définir la notion de *chemin*.

Un chemin est défini comme une suite de *transitions* permettant de regrouper, de proche en proche, l'ensemble des phrases des deux textes en binômes d'alignement. Chaque transition n'est autre qu'un couple d'entiers positifs (*p*:*k*) exprimant le regroupement de *p* phrases contiguës avec *k* phrases contiguës.

Par exemple, pour exprimer l'alignement de deux textes T et T' segmentés en phrases $P_1\ P_2\ P_3\ P_4\ P_5\ P_6$ et $P'_1\ P'_2\ P'_3\ P'_4\ P'_5\ P'_6\ P'_7$, on peut recourir aux deux systèmes de notations :

– avec la notation ensembliste :

$$A = \{(P_1;P'_1\ P'_2)(P_2;\ )(P_3;P'_3)(P_4\ P_5;P'_4)(P_6;P'_5\ P'_6\ P'_7)\}$$

– avec la notation de type chemin :

$$A = ((1:2),(1:0),(1:1),(2:1),(1:3))$$

On peut établir une typologie sommaire des différents types de transition :

1. Conservation :	1:1		
2. Insertion :	simple :	0:1	multiple : 0:2, 0:3, etc.
3. Omission :	simple :	1:0	multiple : 2:0, 3:0, etc.
4. Fusion :	simple :	2:1	multiple : 3:1, 4:1, etc.
5. Scission :	simple :	1:2	multiple : 1:3, 1:4, etc.
6. Coalescence :	simple :	2:2	multiple : 2:3, 3:2, 3:3, 2:4, 4:2, etc.

Les cas de figure 1, 4 et 5 sont cohérents avec l'hypothèse normale de parallélisme.

[128] lorsqu'un alignement n'est pas complet, on parlera d'alignement partiel ou de *préalignement*.

Les cas de figure correspondant à l'insertion et à l'omission (2 et 3) permettent de traiter les écarts par rapport à la bijectivité.

Enfin, les cas de coalescence (6) permettent de tenir compte, localement, du non-respect de la monotonie. Par exemple, un texte du corpus BAF contient un glossaire trié par ordre alphabétique. Or il est bien évident que l'ordre des termes en anglais et l'ordre en français ne se correspondent pas : la solution est de considérer les deux glossaires comme deux blocs, résultant de la coalescence de toutes leurs entrées. L'alignement des deux glossaires est donc codé en une seule transition.

Notons qu'un chemin peut aussi être représenté comme suite de *points d'ancrage* si l'on retient les coordonnées de la première phrase de chaque groupement :

A =((1,1),(2,3),(3,3),(4,4),(6,5))

C'est dans ce sens que nous parlerons, ultérieurement, des points constituant un chemin. Ce type de notation est similaire au format COAL développé par Isabelle & Simard (1996).

En voici un exemple, pour les extraits du tableau 6 :

	Anglais		Français
P_1	*"Here we are at the 10th of August," exclaimed J.T. Maston one morning, " only four months to the 1st of December.*	P'_1	*« Nous voilà au 10 août, dit un matin J.-T. Maston*
		P'_2	*Quatre mois à peine nous séparent du premier décembre !*
		P'_3	*Enlever le moule intérieur, calibrer l'âme de la pièce, charger la Columbiad, tout cela est à faire !*
P_2	*We shall never be ready in time !"*	P'_4	*Nous ne serons pas prêts ! »*

tableau 6 : *exemple d'alignement*

On peut noter l'alignement comme suit :

$T = P_1 P_2$ $T' = P'_1 P'_2 P'_3 P'_4$

$A = \{(P_1;P'_1P'_2),(\varnothing;P'_3),(P_2;P'_4)\}$

ou bien, si l'on représente ces regroupements par une suite de transitions :

$A = ((2:1), (0:1), (1:1))$

II.2.3 Critères quantitatifs d'évaluation

Avant de présenter les différentes méthodes, il nous faut examiner rapidement les mesures permettant d'évaluer quantitativement la qualité d'un alignement obtenu automatiquement. Le principe de ce type d'évaluation est de comparer l'alignement produit avec un alignement effectué « manuellement » (c'est-à-dire par l'humain), pris comme alignement de référence.

II.2.3.1 Alignement de référence

Il est en effet nécessaire de se donner une base de comparaison : on ne peut évaluer un alignement que de manière relative, par rapport à l'alignement considéré comme référence.

Or il est souvent difficile d'établir ce que doit être *l'*alignement de référence. Les divergences qui apparaissent dans la traduction font qu'il est parfois difficile de choisir entre fusion et insertion. Considérons l'extrait suivant, tiré de l'alignement de référence du corpus BAF :

	Anglais		***Français***
P_1	*" Bah!"*	P'_1	*--Il nous recevra mal, murmura Bilsby entre les quatre dents qu'il avait sauvées de la bataille.*
P_2	*growled Bilsby between the four teeth which the war had left him;*		
P_3	*" that will never do!"*		
P_4	*" By Jove!"*	P'_2	*--Par ma foi, s'écria J.-T. Maston, aux prochaines élections il n'a que faire de compter sur ma voix!*
P_5	*cried J. T. Maston, " he mustn't count on my vote at the next election!"*		

tableau 7 : *problèmes d'alignement*

Le choix de considérer P_3 comme un ajout, sans correspondance, est discutable. On aurait aussi bien pu fusionner P_3 avec P_1 et P_2, en correspondance avec P'_1. Car *that will never do* vient compléter et expliciter le *Bah*, même si cette expression ne traduit pas mot à mot l'assertion *il nous recevra mal*. Le problème est le suivant : la compositionnalité est relative, et joue à des niveaux différents allant du mot à la phrase, ou plus. A chacun de ces niveaux, l'équivalence connaît des degrés différents : dès lors, on peut hésiter entre un découpage plus fin et une équivalence plus faible, et ou un découpage plus grossier (en regroupant les phrases) et une meilleure équivalence ; de même lorsque l'équivalence est partielle, comme dans le cas du tableau 7, il faut choisir entre l'appariement vide, ou le regroupement. Dans la mesure où l'équivalence est un phénomène scalaire, le fait de trancher entre une configuration ou une autre implique forcément une part d'arbitraire. Nous approfondirons la question en traitant des correspondances lexicales (cf. Partie III), pour lesquelles ce problème est particulièrement aigu.

Dans la réalisation d'un corpus de référence, il est donc important d'expliciter en détail tous les critères employés pour résoudre les cas litigieux. L'alignement obtenu peut alors être considéré comme *un* alignement de référence, sachant qu'il peut en exister des variantes tout aussi justifiées.

Par ailleurs, il peut être intéressant de confier la tâche à plusieurs annotateurs, afin de mesurer l'accord entre les différents alignements proposés, car seul l'accord intersubjectif peut garantir une certaine objectivité.

II.2.3.2 Précision et Rappel

Une fois l'alignement de référence établi, il est aisé de définir des mesures quantitatives permettant d'évaluer un alignement quelconque. Depuis les campagnes d'évaluation menées au sein du projet ARCADE (Véronis, 1997), destinées à fournir une base de comparaison solide pour évaluer des systèmes d'alignement de conceptions différentes, un certain consensus s'est établi : les mesures quantitatives généralement employées sont le *rappel* et la *précision*.

Rappel = Nombre d'alignements corrects / nombre d'alignements de référence
Précision = Nombre d'alignements corrects / Nombre d'alignements proposés

Le rappel indique la proportion de binômes de l'alignement de référence présents dans l'alignement évalué, et la précision indique la proportion de binômes corrects dans l'alignement évalué. Ces deux indicateurs sont les complémentaires du silence et du bruit :

Silence = 1-Rappel

Bruit = 1-Précision

Traduit dans notre système de notation, on obtient, dans l'évaluation de *A* par rapport à $A_{réf}$ (la notation d'un ensemble entre barres représente le cardinal) :

$$précision(A/A_{rèf}) = \frac{\left|A \cap A_{réf}\right|}{\left|A\right|} \quad (2)$$

$$rappel(A/A_{réf}) = \frac{\left|A \cap A_{réf}\right|}{\left|A_{réf}\right|} \quad (3)$$

En recherche d'information, rappel et précision évoluent souvent de manière antagoniste : plus une méthode obtient un bon rappel, plus la précision est faible. En ce qui concerne l'alignement, un bon rappel permet parfois d'améliorer la précision car il y a consolidation réciproque des binômes, du fait des hypothèses de quasi-bijection et monotonie : la donnée d'un binôme correct permet d'aligner plus sûrement les phrases qui suivent et qui précédent.

Enfin, pour tenir compte en même temps des deux mesures, nous reprendrons la *F-mesure* utilisée dans le projet ARCADE (reprise de van Rijsbergen, 1979), comme indice d'évaluation globale synthétisant rappel et précision :

$$F = \frac{2 \cdot (P \cdot R)}{(P + R)} \quad (4)$$

Cette mesure, qui représente la moyenne harmonique de *P* et *R*, n'est autre que le coefficient Dice appliqué à la comparaison de *A* et $A_{réf}$:

$$F = Dice = \frac{2 \cdot \left|A \cap A_{réf}\right|}{\left|A\right| + \left|A_{réf}\right|}$$

De par son caractère multiplicatif, elle défavorise, à moyenne arithmétique égale, les configurations où *P* et *R* sont éloignés :

si $P = 0,5$ $R = 0,5$ alors *Moyenne* = 0,5 et $F = 0,5$
si $P = 0,3$ $R = 0,7$ alors *Moyenne* = 0,5 et $F = 0,42$

II.2.3.3 Modes de calcul

Il existe plusieurs façons de quantifier les cardinaux de A et $A_{réf}$. Nous en distinguerons trois, suivant que l'on calcule 1/ le nombre de binômes, 2/ les longueurs cumulées des binômes et 3/ les surfaces des binômes.

– *Nombre de binômes*

La méthode la plus simple consiste à calculer le nombre de couples contenus dans A, $A_{réf}$, et $A \cap A_{réf}$. Mais cette mesure ne prend pas en compte l'importance relative des phrases : l'alignement incorrect d'une phrase de deux cents mots sera traité de la même manière que l'alignement incorrect d'une phrase de 2 mots.

– *Longueurs cumulées des binômes*

On peut bien sûr pallier ce défaut par la prise en compte des longueurs de chaque binôme (exprimées en nombre de caractères ou en nombre de mots).

La longueur d'un couple (P,P') pouvant être exprimée par $l(P,P')=l(P)+l(P')$, on a dans l'exemple suivant :

$A_{réf} = \{(P_1;P'_1\,P'_2)(P_2;)(P_3;P'_3)(P_4\,P_5;P'_4)(P_6;P'_5\,P'_6\,P'_7)\}$

$A = \{(P_1;P'_1)(P_3;P'_3)(P_4;P'_4)(P_5;P'_5\,P'_6)(P_6;P'_7)\}$

$L(A_{réf})=l(P_1)+l(P'_1)+l(P'_2)+l(P_2)+l(P_3)+l(P'_3)+l(P_4)+l(P_5)+l(P'_4)+l(P_6)+l(P'_5)+l(P'_6)+l(P'_7)$

$L(A)=l(P_1)+l(P'_1)+l(P_3)+l(P'_3)+l(P_4)+l(P'_4)+l(P_5)+l(P'_5)+l(P'_6)+l(P_6)+l(P'_7)$

$L(A \cap A_{réf})=l(P_1)+l(P'_1)+l(P_3)+l(P'_3)+l(P_4)+l(P'_4)+l(P_6)+l(P'_7)$

Le problème d'une telle mesure est qu'elle n'est pas homogène avec la distributivité des couples d'alignement :

$$l(P_1;P'_1P'_2)=l(P_1)+l(P'_1)+l(P'_2) \neq l(P_1;P'_1)+l(P_1;P'_2)= 2\cdot l(P_1)+l(P'_1)+l(P'_2)$$

– *Surface des binômes*

Pour corriger cette anomalie, les consignes du projet ARCADE, s'inspirant des travaux de Simard & Isabelle (1996), suggèrent de pondérer chaque binôme (P_i,P'_j) par l'aire représentée par le produit des longueurs de P_i et P'_j. Par exemple :

$$\begin{aligned} card(A) &= card(\{(P_1;P'_1),(P_1;P'_2),(P_2;P'_0),(P_3;P'_3),(P_4;P'_4),(P_5;P'_4)\}) \\ &= Aire(P_1;P'_1)+Aire(P_1;P'_2)+Aire(P_2;P'_0)+Aire(P_3;P'_3)+Aire(P_4;P'_4)+Aire(P_5;P'_4) \\ &= l(P_1)\cdot l(P'_1)+l(P_1)\cdot l(P'_2)+l(P_2)\cdot l(P'_0)+l(P_3)\cdot l(P'_3)+l(P_4)\cdot l(P'_4)+l(P_5)\cdot l(P'_4) \end{aligned}$$

où $l(P)$ représente le nombre de caractères de P.

Si l'on prend :

$$A_{réf} = \{(P_1;P'_1P'_2)(P_2;)(P_3;P'_3)(P_4P_5;P'_4)(P_6;P'_5P'_6P'_7)\}$$
$$A = \{(P_1;P'_1)(P_2;P'_2)(P_3;P'_3)(P_4;P'_4)(P_5;P'_5P'_6)(P_6;P'_7)\}$$

l'évaluation de $A / A_{réf}$ peut être exprimée par les rapports des différentes aires représentées figure 17.

Comme le remarquent Simard et Isabelle (1996 : 4), le recours à cette représentation bidimensionnelle permet la comparaison de toute forme d'alignement, quelle qu'en soit la granularité, du moment que l'on calcule les aires au niveau des caractères : « Comme la segmentation en caractères est une donnée qui ne laisse aucune place à l'interprétation, les alignements mettront tous en rapport les mêmes objets, et seront de ce fait directement comparables. »

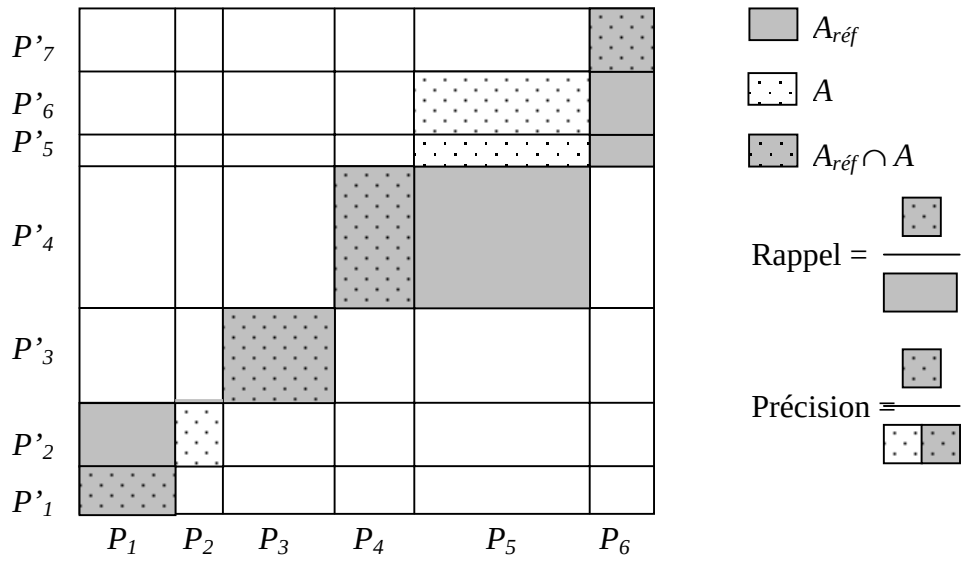


figure 17 : *évaluation d'un alignement basé sur le rapport de surfaces*

Quant aux alignements avec P_0 ou P'_0, plusieurs conventions sont envisageables, suivant qu'on considère ces deux phrases comme faisant partie de l'alignement ou non : soit on considère, dans la négative, que $card(P_0) = card(P'_0) = 0$, et on ne les fait pas intervenir dans le calcul, soit on pose arbitrairement : $card(P_0) = card(P'_0) = 1$, de manière à faire rentrer les appariements manquants dans le calcul du rappel, et les appariements indus dans le calcul de la précision. Dans ce dernier cas les phrases alignées avec P_0 ou P'_0 seront clairement distinguées des phrases non alignées du fait de l'incomplétude de l'alignement. Et leur présence n'affectera pas le rappel pour un algorithme capable de les identifier.

Une version plus stricte de l'évaluation consiste à ne considérer comme recevables que les surfaces coïncidant exactement avec l'alignement de référence. Dans cette optique, un binôme fragmentaire est erroné, car la traduction concerne l'intégralité des deux segments correspondants. Cette dernière mesure mettant sur un même plan les petites irrégularités (par exemple la séparation d'un titre et de sa numérotation) et les alignements totalement erronés, elle procède selon une logique du *tout ou rien* qui reflète mal, dans la réalité, le continuum des degrés de réussite des différents algorithmes : elle ne se prête pas, par conséquent, à un classement progressif des méthodes suivant leurs résultats.

II.2.3.4 Lien entre rappel, précision et granularité

La précision indique, nous l'avons vu, l'exactitude d'un alignement par rapport à l'alignement de référence. Mais elle dépend beaucoup de la *granularité* de l'alignement évalué. Par exemple, si l'alignement évalué s'arrête au niveau des paragraphes, en regroupant ensemble toutes les phrases d'un même paragraphe, la précision sera faible par rapport à un alignement de niveau phrastique, même si l'alignement évalué est correct.

A l'inverse, si la *granularité* de l'alignement de référence est plus grossière que l'alignement évalué, c'est le rappel qui aura tendance à se dégrader. Dans l'exemple d'une partie du corpus BAF comportant un glossaire (documentation de Xerox), l'alignement de référence (établi dans le projet ARCADE) a une granularité importante, puisque toutes les phrases du glossaire sont regroupées en un seul bloc. La surface occupée par ce binôme énorme est importante : 249 phrases en français par 244 phrases en anglais. Un alignement de ce même glossaire au niveau des phrases, même correct, obtiendra un rappel très faible, car il représente une surface correspondant à la diagonale de ce bloc.

II.2.4 Les indices d'alignement

On peut classer les différentes méthodes existantes en fonction de l'information prise en compte lors des traitements. Nous distinguerons principalement deux types d'indice d'alignement : les indices formels, et les indices lexicaux.

II.2.4.1 Les indices formels

Par indice formel, on entend ce qui a trait à la charpente externe du texte, à son articulation de surface : par exemple les découpages en chapitres, sous-chapitres, paragraphes, phrases. Pour des formats plus riches, comme SGML ou HTML, les balises peuvent également fournir des informations précieuses (cf. Bonhomme & Romary, 1996).

Dans la mesure où certaines traductions conservent les découpages à un niveau déterminé, par exemple au niveau des chapitres ou des paragraphes, ces articulations

peuvent constituer une forme de préalignement. Dans la pratique, la conservation des unités textuelles s'arrête, le plus souvent, au niveau des paragraphes.

II.2.4.1.1 Les indices basés sur les longueurs

La non-conservation du découpage formel n'empêche cependant pas d'en tirer parti : les schémas sont seulement un peu plus compliqués. Par exemple, en se situant au niveau des phrases :

- au lieu d'avoir des relations biunivoques, du type : 1 phrase = 1 phrase,
- on a des relations multiples du type : 1 phrase = 3 phrases, 2 phrases = 0 phrase, etc.

Gale & Church (1991) et P. Brown *et al*. (1991) ont proposé deux méthodes similaires pour extraire un alignement tenant compte de la probabilité des correspondances entre les longueurs, pour les regroupements mis en jeu dans chaque binôme. L'hypothèse sous-jacente à ces deux méthodes peut se formuler simplement : il est plus probable de traduire une phrase longue par une phrase longue, et une phrase courte par une phrase courte.

Ainsi, pour chaque binôme :

$$B = (G;G') = (P_n...P_m\,;\, P'_p...P'_q)$$

où G et G' sont respectivement des groupes de phrases de T et T', avec $L(G) = l$ et $L(G')=l'$, on peut estimer la probabilité que G soit aligné avec G' sachant l et l'.

Notons que la fonction de longueur $L(x)$ peut être exprimée en mots (P. Brown *et al*., 1991) ou en caractères (Gale & Church, 1991), sans rien changer au principe de la méthode.

Gale & Church sont partis de l'hypothèse que le rapport des longueurs entre des unités textuelles alignées, exprimées en caractère, suit approximativement une distribution normale. Plus formellement, un caractère du texte source produit X caractères cibles :

$$1\ car.\ source \rightarrow X\ car.\ cibles$$

où *X* est une variable aléatoire ayant une distribution normale, de moyenne *c* et de variance S^2.

Pour une phrase source de *l* caractères, on peut considérer la longueur *l'* de la phrase cible comme résultant de la somme de *l* variables aléatoires X_i.

D'après le théorème de la limite centrale, la variable aléatoire δ définie au niveau de la phrase :

$$\delta = \frac{(X_1 + X_2 + .. + X_l) - l \cdot c}{\sqrt{s^2 l}} = \frac{(l' - l \cdot c)}{\sqrt{s^2 l}} \quad (5)$$

suit une loi normale centrée réduite (moyenne nulle, variance 1).

De la probabilité de l'alignement sachant une valeur donnée de δ, on peut tirer une mesure de distance entre *G* et *G'*:

$$D_{GC}(G,G') = - \textit{log}\ \text{prob}(\text{alignement} / \delta) \quad (6)$$

L'application du logarithme permet de transformer le produit des probabilités en somme, et le signe moins de transformer la similitude en distance. On peut donc calculer la distance totale d'un chemin, correspondant au logarithme du produit des probabilités, comme la somme des distances des binômes constituant ce chemin. Trouver un alignement probable revient alors à un problème de recherche d'extremum : il s'agit de déterminer le chemin maximisant la probabilité totale du chemin, i.e. minimisant la somme des distances des binômes.

D'après le théorème de Bayes, on a :

$$prob(alignement / \delta) = \frac{prob(\delta / alignement) \cdot prob(alignement)}{prob(\delta)} \quad (7)$$

Gale & Church supposent que la $prob(\delta)$ correspond à une constante, qui peut donc être négligée car elle intervient de manière identique pour tous les chemins possibles[129].

[129] Négliger cette probabilité est cependant discutable, car tous les chemins n'impliquent pas nécessairement le même nombre de binômes.

Le deuxième facteur, *prob*(*alignement*) est assimilé à la probabilité générale d'observer telle ou telle transition. Les auteurs lui affectent des valeurs moyennes issues de l'observation[130].

Enfin, comme δ suit une distribution normale on peut donner l'estimation suivante : *prob*(δ/*alignement*)=*2*(*1-prob*(|δ|)), d'où :

$$prob(\delta / alignement) = 2 \cdot (1 - \int_{-\infty}^{\delta} \frac{e^{\frac{-t^2}{2}}}{\sqrt{2\pi}} dt) \quad (8)$$

La méthode développée par Gale & Church fait référence dans de nombreux travaux d'évaluation ; nous y référerons désormais par l'abréviation GC.

II.2.4.1.2 Les transfuges

Par *transfuges*, on désigne toutes les chaînes de caractères invariantes dans le passage à la traduction : les noms propres, les données numériques, certains sigles, les numéros de chapitre, etc. Les transfuges peuvent aussi s'étendre à des noms communs identiques dans les deux langues, quand il y a emprunt. Ces indices peuvent ainsi concerner le lexique, mais sous un aspect purement formel : ce qui définit un transfuge, c'est l'identité de deux chaînes alphanumériques, quelle qu'en soit la nature.

Ces indices formels peuvent permettre d'apparier des points isolés entre les deux textes, et former ainsi un réseau de points de capiton, ou *points d'ancrage*. Leur densité n'étant pas très importante, ils sont utilisés le plus souvent pour obtenir un *préalignement* : ils permettent de découper les deux textes en segments parallèles, parfois appelés « îlots de confiance », à l'intérieur desquels on applique ensuite des algorithmes plus coûteux et plus précis (cf. § II.2.5.3) Les transfuges sont adaptés pour servir de guide à d'autres méthodes, et ils permettent d'augmenter sensiblement l'efficacité en limitant l'espace de recherche avant l'application d'algorithmes de complexité polynomiale (en O(n^2)).

La plupart des méthodes ont recours à ce type d'indice. Par exemple, dans les travaux de Gale & Church, les textes sont censés être déjà préalignés au niveau des

[130] pour leur corpus : *p*(1:1)=0,89, *p*(1:0) = *p*(0:1) = 0,0099, *p*(2:1) = *p*(1:2) = 0,089, *p*(2:2)=0,011

paragraphes : les marqueurs de paragraphes sont alors exploités comme des transfuges et donnent lieu à une série de points d'ancrage.

P. Brown *et al*. (1991 : 170) proposent de hiérarchiser les points d'ancrage. Ils établissent une distinction entre points d'ancrages majeur et mineur, ces derniers étant plus nombreux mais moins robustes vis-à-vis des insertions et des omissions (la distinction est établie de manière *ad hoc* en fonction des caractéristiques de leur corpus : par exemple, sont reconnus comme points d'ancrage mineurs des commentaires du type « *Author = Mr. Speaker* », fréquents dans le corpus *Hansard*). Le préalignement s'effectue donc en deux étapes, une première où l'on tient compte des points d'ancrages majeurs et une seconde où l'on tient compte des mineurs. La première étape fournit une segmentation en sous-sections, à l'intérieur desquelles on compte les points d'ancrages mineurs. Si ceux-ci apparaissent le même nombre de fois et dans le même ordre de part et d'autre, la sous-section est validée, ainsi que toutes les nouvelles sous-sections impliquées par les points d'ancrages mineurs. Lorsque ce n'est pas le cas, la section est purement et simplement rejetée de l'alignement (10 % du corpus est ainsi rejetée).

Comme tout préalignement, les points d'ancrage doivent aboutir à des appariements très sûrs, car la bonne exécution de l'alignement subséquent en dépend. C'est la contrainte majeure qui pèse sur leur prise en compte : ils n'ont d'intérêt que s'ils mènent à une précision voisine de 100 %, même avec un rappel faible.

II.2.4.1.3 Les Cognats

Le recours aux mots apparentés, ou *cognats*, est motivé par la double ressemblance qui les caractérise : ressemblance graphique et sémantique. Etant susceptibles d'être employés comme équivalents traductionnels, ils se rapprochent des indices lexicaux étudiés plus loin. Mais dans la mesure où c'est leur ressemblance graphique qui sert d'indice, nous les rangeons dans les indices formels, au même titre que les transfuges.

Reste que la notion intuitive de *ressemblance* peut recevoir des définitions variables, et aboutir à des délimitations peu claires.

– d'une part, sur le plan du contenu, les écarts sémantiques entre deux formes apparentées décrivent tout un continuum, de l'identité à la différence. Par exemple, les cognats français / anglais du tableau 8 présentent différents degrés d'éloignement sémantique :

Français	**Anglais**
identité	*identity*
route	*route* (« itinéraire »)
injure	*to injure* (« blesser »)
sensible	*sensible* (« raisonnable »)
façon	*fashion* (« mode »)
habit	*habit* (« habitude »)
ménagère	*manager*

tableau 8 : *degrés d'éloignement sémantique liés à la cognation*

Ces distorsions sémantiques sont dues à des évolutions divergentes, mais ces écarts peuvent aussi apparaître en synchronie, les opérations de composition et de dérivation n'étant pas toujours clairement motivées[131]. Ces écarts sont problématiques car il est difficile d'effectuer un *calcul* de la distance sémantique, et la limite entre *ressemblant* et *non ressemblant* est nécessairement arbitraire.

– sur le plan de l'expression, la difficulté est la même : la ressemblance peut être appréhendée le long de différents axes, suivant qu'on s'attache à la structure syllabique, au squelette consonantique, à la présence de certains morphèmes, à des règles de transformation phonologiques, ou simplement à l'identité de certains groupes de caractères.

[131] p. ex. le contenu sémantique de la racine est trop abstrait pour qu'un lien soit perceptible entre *abroger*, *déroger* , *subroger*, *proroger*, *arroger*, *interroger*, *etc.*

Si la notion de cognat est vague, la plupart des méthodes basées sur l'identification des cognats s'appuient sur des approximations encore plus floues. En effet, ce ne sont pas des dictionnaires de cognats qui sont employés, mais des critères purement formels : en général, les cognats sont identifiés par la présence de chaînes de caractères partagées, les *n-grammes* (chaînes de longueur *n*).

Avant de pouvoir juger de la validité de ces approximations, il semble essentiel de se donner une définition rigoureuse de ce qu'on entend par cognat dans le cadre précis de la comparaison de lexies supposées équivalentes. A partir de cette base, nous pourrons évaluer l'opportunité des critères de surface employés, et déterminer en second lieu leur efficacité opératoire quant à l'alignement.

Etant donné les difficultés inhérentes à la notion de ressemblance, tant sur le plan du contenu que sur celui de l'expression, nous nous contenterons d'une définition opératoire visant à en contourner l'aspect subjectif. Deux unités *U* et *U*' sont des cognats si et seulement si :

1. *U* et *U*' ont un lien étymologique (emprunt, origine commune) perceptible dans leur signifiant.
2. on peut trouver deux phrases (*P*,*P*') dont l'une est la traduction de l'autre, et dans lesquelles *U* et *U*' sont en relation d'équivalence.

Les *transfuges*, invariants dans la traduction, peuvent être considérés comme des cognats particuliers.

Le critère 1, ainsi que la notion de transfuge ne posent pas de problèmes significatifs, l'existence de liens étymologiques permettant de donner une assise objective à la notion confuse de ressemblance. Une question se pose néanmoins : tous les étymons doivent-ils entrer en ligne de compte ou bien seulement les racines ? Après tout, si deux lexies sont des équivalents possibles, le fait qu'elles soient apparentées par le biais d'affixes n'enlève rien à leur fonction heuristique d'indice : nous les considérerons donc comme des cognats.

En revanche, vis-à-vis du critère 2, décider de la possibilité de traduire une lexie par une autre implique des difficultés :

– d'une part, un mot peut être traduit par une unité polylexicale : par exemple (angl.) *because* ↔ (fr.) *à cause*. Une solution consiste à ne prendre en compte que des formes simples, en se limitant à celles qui portent l'étymon commun : *because* ↔ *cause*.

– d'autre part, il est parfois difficile de déterminer si un mot *peut* en traduire un autre, car tout dépend du contexte de la traduction. Or, des mots d'étymologie commune mais de sens différents peuvent, dans un certain contexte, se retrouver en relation de traduction : par exemple, les mots (angl.) *importation* et (fr.) *export* peuvent être considérés comme des cognats bien que leurs contenus sémantiques présentent une différence d'orientation. On peut leur imaginer le contexte de traduction suivant :

fr. : *Il fait de l'export vers les USA*
angl. : *He makes importations from France*

Entre (angl.) *translation* et (fr.) *transfert*, l'écart sémantique est plus grand. Et pourtant, ces deux unités sont apparentées (*translatum* est le participe passé latin de *transferre*), et il est possible de trouver des contextes à l'intérieur desquels ils sont en relation de traduction :

fr. : *J'ai effectué un transfert du français vers l'anglais*
angl. : *I did a translation from French to English*

Ici *transfert* est employé en tant que terme *générique* pouvant subsumer le terme anglais plus précis. Ce genre de saut du générique au spécifique n'est pas rare lors de la traduction.

Ainsi, pour déterminer l'équivalence potentielle de deux unités, plusieurs méthodes sont envisageables : la plus rigoureuse consiste à se baser sur l'attestation des traductions existantes ; la plus rapide revient à fabriquer un contexte *ad hoc*, permettant de mettre les unités en relation d'équivalence lorsque c'est possible. Cette dernière méthode comporte

une grande part d'arbitraire dans la mesure où elle repose sur la subjectivité du locuteur, qui doit décider de l'acceptabilité relative de l'équivalence ainsi produite.

Dans la suite de ce travail, nous appliquerons le premier mode de validation des cognats, fondé sur l'observation des équivalences à travers un corpus donné.

II.2.4.1.3.1 Méthode d'identification automatique des cognats

L'exploitation des données étymologiques requérant des dictionnaires spécifiques, il est clair que les systèmes se basant sur les indices superficiels doivent faire l'économie de cette information (sinon, pourquoi ne pas utiliser directement un dictionnaire bilingue ?).

Concrètement, la plupart des techniques tirant parti de cette information se basent sur une approximation très abrupte : tous les mots comportant une sous-chaîne commune de 4 caractères consécutifs, i.e. un 4-gramme, sont considérés comme cognats potentiels. Par exemple, chez Simard *et al*. (1992), les cognats potentiels appartiennent aux catégories suivantes :

1. les couples de chaînes *alphanumériques*[132] identiques.
2. les couples de mots commençant par 4 caractères identiques.
3. les signes de ponctuations identiques.

De même, avec la méthode baptisée *Char-align* (K.W. Church, 1993), tous les couples partageant le même 4-gramme sont des candidats cognats (et participent donc à la production d'un point), à condition que les unités concernées aient une fréquence inférieure à un certain seuil. Dans cette méthode, où chaque couple de cognats est susceptible de produire un point, ensuite filtré au sein du nuage obtenu, les mots les plus fréquents sont jugés inexploitables pour l'alignement.

[132] Par alphanumérique, nous désignerons dorénavant les chaînes contenant au moins un caractère numérique : 1969, A4, RS232, etc.

De manière plus élaborée, Débili & Sammouda (1992 : 522) proposent de calculer un indice de ressemblance superficielle en fonction des sous-chaînes maximales : ils utilisent l'algorithme de Bellman (Bellman, 1957 ; Laurière, 1979) pour la comparaison dynamique de deux mots. Cet indice se calcule sous la forme suivante :

$$N = \left(1 - \frac{|l_1 - l_2|}{(l_1 + l_2)}\right) * \sum_t n(t) * t^2 \qquad (9)$$

où l_1 et l_2 sont les longueurs des chaînes c_1 et c_2 à comparer, $n(t)$ est le nombre de sous-chaînes communes de longueur t. Cette mesure n'indique pas si deux mots sont des candidats cognats ou non, mais elle permet d'évaluer leur *degré* de ressemblance.

II.2.4.1.3.2 Exploitation

Après s'être doté d'indicateurs permettant d'identifier les couples de cognats potentiels, moyennant des approximations se basant sur des données de surface, il reste à intégrer cette information dans un score comptabilisant les cognats au niveau des couples de phrases.

Simard *et al*. (1992) proposent d'utiliser la mesure suivante pour évaluer la *cognation* (de l'anglais "*cognateness*"), c'est-à-dire la densité de cognats apparaissant à l'intérieur d'un binôme (*P*,*P'*) :

$$\gamma = \frac{2c}{l + l'} \qquad (10)$$

où *l* et *l'* sont les longueurs respectives de *P* et *P'*, et *c* est le nombre maximal[133] de couples de cognats identifiables entre *P* et *P'*.

M. Davis *et al*. (1995), dans la même perspective, utilisent une version simplifiée de γ : le nombre total (et non plus maximal) de couple de cognats relevés entre les deux phrases, divisé par le produit des longueurs :

[133] C'est-à-dire, pour tous les appariements des mots de *P* avec des mots de *P'*, où chaque mot n'intervient que dans un seul couple.

$$\gamma = \frac{c_{tot}}{l \cdot l'} \tag{11}$$

En comptabilisant les cognats corrects (identifiés par les auteurs), sur 102 phrases anglaises et 94 phrases françaises (extraites du Hansard) Simard *et al*. observent une valeur moyenne de 0,21, pour γ, entre des phrases équivalentes, et de 0,06 entre des phrases prises au hasard. Lorsque le même calcul est effectué à partir de l'identification automatique (en comptant les nombres, les 4-grammes et la ponctuation), les valeurs moyennes passent à 0,3 pour des phrases équivalentes et 0,09 entre des phrases prises au hasard.

Ces résultats semblent établir deux faits : tout d'abord que la mesure de *cognation* permet effectivement d'établir une discrimination entre les couples de phrases qui sont traductions mutuelles et les autres ; ensuite que l'identification des cognats au moyen d'indices superficiels est opérante, car la discrimination est tout aussi bonne avec ceux-ci, sinon meilleure. Toujours d'après les auteurs, la bonne tenue de cette identification est expliquée par le fait qu'elle permet d'ignorer les mots très fréquents (ceux-ci étant souvent courts), qui sont source de confusion.

A partir de la cognation, Simard *et al*. développent une mesure de distance, de façon analogue à la mesure de distance basée sur le rapport des longueurs.

Pour estimer la probabilité d'obtenir *c* couples de cognats entre deux phrases équivalentes de longueurs *l*, on peut recourir au modèle théorique d'une distribution binomiale[134] :

$$P(c/l,t) = C_l^c \, p_t^{\,c} (1 - p_t)^{l-c} \tag{12}$$

p_t étant la probabilité qu'un mot soit en relation d'équivalence traductionnelle avec un mot apparenté; *l* étant la longueur moyenne des deux segments, et *t* l'événement exprimant que les deux segments sont équivalents.

[134] Le paradigme d'une distribution binomiale le suivant : pour un événement ayant une probabilité *p* à chaque tirage, la probabilité d'obtenir *q* fois l'événement en effectuant *n* tirage est de : $C_n^q \, p^q (1-p)^{n-q}$

La distance proposée est alors issue du rapport de vraisemblance $P(c/l,t) / P(c/l)$, multiplié par la probabilité $P(a)$ du type de transition, comme chez Gale & Church (avec les mêmes transitions 1:1, 1:0, 0:1, 1:2, 2:1 ou 2:2) :

$$Score(a,c,l) = -\log\left[\frac{P(c/l,t)}{P(c/l)}.P(a)\right] = -\left[c.\log\frac{p_t}{p}\right] - \left[(l-c).\log\frac{1-p_t}{1-p}\right] - \log P(a) \qquad (13)$$

p_t et p sont données par les valeurs moyennes de γ déterminées empiriquement :

$$p_t = 0{,}3 \text{ et } p = 0{,}09.$$

De façon plus synthétique, le score peut s'écrire sous la forme :

$$Score'(a, \gamma, l) = l\,(A\,\gamma + B) - log\ P(a) \qquad (14)$$

Dans cette version condensée, on a $A > 0$ et $B > 0$ (puisque $0 \leq p < p_t \leq 1$), ce qui assure la diminution du score avec γ, la pondération par l (la longueur du couple mis en jeu) exprimant l'importance de l'appariement.

Le coût total d'un chemin est par suite calculé comme la somme des scores des appariements correspondants.

II.2.4.2 Les indices lexicaux

Par indices lexicaux, nous désignons les informations relatives à la traduction des lexies, utilisées au palier supérieur afin de déterminer la probabilité d'apparier deux phrases. L'identification des cognats constituait un type particulier d'indice lexical, que nous avons considéré comme un indice formel dans la mesure où il découle d'éléments purement superficiels.

Fait paradoxal, nous verrons que l'extraction des correspondances au niveau lexical est permise par l'alignement au niveau des phrases : on peut alors déceler une circularité dans le fait de baser l'alignement des phrases sur l'alignement des mots en même temps que l'alignement des mots sur celui des phrases. Mais comme l'ont noté Débili & Sammouda (1992 : 521), cette circularité est à la base d'un processus itératif, où les deux étapes sont appliquées successivement : « Pour obtenir un appariement fin des mots il faut

apparier les phrases ; pour apparier les phrases on peut se contenter d'un appariement grossier des mots. » Pour alimenter la première étape de l'alignement des phrases, nous verrons qu'il suffit en effet d'extraire des correspondances relativement peu nombreuses au niveau lexical.

Ces informations traductionnelles peuvent être définies *a priori*, en étant codées à l'intérieur de dictionnaires bilingues. Certains systèmes d'alignement font appel à ce genre de ressource (Fluhr, Bisson & Elkateb in Véronis, 2000 §9). Malgré l'intérêt présenté par ce genre d'approches, nous les avons négligées dans notre étude : du fait de nos orientations méthodologiques nous privilégions les techniques « alinguistiques ».

II.2.4.2.1 Extraction de correspondances lexicales

Le plus souvent, c'est la comparaison des distributions lexicales qui permet d'extraire des informations utiles pour l'étude des correspondances entre unités. Lorsque les textes sont déjà alignés, les distributions de deux unités lexicales *u* et *u*' peuvent être caractérisées par leurs *occurrences* (séparément, dans chaque texte) et leurs *cooccurrences* (ensemble, dans les mêmes binômes). Dans l'exemple de la figure 18, on compte $n_1 = 5$ occurrences de *u*, $n_2 = 4$ de *u*' et $n_{12} = 3$ cooccurrences des deux unités.

(…*u*…, … …)
(…*u*…, …*u*'…)
(… …, … …)
(…*u*…, …*u*'…)
(… …,… …)
(… …, …*u*'…)
(…*u*…, … …)
(…*u*…, …*u*'…)

figure 18 : *occurrences et cooccurrences de deux unités* ($n_1 = 5$, $n_2 = 4$, $n_{12} = 3$)

A partir de ces observations, on peut estimer la probabilité d'obtenir n_{12} cooccurrences par le seul jeu du hasard (hypothèse d'indépendance des occurrences de *u* et

u'). Si cette probabilité est voisine de zéro, on peut alors supposer qu'il existe un lien de dépendance entre *u* et *u*'. A priori, le seul lien sous-jacent aux deux textes étant l'équivalence traductionnelle, on en déduira que *u* et *u*' sont probablement équivalents.

En prenant la réciproque de l'affirmation de Débili & Sammouda, on peut affirmer qu'un alignement grossier au niveau des phrases permet d'obtenir un alignement un peu moins grossier des mots : par exemple, à partir d'un préalignement au niveau des paragraphes (ou de zones plus larges), on peut observer des distributions d'occurrences et de cooccurrences assez marquées pour en extraire des correspondances lexicales. Si un tel préalignement n'est pas disponible, il est néanmoins possible de calculer les cooccurrences en se basant sur des zones plus étendues, « présumées alignées ». Par exemple, pour obtenir un préalignement grossier, Fung & Church (1994) proposent de découper chaque texte en *K* zones d'égale importance : ces zones sont alors considérées comme étant hypothétiquement alignées. Nous verrons plus loin (§ III.2.1.1.1) qu'il existe d'autres façons de calculer les cooccurrences, en se basant sur la distance des points à la diagonale.

Une fois établi le mode d'extraction des informations distributionnelles, on peut calculer des indices statistiques visant à quantifier le degré de dépendance entre deux unités. Le plus répandu de ces indices est l'*information mutuelle*, introduite par Shannon (1949), qui résulte du rapport entre les cooccurrences observées de deux unités *u* et *u*', et les cooccurrences attendues par le simple jeu du hasard. Cet indicateur est couramment utilisé pour mesurer l'association entre deux événements, comme la cooccurrence, tant sur le plan monolingue (Church & Hanks : 1990) que bilingue (Gaussier & Langé :1995).

$$IM = \log \frac{p(u,u')}{p(u) \cdot p(u')} \qquad (15)$$

Le numérateur représente la probabilité observée de cooccurrence de *u* et *u*'. Le dénominateur est la probabilité estimée de cooccurrence, en supposant l'indépendance des deux événements « occurrence de *u* » et « occurrence *u*' ». Lorsque le nombre de cooccurrences observées dépasse de façon significative le nombre de cooccurrences attendues par hasard, *IM* prend une valeur positive importante. L'association est donc forte

entre les deux événements. A l'inverse, si le nombre de cooccurrences observées est très inférieur au nombre de cooccurrences attendues, *IM* prend une valeur négative importante. Dans ce cas, il y a inhibition entre les deux événements. Lorsque les deux événements sont indépendants, *IM* est nulle et les cooccurrences sont imputables au hasard.

Mais l'information mutuelle a un défaut : elle surévalue l'association entre des formes de faible fréquence. Par exemple, deux hapax qui cooccurrent ensemble obtiennent une information mutuelle très importante, alors qu'un événement si ponctuel ne permet pas de tirer de conclusion quant à leur lien de dépendance.

Fung & Church (1994) proposent de recourir à une mesure modifiée, le *t-score* , pénalisant les basses fréquences :

$$TS = \frac{p(u,u') - p(u) \cdot p(u')}{\sqrt{\frac{1}{K} p(u,u')}} \quad (16)$$

où *K* correspond au nombre de tirages (par exemple, le nombre de zones préalignés).

Pour $TS > 1{,}65$, la probabilité d'avoir des cooccurrences fortuites est inférieure à 0,05. Lorsque deux hapax cooccurrent, on obtient $TS \approx 1$, ce qui n'est pas significatif.

Kay & Röscheisen (1993) utilisent un autre score d'association, calculé suivant les phrases considérées comme « alignables » (c'est-à-dire proches de la diagonale et situées entre les points d'ancrages issus des étapes précédentes du processus d'alignement). Le score, entre *u* et *u*', n'est autre qu'un coefficient Dice :

$$Dice = \frac{2c}{n_1 + n_2} \quad (17)$$

où n_1 est le nombre d'occurrence de *u*, n_2 le nombre d'occurrence de *u*' et *c* est la taille de la suite la plus longue de couples disjoints de phrases alignables (*P*,*P*') comportant respectivement *u* et *u*'. Cette mesure, variant entre 0 et 1, indique le degré d'association entre les deux formes.

Pour obtenir des couples de candidats, Kay & Röscheisen ne retiennent que les associations obtenant un score élevé. Les appariements obtenus sont ensuite traités de façon hiérarchisée, suivant des classes de fréquence : sont d'abord pris en considération les

appariements des classes de mots de haute fréquence, pour qui cette mesure est plus fiable. A l'intérieur de chaque classe, les appariements sont traités suivant leur score, du plus fort au plus faible. Les mots de basse fréquence, du type hapax, sont ignorés, car le coefficient Dice n'est pas significatif pour eux, à l'instar de l'information mutuelle.

II.2.4.2.2 Utilisation des correspondances pour l'alignement

Il existe plusieurs manières d'intégrer les correspondances ainsi obtenues dans l'algorithme d'appariement des phrases.

– *Méthodes basées sur un score d'association*

Dans la méthode décrite par Kay & Röscheisen, les couples d'unités sont ordonnés en fonction de leur degré de pertinence, et sont injectés dans l'ordre pour apparier les phrases, en suivant deux critères :

- la paire (*u*,*u*') permet d'aligner (*P*,*P*') si *P* n'est pas alignable avec une autre phrase *P*' contenant aussi *u*'.
- si la paire (*u*,*u*') engendre un alignement incompatible avec les points d'ancrage déjà obtenus (croisement), elle est rejetée.

De façon générale, les couples d'unités obtenus permettent de générer des *points d'ancrage* potentiels, dont on ne retient que ceux qui sont probablement situés sur le chemin d'alignement : ainsi, le plus souvent, on élimine d'emblée tous les points qui se situent loin de la diagonale. Dans la méthode *K-Vec* (Fung & Church, 1994), ce filtrage est effectué sur des bases statistiques, le chemin étant interpolé à partir des zones où la densité des points est la plus forte.

Outre ces nuages de points, les correspondances lexicales peuvent être intégrées dans des mesures exprimant la distance de deux segments, comme celles développées pour les cognats.

– *Méthodes basées sur un modèle de traduction*

Il existe enfin des méthodes plus sophistiquées, où la mesure de distance est basée sur un modèle complet de génération de la traduction mot par mot. En s'inspirant des modèles statistiques élaborés par P. Brown *et al*. (1990, 1993), S. Chen (1993) a développé une méthode d'alignement originale, obtenant des résultats intéressants.

Dans cette dernière méthode, de même que dans le système de Kay & Röscheisen, l'évaluation des chemins potentiels est effectuée en fonction des appariements de mots, eux-mêmes extraits des approximations précédentes du meilleur chemin. Comme chez Brown *et al*. (1990), le calcul de la probabilité d'un chemin se fait suivant un modèle stochastique génératif [135]: un chemin est conçu comme le résultat d'une suite de tirages aléatoires indépendants, partant des unités les plus petites (appariement lexical) pour la construction progressive des unités plus larges (appariements de phrases, puis chemin entier).

L'élaboration de ce modèle suit un certain nombre d'étapes :

– dans la génération de deux phrases, on tire d'abord le nombre de couples de mots appariés, i.e. la longueur de l'alignement des mots, notée *l*.

– puis on tire *l* couples pour aboutir à un alignement des mots : on construit $B = \{b_1,b_2,..,b_l\}$ où chaque b_i (pour « *word bead* ») correspond à un appariement de mots de type 1:0[136] si $b_i = (e;)$ ou 1:1 si $b_i = (e;f)$ ou 0:1 si $b_i = (;f)$, où *e* représente un mot en anglais et *f* un mot en français. Nous nommerons *assignement* un ensemble d'appariements de mots entre deux phrases. La probabilité globale d'un assignement *B* résulte du produit des probabilités des appariements des mots b_i (considérés comme des tirages aléatoires indépendants), multiplié par la probabilité de la longueur de cet alignement. On obtient donc :

[135] sauf que dans le modèle de Brown et al., la génération d'une phrase cible est fonction d'une phrase source, tandis que dans le présent modèle les deux phrases sont générées simultanément.

[136] on utilisera les parenthèses pour distinguer les transitions (appariements de phrases) des appariements lexicaux.

$$P(B) = \frac{p(l)}{N_l} \prod_{i=1}^{l} p(b_i) \quad (18)$$

où N_l est une constante de normalisation.

– pour un assignement donné *B*, on peut ensuite construire deux phrases (*E* ;*F*) correspondant au tirage d'un ordre spécifique des mots de *B* (car *B* est un ensemble de couples non ordonnés). On peut ainsi évaluer la probabilité du couple de phrases ainsi construit :

$$P((E;F),B) = \frac{p(l)}{N_l n!m!} \prod_{i=1}^{l} p(b_i)$$

où *n!* et *m!* correspondent au nombre de permutations possibles des mots de *E* et *F*. On fait ainsi l'hypothèse simplificatrice que tous les ordres syntaxiques sont équiprobables.

– on en déduit la probabilité générale de l'appariement (*E* ;*F*), comme la moyenne des probabilités de tous les assignements *B* possibles, à l'intérieur de *E* et *F* :

$$P((E;F)) = \sum_{B} \frac{p(l)}{N_l n!m!} \prod_{i=1}^{l(B)} p(b_i)$$

– la génération d'un chemin entier peut ensuite être conçue comme une succession d'appariements de phrases (des binômes, dans notre propre terminologie) tirés indépendamment. De manière générale, pour l'appariement des phrases 5 cas de figures sont retenus : (1:0), (0:1), (1:1), (2:1) et (1:2). Dans le calcul de la probabilité du chemin, chaque tirage d'un binôme doit être pondéré par la probabilité du cas de figure : $p_{1:0}$, $p_{0:1}$, $p_{1:1}$, $p_{2:1}$ et $p_{1:2}$. On obtient alors, dans le cas général :

$$P((E;F)) = p_{1:1} \sum_{B} \frac{p_{1:1}(l)}{N_l n!m!} \prod_{i=1}^{l(B)} p(b_i)$$

$$P((E;\ )) = p_{1:0} \frac{p_{1:0}(n)}{N_n n!} \prod_{i=1}^{n} p(e_i) \qquad P((\ ;F)) = p_{0:1} \frac{p_{0:1}(m)}{N_m m!} \prod_{i=1}^{m} p(f_i)$$

dans les cas d'appariements vides ((1:0) ou (0:1)), il n'existe bien sûr pas différentes solutions d'appariement lexical : tous les b_i sont nécessairement du type (e;) ou (;f).

$$P((E_1, E_2; F)) = p_{2:1} \sum_B \frac{p_{2:1}(l)}{N_l n_1! n_2! m!} \prod_{i=1}^{l(B)} p(b_i)$$

$$P((E; F_1, F_2)) = p_{1:2} \sum_B \frac{p_{1:2}(l)}{N_l n! m_1! m_2!} \prod_{i=1}^{l(B)} p(b_i)$$

Reste à estimer les paramètres du modèle. Les distributions des longueurs des phrases sont modélisées suivant une loi de Poisson :

$$p_{1:0}(l) = \frac{\lambda_{1:0}^{\ l}}{l! e^{\lambda_{1:0}}}$$

avec un lien de dépendance entre les différents cas de figure :

$$\lambda_{1:0} = \lambda_{0:1} = \frac{\lambda_{1:1}}{2} = \frac{\lambda_{2:1}}{3} = \frac{\lambda_{1:2}}{3}$$

Enfin, Chen fait l'hypothèse que les distributions des appariements de mots sont identiques dans les appariements de phrases (1:1), (1:2), (2:1). De même, on assimile ces probabilités dans les appariements de phrases (1:0) et (0:1). Ces derniers sont d'ailleurs déduits des appariements de mots 1:0 ou 0:1, dans le cas des appariements de phrases (1:1), (1:2), (2:1) :

$$p_E(e) = \frac{p_b(e)}{\sum_{e' \in B_e} p_b(e')}$$

$$p_F(f) = \frac{p_b(f)}{\sum_{f' \in B_e} p_b(f')}$$

où p_E représente la probabilité de (e;) dans le cas (1:0), p_F celle de (;f) dans le cas (0:1), et p_b celles de (e;) ou (;f) dans les cas (1:1), (1:2), (2:1).

Chen recourt ensuite à une approximation permettant de simplifier les calculs, en assimilant la somme des probabilités de tous les assignements possibles à la probabilité du meilleur assignement[137] :

$$P((E;F)) = p_{1:1} \sum_{B} \frac{p_{1:1}(l)}{N_l n! m!} \prod_{i=1}^{l(B)} p(b_i)$$

$$\approx p_{1:1} \max_{B} \left(\frac{p_{1:1}(l)}{N_l n! m!} \prod_{i=1}^{l(B)} p(b_i) \right)$$

Il est trop coûteux de calculer ce maximum (la complexité est une fonction factorielle de *l*), mais une heuristique simple permet d'atteindre un résultat approché :

1. on commence avec un assignement où tous les appariements sont vides,
2. puis l'on cherche les deux appariements de mots 1:0 et 0:1 qui aboutissent au meilleur gain de probabilité s'ils sont remplacés par un appariement 1:1,
3. on réitère l'opération jusqu'à ce qu'aucun gain ne soit plus possible.

Nous verrons plus loin comment les paramètres sont initialisés puis réestimés dans un processus itératif de type EM (de l'anglais *Expectation-Maximisation*, *cf.* Dempster *et al*., 1977).

II.2.5 Architectures

Nous avons passé en revue les différents types d'information, ou indices, utilisables dans le calcul d'un alignement probable. Nous dissocions délibérément les indices des algorithmes dans lesquels ils sont usuellement implémentés, car ils peuvent tous s'intégrer indépendamment dans chaque type d'architecture.

Nous distinguerons trois familles d'algorithmes : matriciels, itératifs, et dynamiques. Nous chercherons à déterminer à quel type de tâche ils sont adaptés, quel est leur coût, et comment ils peuvent se combiner.

[137] Cette approximation est classique : elle s'apparente au modèle de P. Brown *et al*. (1993 : 293) et à la recherche de l' « assignement de Viterbi », proposée par Melamed (1998a).

II.2.5.1 Algorithme matriciel

Par architecture matricielle, nous désignons tous les systèmes qui se basent sur le calcul d'une matrice (M_{ij}) où sont stockés les coefficients d'association entre les éléments de T et T' : ces éléments peuvent être les phrases P_i et P'_j, ou les unités elles-mêmes u_i et u'_j, numérotées par leur ordre d'apparition (leurs coordonnées dans les textes). Ces coefficients d'association sont ensuite filtrés, de manière statistique, de manière à aboutir à un ensemble de points situés sur le chemin. Ces points peuvent ultérieurement servir de points d'ancrage pour un alignement complet, ou générer d'emblée un alignement entre les sections qu'ils délimitent.

La méthode *Char-align* présentée par Church (1993) illustre parfaitement le principe matriciel : l'algorithme effectue un comptage des 4-grammes observés entre tous les mots situés à l'intérieur d'une bande autour de la diagonale. La matrice se situe donc au niveau des mots, et est basée sur une mesure d'association du type tout ou rien : si deux mots contiennent un 4-gramme commun, ils génèrent un point, sinon rien. On obtient ainsi une matrice de points appelée « *dotplot* ». Des techniques de traitement du signal (filtrage basse fréquence « *low-pass* » et seuillage) permettent de filtrer les points pour retenir un nuage resserré autour du chemin d'alignement. Une pondération des points, inversement proportionnelle à la fréquence des 4-grammes, permet de diminuer l'effet de dispersion dû aux mots très fréquents.

Au final, le chemin candidat est celui qui totalise le plus grand poids moyen (c'est-à-dire la somme de tous les poids des points situés sur le chemin divisée par la longueur du chemin). Cette heuristique à pour effet de maximiser la cognation moyenne des mots appartenant au chemin, tout en favorisant les chemins les plus courts (ce choix consiste à minimiser l'entropie moyenne, c'est-à-dire les irrégularités du chemin, étant donné un ensemble de contraintes représentées par les points).

Le problème de cette technique est sa complexité élevée : la comparaison des mots est en $O(n^2)$, et sur des textes de taille importante le temps et la mémoire nécessaires deviennent rapidement rédhibitoires.

Lorsque l'espace mémoire est insuffisant, Church propose un principe itératif, basé sur une modification de la granularité : au lieu de calculer la matrice au niveau des mots,

l'algorithme commence à comparer des blocs de mots, à l'intérieur d'une bande très large autour de la diagonale (notons que seule la complexité en espace est diminuée : du point de vue de la complexité en temps, le même nombre de comparaisons est effectué). Après filtrage des points, la bande de calcul est resserrée autour de la zone de plus forte densité, et la granularité, i. e. la taille des blocs, diminue. Le rapport entre granularité et largeur de la bande restant constant, l'espace mémoire requis reste stable. Les itérations se succèdent jusqu'à ce que la granularité ne puisse plus être raffinée (taille des blocs = 1 mot) ou que la largeur de la bande ne puisse plus être diminuée (sous peine de tronquer les irrégularités du chemin). Le chemin optimal est alors calculé sur la base des points obtenus.

A partir du même type de matrice *dotplot*, Pascale Fung (1994) a cherché à tirer parti des distributions des formes simples. Le principe de la méthode est le suivant : on découpe des deux textes T et T' en K segments d'égales dimensions (indépendamment de la segmentation en phrases). A chaque forme est ainsi attribué un « K-vecteur » représentant sa distribution dans le texte : $Vf=(d_i)_{i=1..K}$. avec $d_i=1$ si u appartient au segment numéro i, 0 sinon. La comparaison de ces vecteurs entre les formes des deux textes permet d'effectuer des appariements entre elles. Cette comparaison peut être effectuée à partir des mesures statistiques précédemment étudiées (information mutuelle ou t-score) : tous les mots dont le score d'association est supérieur à un certain seuil génèrent un point. La matrice *dotplot* ainsi obtenue est ensuite traitée suivant les mêmes techniques que dans *Char-align*. Du choix de K dépend la réussite de la méthode : en effet, si K est trop grand, les sections découpées seront étroites, et les cooccurrences des formes correspondantes risquent d'être perdues, avec augmentation du silence. A l'inverse, avec un K trop petit, on risque d'obtenir de nombreuses cooccurrences dues au hasard, à l'intérieur de sections trop larges : cette perte de discrimination se traduit alors par une augmentation de bruit. Suite à une étude empirique, Fung propose de choisir K comme la racine carrée du nombre (moyen) de phrases des textes.

Dans la mesure où cette méthode fournit un ensemble de correspondances lexicales brute, l'auteur propose de l'utiliser pour amorcer un algorithme d'alignement au niveau lexical tel que celui décrit par Dagan *et al*. (1993).

Notons enfin l'architecture originale présentée par Débili & Sammouda (1992 : 522) : toutes les phrases sont comparées à l'intérieur d'une fenêtre mouvante, qui avance à mesure que les appariements sont considérés comme acquis. Le score d'association entre chaque phrase est obtenu par le produit de trois facteurs :

– $\alpha = 1 - \frac{|i-j|}{(i+j)}$ représente la proximité des phrases P_i et P'_j par rapport à la diagonale.

– $\beta = 1 - \frac{|l-l'|}{(l+l')}$ représente la proximité de l et l', les longueurs respectives des phrases P_i et P'_j

– la similarité lexicale de surface entre P_i et P'_j, est calculée en fonction d'une matrice *Matmot*, qui contient les mesures de similarité entre chaque couple de mots de P_i et P'_j en fonction des sous-chaînes maximales (en utilisant l'algorithme de Bellman déjà évoqué). Pour les unités u_p et u'_q, on calcule donc :

$$N_{pq} = 1 - \frac{|l_p - l'_q|}{(l_p + l'_q)} \sum_t n(t) \cdot t^2 \qquad (19)$$

La matrice *Matmot* ainsi obtenue est convertie en un scalaire calculé comme la somme des maxima verticaux et horizontaux divisée par 2.

Ce type d'architecture matricielle est assez souple pour intégrer tous les types d'indice, voire de les combiner simultanément sous la forme d'un produit de différents facteurs.

II.2.5.2 Algorithme itératif

La plupart des architectures décrites dans la littérature fonctionnent selon un principe itératif. La nature même de la tâche de l'alignement s'y prête parfaitement : lorsqu'un algorithme a permis d'obtenir un certain nombre de points d'ancrage, il peut être intéressant d'utiliser les informations dégagées par ce préalignement pour aligner de

nouveau, avec plus de finesse. Chaque étape apporte un lot d'informations nouvelles qui est ensuite exploité par l'étape suivante, et l'on répète l'opération jusqu'à avoir atteint la stabilité.

Dans ce type de processus, on observe généralement un « va-et-vient » entre deux types d'informations complémentaires : par exemple un alignement grossier peut donner lieu à l'extraction de correspondances lexicales, qui sont ensuite utilisées pour fournir un alignement un peu plus fin, etc. Une architecture itérative typique pourra donc se représenter de la sorte :

0. *Initialisation* : calcul d'un préalignement grossier (par exemple au niveau des paragraphes)
1. calcul des correspondances lexicales en fonction de l'alignement obtenu.
2. calcul d'un alignement en fonction des dernières correspondances lexicales.
3. retour en 1 si le nouvel alignement est plus fin que le précédent, sinon
4. *Terminaison*

La méthode développée par Kay & Röscheisen (1993) suit ce fonctionnement circulaire : l'alignement des phrases permet d'apparier des mots en fonction de leurs distributions, et l'on en déduit une segmentation en portions alignées. Ce nouvel alignement plus fin permet un nouveau calcul sur les distributions et donc de nouveaux appariements lexicaux.

Trois tables sont mises à jour au cours de l'algorithme :

- la table des phrases alignables (en anglais abrégé AST).
- la table de l'alignement des mots (WAT).
- la table de l'alignement des phrases (SAT).

Les étapes de l'algorithme peuvent être schématisées ainsi :

0. *Initialisation* : création de la table AST

1. AST → création de WAT : appariement des mots en fonction de AST.
2. WAT → création de SAT : les appariements lexicaux aboutissent à des appariements de phrase.
3. SAT → m. a. j. de AST : on met à jour de la table des phrases alignables en fonction des nouveaux points d'ancrage.
4. retour en 1 si AST a changé, sinon :
5. *Terminaison*

Au départ, la table AST est initialisée de telle sorte que sont considérés comme alignables tous les couples situés autour de la diagonale (l'écart autorisé par rapport à la diagonale est variable : il est minimal aux deux extrémités, et maximal au milieu de la diagonale). La déviation maximum autorisée est fonction de la racine carrée de la longueur des textes.

D'abord, l'étape (1) débouche sur une mise à jour de la table WAT, sur la base du score d'association précédemment décrit (cf. *Dice coefficient*, p 263).

Puis cette mise à jour permet de passer à l'étape (2) et d'enrichir l'ensemble des phrases alignées de SAT. Les appariements de mots sont utilisés pour apparier les phrases suivant les critères précédemment exposés (cf. p. 264). Ce système reposant essentiellement sur les associations lexicales, les unités lexicales sont préalablement lemmatisées afin de gommer les variations morphologiques et de consolider les phénomènes de cooccurrence.

Enfin dans la table SAT ainsi obtenue, chaque appariement de phrase est comptabilisé. Les couples qui ont été appariés un certain nombre de fois sont retenus comme points d'ancrage pour la mise à jour de la table AST, en suivant les mêmes principes qu'au départ : les points d'ancrage établissent de nouvelles diagonales autour desquelles on calcule le champ des phrases alignables. A l'issue de cette dernière phase le processus itératif peut recommencer (les tables WAT et SAT sont réinitialisées à chaque étape). L'algorithme se termine lorsque la table AST devient stable d'une itération à l'autre.

II.2.5.3 Programmation dynamique

Le nombre de chemins possibles à travers l'espace bidimensionnel décrit par les deux textes est une fonction exponentielle de la taille des textes : pour trouver le *meilleur* chemin (i.e. celui qui obtient le meilleur score, la plus petite distance), il est donc exclu de tester tous les chemins possibles. L'algorithme de Viterbi, dit de programmation dynamique (R. Bellmann, 1957 , E. Charniak, 1993) apporte une solution élégante à ce problème d'optimisation : plutôt que de tester *tous* les chemins possibles, il est plus intéressant de ne calculer, pour chaque point de l'espace, que le *meilleur* sous-chemin qui relie ce point à l'origine. Or le meilleur sous-chemin parvenant à un point est fonction des meilleurs sous-chemins qui arrivent aux points situés juste avant : il peut donc être calculé de proche en proche, de façon récursive. Le chemin optimal – l'alignement final – ne sera autre que le meilleur sous-chemin reliant l'origine à la fin de l'espace. Pour des textes de taille moyenne n il y a n^2 sous-chemins à calculer, et la complexité globale de l'algorithme devient donc polynomiale.

Cette technique est employée dans toutes les méthodes basées sur la recherche de la meilleure suite de transitions, comme celles de Gale & Church (1991) ou de Brown *et al*. (1991).

Concrètement, pour un point de coordonnées (i,j), la distance du meilleur chemin menant à (i,j) est calculée en fonction des sous-chemins immédiats (i. e. qui rejoignent (i,j) moyennant une transition supplémentaire).

Dans le cas suivant, si l'on ne tient compte que des transitions (1:1), (1:0), (0:1), (2:1), (1:2), (2:2), on a :

$$D(i,j) \begin{cases} (1:1) \text{ -> } D(i\text{-}1;j\text{-}1) + d(P_{i-1};P'_{j-1}) \\ (0:1) \text{ -> } D(i;j\text{-}1) + d(\ ;P'_{j-1}) \\ (1:0) \text{ -> } D(i\text{-}1;j) + d(P_{i-1};\ ) \\ (2:1) \text{ -> } D(i\text{-}2;j\text{-}1) + d(P_{i-2}P_{i-1};P'_{j-1}) \\ (1:2) \text{ -> } D(i\text{-}1;j\text{-}2) + d(P_{i-1};P'_{j-2}P'_{j-1}) \\ (2:2) \text{ -> } D(i\text{-}2;j\text{-}2) + d(P_{i-2}P_{i-1};P'_{j-2}P'_{j-1}) \end{cases} \quad (20)$$

Où $D(i,j)$ est la distance du sous-chemin optimal menant à (i,j)

et $d(P_1P_2..P_n;P'_1P'_2..P'_m)$ est l'incrément de distance liée au binôme $(P_1P_2..P_n;P'_1P'_2..P'_m)$.

L'algorithme peut être appliqué à l'identique sur une plus grande étendue de transitions, du type (1:3), (3:2), (2:3), (3:2), etc. ce qui élargit la couverture des cas possibles, mais augmente aussi les branchements récursifs.

En ce qui concerne le calcul de la distance le long d'un chemin, Davis *et al*. (1995) ont fourni un cadre théorique général pouvant s'appliquer à indifféremment à tous les types d'indices (rapport des longueurs, cognation, contenu lexical, points d'ancrage, etc.). Le raisonnement est analogue à celui de Gale & Church (1991), mais plus général.

On calcule la probabilité d'un alignement comme le produit des probabilités de chaque binôme, celles-ci étant dépendantes d'un certain nombre d'indices $\delta_1, \delta_2,\ldots, \delta_k$:

$$P(A/T1,T2) = \prod_{B_i} p(B_i / \delta_1, \delta_2,..., \delta_k)$$

l'événement B_i signifiant « $(P_{mi...ni}, P'_{pi...qi})$ sont alignés ».

Avec le théorème de Bayes, on peut donc estimer la probabilité des indices sachant B_i :

$$P(B_i / \delta_1, \delta_2,..., \delta_k) = \frac{p(\delta_1, \delta_2,..., \delta_k / B_i)\, p(B_i)}{p(\delta_1, \delta_2,..., \delta_k)}$$

Si l'on présume l'indépendance statistique des indices, on a :

$$P(B_i / \delta_1, \delta_2,..., \delta_k) = \frac{\prod_q p(\delta_q / B_i)}{\prod_q p(\delta_q)}\, p(B_i)$$

Enfin, on distingue les distributions de δ_q suivant qu'elles sont observées entre les binômes équivalents (B_i), ou pour les binômes non équivalents ($\neg B_i$):

$$P(B_i / \delta_1, \delta_2,..., \delta_k) = \frac{\prod_q p(\delta_q / B_i)}{\prod_q p(\delta_q / B_i)\, p(B_i) + p(\delta_q / \neg B_i)\, p(\neg B_i)}\, p(B_i) \qquad (21)$$

On notera:

$p(\delta_q/B_i) = p_a(\delta q)$ la probabilité observée de l'indice δ_q au sein des binômes équivalents ;

$p(\delta_q/\neg B_i) = p_{\neg a}(\delta_q)$ la probabilité observée de l'indice δ_q au sein des binômes non équivalents .

Pour l'estimation de ces distributions, on peut se baser sur des données empiriques.

La véritable signification de $p(B_i)$ est : « probabilité que les phrases $P_{mi...ni}$ soient alignées avec les phrases $P'_{pi...qi}$ ». Dans l'absolu, cette probabilité est difficile à estimer, car elle dépend du nombre de regroupements possibles au sein de l'espace de calcul. Pour simplifier, comme chez Gale & Church, on assimilera la probabilité de l'événement B_i à la probabilité de la transition $T_i = (m_i..n_i:p_i...q_i)$. L'équation (21) permet donc d'évaluer la probabilité de l'hypothèse d'alignement en fonction de données observables : les indices δ_q et la transition requise.

On construit la mesure de distance en prenant l'opposé du logarithme de cette probabilité. La distance totale liée à un alignement A est donc :

$$\begin{aligned} P(\boldsymbol{A}/T,T') &= \sum_i -\log P(B_i/\delta_1,\delta_2,...,\delta_k) \\ &\approx \sum_i \left[\sum_q -\log p_a(\delta_q) - \sum_q -\log\left(p_a(\delta_q)p(t_i) + p_{\neg a}(\delta_q)p(\neg t_i)\right) - \log p(t_i) \right] \end{aligned} \quad (22)$$

Cette mesure présente l'intérêt de faire intervenir des informations positives et négatives vis-à-vis de la présomption d'alignement : les statistiques des indices δ sont calculées en fonction des deux hypothèses de l'équivalence et de la non-équivalence. Plus un indice est discriminant, plus les deux distributions p_a et $p_{\neg a}$ divergent, et plus la mesure de distance est efficace pour faire ressortir le chemin correct.

Ces équations définissent un cadre général pour combiner, de manière cohérente, des informations diverses. Par exemple Davis *et al*. utilisent 4 indices :

– δ_l : rapport des longueurs (comme chez Gale & Church, 1991)

- δ_2 : nombre de 4-grammes dans la comparaison des formes d'un binôme, divisé par le produit des longueurs des deux segments.
- δ_3 : longueur de la sous-chaîne maximale commune au deux segments[138], divisée par la somme des longueurs des segments.
- δ_4 : nombre de chaînes numériques communes, divisé par le nombre total de chaînes numériques dans les deux segments.

Testé sur des documents problématiques respectant mal les conditions de parallélisme (traductions approximatives de textes anglais / espagnol issu de la *Pan American Health Organisation*), cette combinaison d'indices parvient à améliorer nettement les résultats de la méthode de Gale & Church (1991) :

	Précision	***Rappel***
GC	33,2 %	23,1 %
Combinaison	62,2 %	49,1 %

Chez Simard *et al*. (1994), la combinaison des indices ne se fait pas de manière simultanée. L'indice basé sur la cognation est conçu comme complémentaire de la mesure basée sur les longueurs, et s'applique dans un deuxième temps. Les auteurs remarquent que si l'on compare les scores des meilleurs chemins (pour la mesure basée sur les longueurs), on observe deux cas de figure : dans le cas où le meilleur alignement proposé est correct, la distance totale du chemin est en moyenne 100 fois inférieure à celle du second meilleur alignement. Si au contraire l'algorithme n'a pas fourni d'alignement correct, ce rapport tombe en moyenne à 2. Les auteurs proposent de se servir de ce « rapport de distinction » pour détecter les cas litigieux, auxquels on applique, dans une deuxième étape, la distance basée sur la cognation.

Simard *et al*. soulignent par ailleurs que le « rapport de distinction » fournit une méthode de filtrage efficace permettant de détecter les erreurs, d'éliminer les cas litigieux ou d'appliquer d'autres méthodes plus adaptées lorsque celle-ci a échoué.

[138] Par conséquent l'ordre des mots intervient, à la différence de l'indice précédent.

Enfin, il faut signaler l'algorithme implémenté par Chen (1993), qui présente des caractéristiques intéressantes. C'est une version incrémentale de l'algorithme EM (Dempster *et al*. 1977), de type itératif.

En principe, l'algorithme EM s'articule autour des deux phases suivantes, réitérées jusqu'à stabilité des paramètres :

- phase E (pour *Expectation*) : estimation du meilleur chemin en fonction des paramètres du modèle (p_b, p_e, p_f, $p(l)$).
- phase M (pour *Maximisation*) : réestimation des paramètres afin que soit maximisée la probabilité de ce meilleur chemin.

Dans la version incrémentale, les paramètres sont réestimés à mesure que le meilleur chemin est calculé. L'espace de calcul est borné autour de la diagonale, et un algorithme de Viterbi (du type de celui développé précédemment, avec exploration en largeur d'abord), permet de calculer de manière récursive tous les sous-chemins de même longueur. Un élagage progressif des sous-chemins les plus improbables (i.e. dont la probabilité s'éloigne le plus du meilleur sous-chemin) permet de construire, progressivement, le « préfixe » du meilleur chemin.

Ainsi, à chaque nouvelle transition ajoutée à ce « préfixe » de chemin, on met à jour les paramètres liés aux probabilités des appariements de mots de la manière suivante :

$$p_b(b) = \frac{c(b)}{\sum_{b'} c(b')}$$

Où $c(b)$ représente le nombre de fois que l'appariement b est apparu dans l'assignement d'un appariement de phrases du « préfixe » déjà construit.

A l'initialisation du chemin, les $c(b)$ sont fixés à 1 lorsque b est du type (e ;) ou (;f). Afin de permettre aux autres cas de figure d'apparaître, on initialise la probabilité $p((e,f))$ à une valeur non nulle à chaque fois que (e;) et (;f) apparaissent simultanément dans un même assignement. Enfin, pour amorcer le processus, et fixer les paramètres initiaux à des valeurs correctes, on « entraîne » l'algorithme sur un petit corpus d'apprentissage (environ une centaine de binômes) déjà aligné. $\lambda_{1:0}$ est calculé en fonction du nombre moyen de mots par phrase.

Un mécanisme de détection des insertions ou des omissions massives est ajouté (car celles-ci finissent par fausser les paramètres en même temps qu'elles détournent le préfixe

du chemin correct) : en principe le meilleur sous-chemin doit être beaucoup plus probable que tous les autres sous-chemins possibles. Ceux-ci doivent ainsi être élagués très rapidement. Si l'algorithme bute sur une insertion massive, les mesures de probabilité ne discriminent plus entre les sous-chemins divergents : le nombre de sous-chemins candidats doit donc augmenter, et le champ de calcul s'élargit. Lorsque le champ de calcul atteint une certaine largeur, on peut donc supposer qu'il y a insertion ou omission dans la traduction.

Pour identifier la fin de la zone d'insertion une technique originale est employée : les mots rares sont cherchés de façon linéaire de part et d'autre du corpus. Lorsqu'un mot rare a été trouvé ainsi que sa traduction, le point correspondant est considéré comme candidat de la fin de la zone. Les quarante premières phrases suivant ce point sont examinées afin de déterminer s'il en existe un alignement probable. Si c'est le cas, l'algorithme reprend son chemin, sinon un autre candidat est recherché.

II.2.6 Résultats des méthodes décrites

Nous avons présenté un certain nombre de techniques regroupées sous deux rubriques : les indices et les algorithmes. Ces techniques ont toutes fait l'objet de comparaisons et de mises à l'épreuve, mais avant d'en donner un rapide résumé, il nous faut expliciter quelques critères d'évaluation.

II.2.6.1 Critères d'évaluation

Nous retiendrons trois principaux axes d'évaluation destinés à mettre en perspective les performances des techniques en fonction de leurs utilisations potentielles : coût, qualité, champs d'application.

– *Coût*

Il faut distinguer différents types de coûts :

- coût d'implémentation (simplicité de la mise en œuvre)
- complexité de calcul en temps et en espace (linéaire, polynomiale, etc.).

- coût de fabrication des ressources additionnelles (par exemple lorsqu'un système utilise une base de donnée dictionnairique, etc.).

– *Qualité du résultat* (*en fonction des réquisits de l'utilisation ultérieure*) :

Comme on l'a vu, la qualité d'un alignement automatiquement extrait est mesurable à l'aide de mesures quantitatives, lorsqu'on dispose d'un alignement de référence :

- précision
- rappel
- F-mesure

Ces mesures ne sont pertinentes que si l'on spécifie également la *granularité* de l'alignement évalué, qui doit être comparable à la granularité de l'alignement de référence.

Plus généralement, pour évaluer les techniques elles-mêmes, il est nécessaire de les tester sur des corpus aux caractéristiques variées, comme dans le projet ARCADE : cela permet de mettre en évidence la *robustesse* d'une méthode, c'est-à-dire sa capacité à fournir des résultats constants, sans dégradation catastrophique sur certains textes. En outre, avec des textes variés et des couples de langues différents, on peut mieux dégager le *profil* d'une méthode, c'est-à-dire les caractéristiques des corpus (et des langues) auxquels elle convient le mieux.

– *Champ d'application, contraintes d'utilisation*

Il est également important de dégager le champ d'application d'une méthode, et les contraintes qu'elle impose. On peut par exemple évaluer :

- le degré d'automatisation : il peut être total ou bien nécessiter l'intervention humaine (par exemple, lors d'une phase d'entraînement, avec un corpus aligné par l'homme).
- l'indépendance vis-à-vis des langues. Certaines méthodes exploitent des propriétés étroitement liées au couple de langues (par exemple celles qui intègrent la cognation), d'autres non.

- l'indépendance vis-à-vis du type de texte. Les indices peuvent être liés au format et au type des textes (comme les balises, les paragraphes préalignés, les contraintes de parallélisme, etc.) ou non.
- la nécessité de ressources additionnelles *ad hoc*. Certaines techniques utilisent des glossaires bilingues, des étiqueteurs morphosyntaxiques, ou d'autres ressources linguistiques liées à un couple de langues donné et / ou au domaine d'application.

La synthèse de ces différents critères permet d'effectuer une évaluation globale, dans la perspective des applications concrètes des techniques d'alignement.

II.2.6.2 Résultats

Au vu de ces critères, on peut maintenant comparer les résultats des méthodes précédemment décrites.

- *Gale & Church* (*1991*)

Dans leur implémentation, les auteurs ajoutent une première étape pour préaligner les textes au niveau des paragraphes : ensuite la méthode des longueurs est employée à l'intérieur des paragraphes préalignés. Comme le remarque Ingeborg Blank (1995), sans cette étape préliminaire, l'algorithme est inapplicable : sur une grande quantité de phrases non préalignées, l'algorithme « perd » très vite le chemin correct. La longueur des phrases n'est pas une information assez discriminante pour que le chemin optimal coïncide avec le chemin correct malgré les insertions, les omissions et tous les écarts par rapport aux conditions du parallélisme. Même avec un préalignement, on note que la précision décroît fatalement avec la longueur des paragraphes.

Gale & Church (1991 : 182) notent également une mauvais prise en compte des transitions de la catégorie des insertions et omissions (100 % d'erreurs). Les meilleurs résultats sont obtenus avec les transitions du type (1:1), qui entraînent moins d'erreurs que les transitions (1:2), (2:1), et surtout (3:1) et (1:3).

La précision obtenue sur le corpus Hansard se situe aux alentour de 95 %.

Les résultats sont meilleurs lorsque la longueur des phrases est exprimée en caractères. Avec des longueurs en nombre de mots, le taux d'erreur augmente de 4,2 % à 6,5 %.

Enfin, il faut noter que le score est un bon indicateur de la fiabilité. En conservant 80 % des alignements obtenant les meilleurs scores, le taux d'erreur tombe de 4 % à 0,7 %. Cependant, Blank (1995) tempère ce jugement, car avec un autre corpus la mesure de distance ne permet pas toujours d'éliminer les alignements incorrects. Elle reste basée sur des critères trop superficiels.

– *Simard et al.* (*1994*)

A la méthode précédente, Simard *et al*. ajoutent un nouvel indice : la cognation.

Seule, la mesure basée sur les cognats n'améliore pas les résultats (on passe de 1,8 % d'erreurs à 2,4 %) ; mais la combinaison des deux réalise une légère progression (de 1,8 % à 1,6 %) avec une augmentation du rappel (de 128 binômes manquants, on passe à 114). Cette amélioration est probablement due au fait que les deux indices ne font pas les mêmes erreurs aux mêmes endroits : lorsque l'information liée à la cognation est insuffisante, celle liée aux longueurs prend parfois le relais, et réciproquement. Ainsi, la méthode combinée est plus robuste, dans la mesure où elle résiste mieux aux cas de décalages en cascade[139].

En analysant les erreurs, les auteurs remarquent également que ce genre de méthode est incapable de détecter correctement les cas d'omission et d'insertion. Gale & Church (1991) pensent que ces irrégularités requièrent un traitement « *language specific* », basé sur l'analyse du contenu linguistique.

– *K.W. Church* (*1993*)

L'avantage de la méthode *Char-align*, d'architecture matricielle, est sa résistance au bruit. Elle est adaptée à des textes parallèles comportant de nombreuses erreurs de mise en

forme et d'impression : les auteurs donnent l'exemple d'un corpus numérisé par reconnaissance de caractères, comportant des mots et des signes de ponctuation mal reconnus, ainsi que des notes flottant dans le corps du texte sans respecter les conditions de parallélisme. Le traitement matriciel de cette méthode autorise les discontinuités, voire les croisements, pour de grandes portions textuelles : il présente donc une grande robustesse. La mesure du rappel n'a pas beaucoup de signification dans le cas présent car l'alignement se fait au niveau des caractères, et non entre segments de textes.

Les résultats n'ont pas été chiffrés mais la précision semble, au vu des graphiques fournis par l'auteur, assez élevée.

– *Kay & Röscheisen* (*1988*)

Etant basée sur le lexique, cette méthode est totalement indépendante du couple de langues, à la différence des méthodes basées sur les cognats. A l'instar de *K-vec* (Fung & Church, 1994), elle fournit un glossaire brut minimal qui peut être réutilisable ailleurs.

Les résultats obtenus sur un corpus anglais - allemand (textes scientifiques dans le domaine des sciences physiques) d'environ 300 phrases sont intéressants. Au bout de la 4ème itération, la précision est de 99,7 % pour un rappel de 96 %. Sur des textes plus longs, la précision tend vers 100 %. En deçà de 150 phrases, elle se dégrade, car l'information lexicale devient insuffisante. La complexité est majorée par $O(n \sqrt{n})$.

D'après les auteurs, la complexité peut devenir linéaire si la méthode est appliquée à la suite d'une technique de préalignement en $O(n)$. Les résultats peuvent en outre être améliorés en tenant compte de correspondances multiples et des unités polylexicales les plus courantes.

Dans un travail d'évaluation, Blank (1995) remarque que certaines modifications de la méthode permettent d'améliorer simultanément la précision et l'efficacité de l'implémentation :

[139] Par exemple, dans le cas d'une liste, où tous les items seraient *grosso modo* de même longueur, un décalage de l'alignement au début ne serait pas rattrapé avant la fin de la liste, si l'on utilise seulement les longueurs.

- la prise en compte de points d'ancrage additionnels augmente la précision. Ces points d'ancrage peuvent être fournis par les marques de paragraphes, les transfuges et les cognats (s'ils apparaissent avec la même fréquence dans les deux textes, et fournissent des points dans les portions alignables).

- l'exclusion des mots outils (« *function words* ») et des formes très fréquentes accroît la vitesse du traitement et améliore la précision.

- dans la génération de la table AST (table des phrases alignables), la précision est meilleure lorsque la déviation autorisée par rapport à la diagonale est réduite par rapport au paramétrage initial de Kay & Röscheisen.

Avec ces modifications, la précision s'échelonne entre 75 % et 89,1 % sur le corpus étudié (articles scientifiques anglais - allemands, documentation de logiciel français - anglais et documentation concernant les licences commerciales « *patent documentation* » français - anglais - allemand).

La méthode ainsi modifiée a été comparée à la méthode GC, sur un échantillon du corpus trilingue de « *patent documentation* » (350 000 mots) les résultats sont les suivants :

	Précision	***Rappel***
GC	95 %	100 %
Kay & Röscheisen	89,59 %	88,69 %

Le rappel, dans la méthode de Kay & Röscheisen, semble intrinsèquement limité, parce que l'algorithme ne retient que les couples de phrases possédant un bon score (à la différence d'un algorithme du type Viterbi, où le chemin ne connaît pas de rupture).

- *Chen* (*1993*)

La méthode inspirée du modèle de traduction statistique obtient de très bons résultats sur le corpus Hansard (3 millions de phrases en anglais et français) : Chen compte seulement 0,4 % d'appariements erronés, sans supprimer de zone textuelle (comme les

10 % éliminés par Brown *et al*., 1991). La méthode est donc précise et robuste. En outre, il semblerait que la plupart des erreurs soient dues à une mauvaise segmentation des phrases.

Du point de vue de l'efficacité, Chen note un rendement de 2 000 à 5 000 phrases par heure, i.e. 10 fois plus lent que l'algorithme GC avec le même matériel. Mais le résultat n'est pas le même, car l'alignement est plus fin, puisqu'il se situe au niveau lexical. Globalement, le coût des calculs est proportionnel à la taille du lexique des textes traités.

II.3 Expérimentation

Ce rapide tour d'horizon de l'« état de l'art », loin d'être exhaustif, nous a permis de présenter les principales techniques dévolues à la tâche de l'alignement automatique.

Nous proposons maintenant d'implémenter certaines de ces techniques afin d'en donner une évaluation approfondie, en prêtant une attention toute particulière aux propriétés formelles des corpus mis en jeu, puisque ce sont ces propriétés qui conditionnent en grande partie les résultats.

Ce travail d'évaluation poursuit les objectifs suivants :

- comparer les différentes méthodes sur un même corpus ;
- comparer les différents types de corpus en fonction des résultats obtenus ;
- corréler les résultats avec certaines caractéristiques textuelles.

Nous chercherons ainsi a déterminer les traits textuels, qualitatifs ou quantitatifs, susceptibles d'indiquer si telle méthode peut ou non s'appliquer, et avec quelle chance de succès.

En ce qui concerne le développement des techniques, nous étudierons certaines possibilités d'améliorations, par la détermination des paramétrages les plus adaptés. Enfin, nous tenterons de déterminer comment les différentes méthodes peuvent se combiner, afin de minimiser les calculs et d'optimiser la qualité des résultats.

II.3.1 Corpus d'étude

Nous devons rendre hommage au projet ARCADE (Véronis, 1997), qui nous a permis d'utiliser un corpus important, riche, et très utile comme base de comparaison. En effet, ce corpus a servi d'étalon pour la « compétition » amicale de plusieurs systèmes d'alignement (dont un développé par nous). Un grand nombre de résultats chiffrés sont disponibles pour toute la gamme des systèmes qui y ont concouru. Nos propres résultats peuvent ainsi être situés par rapport à ceux obtenus par d'autres systèmes, et fournir un élément de comparaison intéressant, au-delà de ce que nos propres expérimentations ont pu livrer comme observations.

Le corpus BAF (pour *Bi-texte anglais - français*) a été initialement constitué par des chercheurs du CITI, de Laval (Canada), puis repris par le RALI de l'Université de Montréal, dans le cadre d'un projet initié et financé par l'AUPELF-UREF. La finalité de ce corpus est de constituer un étalon de mesure, pour la comparaison des techniques basées sur les bi-textes, « un banc d'essai commun, sous la forme de corpus de référence pour l'alignement »[140] (Simard, 1998 : 490).

Le corpus BAF est essentiellement composé de textes institutionnels, mais contient aussi des textes scientifiques, techniques et littéraires (cf. tableau 83 de l'annexe). Il offre par conséquent une certaine variété typologique.

Le tableau 9 en donne la composition, ainsi que les tailles de ses diverses parties (les statistiques de mots et de caractères ont été obtenues sous Word 97[141], le compte des phrases est lié à la segmentation de référence du corpus).

[140] "a common test-bed, in the form of reference alignment corpora".

[141] Les chiffres et les signes de ponctuation sont comptés comme des mots. Les espaces, les tabulations ou les marques de paragraphes ne sont pas comptés dans les caractères.

Genre	*Nom*	*Français* : *taille*			*Anglais* : *taille*		
		phrases	***mots***	***car.***	***phrases***	***mots***	***car.***
Institutionnel	*Cour*	1 453	33 176	171 165	1 405	31 096	154 906
	Hans	3 075	59 102	311 600	3 162	55 729	270 031
	Ilo	7 606	153 061	836 535	7 246	145 346	768 926
	Onu	2 591	54 869	300 162	2 647	48 407	267 869
Scientifique	*TAO1*	372	7 637	41 955	370	6 572	35 544
	TAO2	314	7 158	39 172	325	6 985	35 841
	TAO3	181	3 298	18 339	202	3 243	17 231
	CITI1	645	12 574	72 443	653	12 480	68 304
	CITI2	1 622	23 068	124 398	1568	20 664	112 194
Technique	*Xerox*	3 319	46 828	257 608	2 554	39 328	199 779
Littéraire	*Verne*	3 871	53 645	282 629	3 766	40 173	197 997
Total	11 textes	25 049	454 416	2 456 006	23 898	410 023	2 128 62

tableau 9 : *composition du corpus BAF*

Le détail des principes de la segmentation en phrases a déjà été esquissé (cf. p. 237). Pour la segmentation des phrases en mots, nous avons utilisé une définition opératoire du mot : toute chaîne de caractères (lettre ou chiffre) comprise entre deux séparateurs (caractère non-alphanumérique[142]). Cette règle admet quelques exceptions : le point n'est pas un séparateur s'il est immédiatement précédé et suivi d'une lettre majuscule (pour la reconnaissance des sigles et des acronymes), le point ou la virgule ne sont pas des séparateurs s'ils sont immédiatement précédés et suivis d'un chiffre (reconnaissance des nombres). Dans le calcul du nombre de mots d'une phrase, chaque séparateur compte pour un mot.

Pour l'alignement manuel, un certain nombre de critères ont été explicités, afin de guider les « aligneurs » humains dans leur tâche (M. Simard, 1998 : 491) :

– critère de traduction : deux segments sont considérés comme équivalents s'ils véhiculent le même contenu conceptuel.

[142] La liste des séparateurs est : espace, retour chariot, tabulation, les signes de ponctuation (point, virgule, point virgule, point d'exclamation, point d'interrogation, guillemet), l'apostrophe, le tiret, les signes de parenthésage (parenthèse, crochets, accolades), les autres caractères non alphanumériques (pourcentage, symboles monétaires, etc.).

- les segments supprimés ou insérés sont alignés avec des segments vides, sauf dans les cas où le segment manquant est pris entre deux phrases alignées avec une seule : si *A* est traduite par A'_1 et A'_2, et si *B*' est insérée entre A'_1 et A'_2, alors *A* est aligné avec $A'_1 B' A'_2$.

- lorsque l'ordre de deux phrases est inversé (par exemple A_1A_2 est aligné avec $A'_2A'_1$) on utilise une transition (2:2). Mais si l'inversion est plus complexe (du type $A_1A_2A_3A_4$; $A'_4A'_1A'_2A'_3$) on effectue une double omission (A_4 et A'_4 sont alignés avec le segment vide).

- l'alignement doit être fait avec une granularité minimale, mais sans aller en deçà du niveau de la phrase.

Ces critères ont été établis dans un souci d'efficacité plus que de cohérence absolue. En cas de doute les « aligneurs » ont la consigne de faire le choix le plus « utile »[143], dans la perspective de l'utilisation ultérieure du bi-texte : on privilégie ainsi les configurations les plus simples, en éliminant les cas litigieux dans des alignements vides.

Deux « aligneurs » humains différents ont traité le corpus séparément puis ont confronté leurs résultats, afin de minimiser les erreurs et l'aspect subjectif de l'interprétation des consignes. Simard note que la plupart des discordances entre les deux alignements étaient dues à des problèmes de segmentation plus qu'à des problèmes de contenu conceptuel.

Nous avons également rassemblé un corpus parallèle, constitué des versions françaises et anglaises de rapports du parlement européen, disponibles en ligne sur le serveur d'Europarl[144]. Nous avons initialement travaillé sur ce corpus, avant d'avoir accès au corpus BAF. Ce dernier existant en version alignée et représentant un étalon plus fiable, nous avons peu utilisé le corpus Europarl dans nos expérimentations.

---

[143] *"to do the most "useful" thing"* (Simard, 1998 : 492)
[144] à l'adresse : http://www.europarl.eu.int

II.3.2 Méthodologie expérimentale

Toute démarche expérimentale se fonde sur un principe fondamental : la reproductibilité des faits observés lorsqu'un certain nombre de conditions sont satisfaites, *mutatis mutandis*. La définition de ces conditions aboutit à la détermination rigoureuse du protocole expérimental. Pour garantir la reproductibilité, ce protocole doit définir la totalité des facteurs intervenant de manière causale dans l'observation des faits : une fois déterminés, ces facteurs peuvent ensuite être neutralisés, en étant fixés, ou bien être exploités comme variable pour être mis en relation (de causalité ou de concomitance) avec les variations observées au niveau des phénomènes.

La méthode hypothético-déductive vise à faire entrer les conditions définies par le protocole expérimental dans le cadre d'un modèle permettant d'anticiper par déduction les observations elles-mêmes : la vérification expérimentale des prédictions déduites du modèle intervient alors comme une validation de celui-ci (sous condition, mais scientifiquement établie) comme interprétation de la réalité empirique.

Dans la démarche empirique fondée sur l'exploitation de corpus linguistique nous nous confrontons à des difficultés inhérentes à la nature des faits :

- les facteurs sont difficiles à abstraire du corpus. En effet, bien qu'on puisse aisément garantir la reproductibilité des résultats sur le même corpus, il est délicat de les généraliser et de les étendre à d'autres corpus. Pour ce faire, il faudrait arriver à extraire tous les caractères (généralisables à d'autres textes) dont les observations dépendent. Nous verrons par la suite que la détermination de ces facteurs devient épineuse dès lors que l'on travaille avec un matériau de nature sémantique : pour garantir la validité du protocole, il faudrait s'appuyer sur des informations intersubjectives issues d'un échantillon représentatif de locuteurs.

- un corollaire de cette difficulté est l'absence de modèle prédictif susceptible de valider ou d'infirmer certaines hypothèses. Nous verrons, à l'occasion de la modélisation des correspondances lexicales, que les modèles mathématiques

deviennent rapidement d'une complexité rédhibitoire dès que l'on cherche à intégrer les facteurs syntactico-sémantiques.

Autant que possible, nous chercherons à pallier ces difficultés au cours de nos propres expériences. Parallèlement à notre effort d'optimisation des méthodes employées, nous tendrons ainsi à en démontrer l'assise scientifique.

Il nous faut reconnaître, cependant, qu'un grand nombre de résultats demeurent peu assurés, tant pas le manque de déterminations empiriques que par l'absence d'interprétation des phénomènes observés : nous le signalerons le cas échéant.

II.3.3 Critères heuristiques

Le nombre d'approches différentes et la quantité importante de paramètres à faire varier, ajoutés à l'application des méthodes sur des corpus souvent lourds, condamne d'emblée toute tentative d'exhaustivité. Il nous faut, dès le départ, faire un certain nombre de choix guidés par des critères de nature stratégique et heuristique :

- nous avons cherché à rendre compte des méthodes les plus répandues, et qui servent bien souvent de *tertium comparationis* dans les travaux d'évaluation ;

- nous avons tenu à mettre en œuvre tous les indices précédemment décrits : transfuges, cognats, rapport des longueurs, lexique. Notre pourrons ainsi vérifier l'intuition que le système fournissant les meilleurs résultats (d'un point de vue qualitatif), est vraisemblablement celui qui intègre toutes les informations pertinentes, combinées de manière appropriée.

- d'un point de vue heuristique, nous avons été guidé par un principe simple permettant de minimiser les calculs tout en garantissant de bons résultats, suivant l'idée que les informations les plus fiables doivent être utilisées d'abord : c'est le principe de « *principe de précision d'abord* » selon lequel les méthodes doivent exploiter d'abord les hypothèses les plus probables. Ainsi l'alignement final doit être obtenu comme le terme d'une suite de préalignements successifs, partant d'un préalignement à la précision maximale pour aboutir à un alignement au rappel

maximal. Chaque étape doit voir le rappel augmenter sans dégradation (si possible) de la précision.

Nous verrons comment ce principe permet, par une réduction drastique de l'espace de recherche dès les premières étapes, de faire fonctionner les techniques les plus coûteuses (programmation dynamique) en un temps linéaire par rapport à la taille des textes.

II.3.4 Traitements préliminaires

Avant la mise en œuvre des algorithmes d'alignement, les textes sont convertis sous la forme de bases de données. Chaque texte donne lieu à la construction de deux tables : la table des formes et la table des segments. L'extraction des tables se fait segment après segment[145], au cours de la lecture du texte. En ce qui concerne le corpus BAF, la segmentation est déjà réalisée, et il n'y a pas d'analyse à faire pour le découpage. Pour le corpus Europarl, les règles de segmentation sont celles décrites au chapitre II.2.1.1.

Au cours de l'étape de segmentation, les segments sont numérotés dans leur ordre d'apparition. Le tableau 10 résume les principales caractéristiques de chaque table :

Les tables sont indexées au moyen d'arbres binaires de recherche permettant des recherches en O(log(*t*)), si *t* est la taille de la table. Ces deux tables sont conçues pour établir des liens réciproques entre leurs éléments, et certaines redondances, telles que les champs *Segment* et *Chaînes*, sont motivées par la recherche d'efficacité dans les traitements.

L'extraction de ces tables est en O(*n* log(*n*))[146].

[145] En l'occurrence des phrases : nous employons le terme de segment dans un sens générique, un segment pouvant désigner une phrase ou un groupe de phrases.
[146] Quand nous indiquons O(*n*) sans autre précision, le *n* se réfère à la taille des textes.

Nom	Descriptif	Liste des champs
UNITES	Chaque enregistrement regroupe les informations relatives à une unité (mot) du texte. Indexée par ordre alphabétique des mots.	- *Réf* : clé primaire - *Unité* : chaîne de caractère correspondant à une forme simple. - *Occ* : nombre d'occurrences de la forme dans le texte (fréquence). - *VecOcc* : Vecteur d'occurrence : liste des numéros de segment dans lesquels la forme apparaît.
SEGMENTS	Chaque enregistrement regroupe des informations relatives aux segments du texte. Indexée suivant les numéros de segment.	- *Réf* : clé primaire - *Segment* : liste des références des formes composant le segment. - *Chaînes* : liste des chaînes de caractères des formes composant le segment. - *Lf* : longueur du segment en nombre de formes. - *Lc* : longueur du segment en nombre de caractères.

tableau 10 : *structure des bases de données textuelles*

Pour l'alignement des textes T et T', on notera les tables *U* (pour *Unités*), *S* (pour *Segment*), *U'* et *S'*.

II.3.5 Exploitation des transfuges

On appelle transfuge toute chaîne de caractères (issue de la segmentation en mots) apparaissant telle quelle dans les deux textes T et T' (quel que soit son emplacement et sa fréquence). Ainsi conçus les transfuges sont des indices superficiels : on ne tient pas compte de leur contenu sémantique, et tous les homographes sont identifiés comme des occurrences différentes des mêmes transfuges.

II.3.5.1 Description de l'algorithme

L'algorithme que nous avons mis en œuvre à cette première étape correspond à une implémentation rigoureuse du principe précédemment énoncé. Il est itératif dans son déroulement : chaque itération intègre les données fournies par l'itération précédente pour aller un peu plus loin dans l'alignement, et l'algorithme se termine lorsqu'il y a stabilisation du rappel.

Il s'appuie sur la prise en compte prioritaire des indices supposés les plus fiables : les chaînes numériques d'abord, puis les transfuges avec une majuscule (souvent des noms propres), puis les transfuges simples. En tant que préalignement, le résultat de cet algorithme doit être un ensemble de points au sens précédemment défini p. 242.

Nous noterons $A = \{(x_1,y_1),(x_2,y_2),\ldots,(x_p,y_p)\}$[147] un tel ensemble. Nous noterons $S_1 S_2 \ldots S_{p+1}$ et $S'_1 S'_2 \ldots S'_{p+1}$ les *p+1 sections de textes* (suites de segments) découlant de *A*.

$S_1 = [P_1 P_2 .. P_{x_1-1}]$, $S_2 = [P_{x_1} P_{x_1+1} \ldots P_{x_2-1}]$, … $S_{p+1} = [P_{x_p} P_{x_p+1} \ldots P_n]$

$S'_1 = [P'_1 P'_2 .. P'y_{1-1}]$, $S_2 = [P'_{y_1} P'_{y_1+1} \ldots P'_{y_2-1}]$, … $S_{p+1} = [P'_{y_p} P'_{y_p+1} \ldots P'_m]$

– *Algorithme* :

A représente l'ensemble des points de l'alignement en construction, A_{cand} représente un ensemble de points candidats créé à chaque nouvelle itération. Les commentaires sont en italique.

0. Initialisation : $A \leftarrow \varnothing$

1. Sélection des transfuges :

Initialisation de l'ensemble des points candidats : $A_{\text{cand}} \leftarrow \varnothing$
Pour $u_i \in U$:

[147] Notons que par convention, les indices sont attribués en fonction de l'ordre des points. Plus précisément, on impose : $i \geq j \Rightarrow x_i \geq x_j$, $y_i \geq y_j$ la relation d'ordre portant sur l'abscisse d'abord, et l'ordonnée ensuite. Dans la construction de l'alignement candidat, on imposera seulement la condition : $i \geq j \Rightarrow x_i \geq x_j$, car au cours de l'algorithme il se peut qu'il y ait des croisements, ensuite éliminés par filtrage des points.

Si u_i est une chaîne alphanumérique (cf. *note 132*) et si $u_i \in U$' alors :
Pour chaque section (S_j, S'_j) de A :
Si le nombre d'occurrences de u_i dans S_j est égal au nombre d'occurrences de u_i dans S'_j alors :
On construit k points $a_{ij1,aij2}..a_{ijk}$ à partir des numéros de segment des occurrences de u_i dans S_j et dans S'_j.
Si $a_{ij1,}a_{ij2}..a_{ijk}$ sont situés au voisinage de la diagonale alors : (1)
On les ajoute à l'ensemble des points candidats : $A_{cand} \leftarrow A_{cand} + \{a_{ij1,}a_{ij2,}..,a_{ijk}\}$.

2. filtrage des points de A_{cand}

On élimine tous les points qui rentrent en conflit : (3)
Si $a = (x,y)$ et $a' = (x',y)$ avec $x \neq x'$, alors : $A_{cand} \leftarrow A_{cand} \setminus \{a,a'\}$
Si $a = (x,y)$ et $a' = (x,y')$ avec $y \neq y'$, alors : $A_{cand} \leftarrow A_{cand} \setminus \{a,a'\}$
On ne conserve que les points donnés par au moins deux transfuges :
Si a n'a été ajouté dans A_{cand} que pour un seul transfuge alors : $A_{cand} \leftarrow A_{cand} \setminus \{a\}$
Pour tous les points a_i de A_{cand}, ($i = 2..q$) :
Si a_i n'induit pas une déviation trop forte avec le dernier point filtré a_f (2)
et si a_i ne croise pas le dernier point filtré a_f (*i.e. si* $a_i=(x_i,y_i)$ *et* $a_f=(x_f,y_f)$ *et* $x_i > x_f$ *et* $y_i > y_f$) ni les deux points suivants a_{i+1} et a_{i+2} alors (3)
On retient le point a_i *dans l'alignement final* : $A \leftarrow A \cup \{a_i\}$.

3. Retour :
Si de nouveaux points ont été rajoutés à la précédente itération, on retourne en 1

4. Terminaison **:** *A* contient le résultat.

Pour le filtrage des points, nous avons mis en œuvre l'algorithme de façon séquentielle, chaque point étant filtré en fonction des points (déjà filtrés) considérés comme acquis. Cette méthode comporte des lacunes : si un point erroné a été conservé, il se peut qu'on élimine des points corrects à la suite, ce qui engendre une propagation de l'erreur et une diminution du rappel. Mais la contrainte de proximité de la diagonale diminue fortement les effets de ce genre de défaillance.

Notons que cet algorithme est basé sur un critère de fiabilité attaché au transfuge lui-même : si un transfuge ne possède pas un nombre d'occurrences identique dans les deux versions d'une même section, ou si une seule de ces occurrences est considérée comme suspecte (par exemple, un point est situé loin de la diagonale), tous les points générés par

ce transfuge sont ignorés. On suppose en effet, que dans toute traduction, certains transfuges sont plus stables que d'autre. Par exemple, il se peut que toutes les données numériques aient été transférées intégralement, ou bien les marqueurs de paragraphe ou de chapitre, ou encore certains noms propres. Dans ce cas le parallélisme des occurrences doit être parfait. Si la moindre divergence apparaît (par exemple un nom propre apparaissant une fois de plus dans le texte traduit), l'indice est jugé peu fiable, et abandonné.

En fait, nous avons quelque peu affaibli cette hypothèse, en réexaminant, à chaque itération, les occurrences d'un transfuge à l'intérieur de chaque section issue de l'alignement provisoire : un même transfuge peut être éliminé dans une section et accepté dans une autre si ses occurrences respectent le parallélisme.

Afin de minimiser les possibilités d'erreur, nous avons en outre appliqué une méthode de renforcement des indices :

- les indices discordants sont écartés : deux points entrant en conflit sont éliminés.
- chaque point doit être surdéterminé : un point n'est retenu que s'il a été généré au moins deux fois (i.e. par au moins deux transfuges différents).

L'accumulation de ces contraintes vise à produire un faisceau d'indices convergents qui sera d'autant plus fiable qu'il sera dense. Plus la trame de l'alignement ainsi produit est serrée, plus il est improbable qu'il soit globalement erroné.

On pourrait tenter une estimation, par le calcul, des probabilités de ce genre d'événement. Par exemple, on peut estimer la probabilité qu'un transfuge apparaissant n fois dans chaque texte puisse vérifier la contrainte de diagonalité seulement par le fait du hasard : pour une suite de n points pris au hasard, on peut calculer la probabilité que ces n points soient à l'intérieur d'une bande entourant la diagonale et occupant $1/10^{ème}$ de l'espace de recherche. On obtient une probabilité en 10^{-n}, ce qui est évidemment très faible pour $n > 2$. Si l'on observe un tel événement pour $n = 10$, par exemple, on peut rejeter l'hypothèse nulle (c'est-à-dire que cet événement soit dû au hasard).

Mais dans la réalité, le calcul est plus complexe, car lorsqu'on apparie n abscisses $x_{i=1..n}$ avec n ordonnées $y_{i=1..n}$ donnés dans un ordre croissant, on considère par construction des suites de points *monotones*. Il faudrait donc calculer la probabilité de l'événement « n points monotones tirés de façon aléatoire sont situés à l'intérieur de la bande autour de la

diagonale» : le calcul est beaucoup plus complexe et dépasse largement nos compétences mathématiques. Si cette bande est très fine, le calcul précédent peut en fournir une estimation, mais pour une bande correspondant à 36 % de l'espace de recherche (comme c'est le cas dans nos tests) il n'est plus valide.

Nous n'offrirons par conséquent aucun modèle permettant d'établir à partir de quelle valeur de *n* on peut rejeter l'hypothèse nulle. Nous nous appuierons sur des résultats empiriques que nous essaierons de rendre aussi généralisables que possible.

En résumé, le filtrage des points est fondé sur la conjonction des critères suivants :

(*1*) *Redondance*

Chaque point d'ancrage doit être donné par au moins deux transfuges différents.

(*2*) *Diagonalité*

Un point d'ancrage n'est retenu que s'il est situé à l'intérieur d'une bande autour de la diagonale. C'est en quelque sorte un espace de confiance au-delà duquel tous les points deviennent suspects. Le couple (*n*,*m*) représentant les coordonnées du point final de l'espace d'alignement, on calcule la distance de (*x*,*y*) à la diagonale de la manière suivante :

$$d(x, y) = \left| \frac{x}{n} - \frac{y}{m} \right| \qquad (23)$$

On peut ainsi éliminer tous les points dont $d(x,y) > Seuil_{diag}$. Pour donner un ordre d'idée, avec un $Seuil_{diag}$ égal à 0,2, la bande occupe une surface égale à 36 % de l'espace de recherche[148].

[148] Avec un seuil $Seuil_{diag}$, la proportion de la surface occupée est égale à :

$$S = 2\, Seuil_{diag} - Seuil_{diag}^{2}.$$

(*3*) *Continuité*

La contrainte de continuité traduit la propriété de quasi-bijection : il ne doit pas y avoir d'omission ou d'insertion trop importante dans le corpus, réalisant des ruptures brutales dans le chemin d'alignement. Autrement dit, un point donné ne doit pas effectuer une déviation trop forte par rapport au point précédent. La déviation entre (x,y) et le point précédent (x_{i-1},y_{i-1}), est donnée par :

$$Déviation = \frac{(x_i - x_{i-1})}{(y_i - y_{i-1})} \times \frac{m}{n} \quad (24)$$

Là encore, on rejette tous les points qui ne sont pas compris dans un certain intervalle autour de 1. Notons néanmoins que cette déviation n'a pas le même sens suivant les valeurs de (x_i-x_{i-1}) et (y_i-y_{i-1}). En effet, plus les points sont rapprochés, plus ils sont passibles de déviations importantes ; un « petit » saut de (1,1) à (2,5) est courant et produit pourtant la même déviation qu'un saut de (1,1) à (11,41), pourtant beaucoup plus rare. On utilisera donc des intervalles différents suivant l'écart entre les points : plus larges pour les points rapprochés, et plus près de 1 pour les points éloignés.

(*4*) *Monotonie*

La contrainte de monotonie implique que les points ne peuvent avoir de configuration croisée. Pour (x,y) et (x',y'), on impose les deux conditions suivantes :

$$x > y \Rightarrow x' > y' \text{ et } x = x' \Leftrightarrow y = y' \quad (25)$$

La deuxième de ces contraintes est propre au préalignement : les chevauchements seront admis lors de la détermination d'un chemin complet s'il s'agit de points contigus (pour intégrer les transitions (1:0) ou (0:1)).

Ces critères sont généraux, dans la mesure où ils expriment chacun un aspect du parallélisme. On les retrouvera dans les autres formes d'alignement.

II.3.5.2 Résultats expérimentaux

Nous avons appliqué l'algorithme précédemment décrit sur la totalité du corpus BAF. Nous avons pris en compte tous les transfuges d'au moins un caractère, y compris ceux correspondant à des caractères non alphanumériques, comme les guillemets, les parenthèses, les signes monétaires. Seuls ont été ignorés les signes de ponctuation principaux, n'apportant pas d'indice fiable vis-à-vis du parallélisme : le point, la virgule et le point virgule.

Nous avons fait l'hypothèse que deux facteurs étaient déterminants vis-à-vis des résultats : la densité des points d'ancrage utilisés, et leur nature (chaînes alphanumériques, mots avec majuscules sauf en début de phrase, et transfuges simples).

– *Première expérience*

Dans une première expérience, nous avons testé de façon indépendante chaque type d'ancrage. Les résultats numériques complets sont dans le tableau 84 de l'annexe. Le tableau 11 donne les valeurs moyennes obtenues (en excluant le sous-corpus *Xerox*, qui ne vérifie pas l'hypothèse de parallélisme au niveau de son glossaire, et pour lequel le rappel est toujours faible – nous noterons désormais BAF* le corpus BAF sans *Xerox*). Cette évaluation, comme toutes celles qui suivront, est basée sur les alignements de référence et les mesures établies dans le projet ARCADE.

	Précision		***Rappel***		***F-mesure***	
	Moy.	**Min.**	**Moy.**	**Min.**	**Moy.**	**Min.**
Alphanumériques	99,9 %	99,2 %	5,1 %	0,1 %	9,3 %	0,1 %
Majuscules	99,7 %	97,1 %	17,1 %	7,8 %	28,2 %	7,8 %
Transfuges	99,5 %	97,8 %	53,0 %	25,6 %	67,7 %	25,6 %
*Transfuges**	*97,6 %*	92,8 %	*72,4 %*	40,6 %	*82,3 %*	40,6 %

tableau 11 : *résultats moyens pour les différents types de points d'ancrage* (*corpus BAF**)

Pour *Xerox*, on obtient les résultats suivants :

	Précision	***Rappel***	***F-mesure***
Alphanumériques	100,00 %	0,02 %	0,04 %
Majuscules	99,66 %	0,34 %	0,67 %
Transfuges	99,83 %	2,38 %	4,64 %
*Transfuges**	97,04 %	3,13 %	6,06 %

tableau 12 : *résultats moyens pour les différents types de points d'ancrage* (*sous-corpus Xerox*)

Nous avons indiqué pour chaque colonne la valeur moyenne et le minimum obtenu. Les transfuges quelconques (qui englobent les deux autres catégories) sont notés *Transfuges*. La dernière ligne marquée d'un astérisque désigne la prise en compte des transfuges quelconques, sans appliquer la condition de surdétermination des points (redondance des transfuges), afin d'augmenter le rappel. La validité de nos hypothèses semble être confirmée : les données numériques engendrent un peu moins de bruit que les mots en majuscules, qui eux-mêmes sont plus fiables que des transfuges quelconques.

Notons que la complexité en temps de cet algorithme est presque linéaire, puisqu'inférieure à $O(n\ log(n))$, où n est la taille des textes[149].

Avec les résultats de la troisième ligne, on constate que cette méthode remplit les prérequis d'un préalignement efficace :

- les précisions obtenues sont très élevées puisqu'au-dessus de 99,5 % en moyenne (avec un minimum de 97,8 %) ;
- le rappel est important car supérieur à 50 % en moyenne (avec un minimum de 25 %).

Enfin les résultats de la quatrième ligne montrent que l'utilisation des transfuges peut aussi constituer une méthode d'alignement simple et de bonne qualité, produisant une F-mesure moyenne (sur ce corpus) supérieure à 80 %.

[149] Le terme $log(n)$ est dû aux recherches dans les index, sous forme d'arbres binaires.

– *Deuxième expérience*

Dans un deuxième temps, nous avons testé notre heuristique de précision d'abord : nous avons utilisé les alphanumériques, les majuscules puis les autres transfuges *successivement*, chaque étape reprenant les résultats de l'étape précédente (cf. tableau 84 de l'annexe).

	Précision	***Rappel***	***F-mesure***
Alphanumériques	99,9 %	5,1 %	9,3 %
+ Majuscules	99,2 %	20,4 %	32,1 %
+ Transfuges	99,5 %	57,1 %	71,4 %

tableau 13 : *résultats moyens pour les différents types de points d'ancrage utilisés successivement* (*corpus BAF**)

Au final, on gagne près de 4 % de F-mesure moyenne. Une certaine incidence de l'ordre dans lequel on exploite les différents indices semble donc se confirmer. Fait notable, c'est sur le rappel que cette incidence est la plus forte. Notre heuristique est basée sur l'exploitation des indices les plus précis d'abord : nous pensions ainsi limiter la propagation des erreurs générées dès les premières itérations ; mais il apparaît que les erreurs initiales ont plutôt un effet d'inhibition sur le rappel, en contribuant à l'élimination de points valides, et n'engendrent pas d'autres erreurs en cascade.

II.3.5.3 Corrélations avec les caractéristiques textuelles

Nous avons ensuite cherché à corréler les résultats obtenus avec la densité de transfuges. Nous avons calculé deux versions de cette densité : le nombre de transfuges par phrase (moyenné sur les deux textes T et T'), rapporté au nombre total de phrases ($d_{transfuge1}$) ou au nombre total d'unités ($d_{transfuge2}$). Pour chaque unité u apparaissant à l'identique dans les deux textes, et comptant respectivement $Occ(u)$ et $Occ'(u)$ occurrences dans T et dans T', on somme $min(Occ(u), Occ'(u))$, et l'on divise le total par la moyenne des tailles de T et T' (respectivement n et m) :

$$d_{transfuge} = \frac{\sum_{u=u'} \min(Occ(u), Occ'(u'))}{\frac{(n+m)}{2}} \qquad (13)$$

Pour donner un ordre d'idée, la densité par phrase oscille entre 1 et 2,5 pour les textes du corpus *BAF*.

Le tableau 14 montre les corrélations linéaires entre ces densités et les résultats obtenus à chaque étape de l'algorithme (*Alphanumériques*, *Majuscules*, *Transfuges*, *Transfuges**) :

	$d_{transfuge1}$	$d_{transfuge2}$
P	-0,13	-0,27
R	0,90	0,85
F	0,88	0,89

tableau 14 : *corrélation linéaire entre les résultats et la densité de transfuges*

On constate une forte corrélation entre la densité et le rappel : ce résultat paraît logique, dans la mesure où plus il y a de transfuges, plus on peut en tirer de points d'ancrage. En outre, cette corrélation est plus marquée avec la densité par phrase qu'avec la densité par unité : la densité phrastique coïncidant avec le niveau de granularité de l'alignement, elle est donc mieux corrélée à l'utilisation des transfuges dans notre algorithme.

Par ailleurs, la corrélation négative entre densité et précision n'est pas significative : l'augmentation du rappel au cours de l'algorithme n'engendre pas une dégradation importante de la précision.

Ces remarques sont illustrées par les nuages de points suivants, où sont représentées les distributions de *P*, *R* et *F* en fonction de d_{phrase} (à chaque texte correspondent 3 points, correspondant aux trois types de transfuges : *Alphanumériques*, *Majuscules*, *Transfuges* ; cf. tableau 85 de l'annexe):

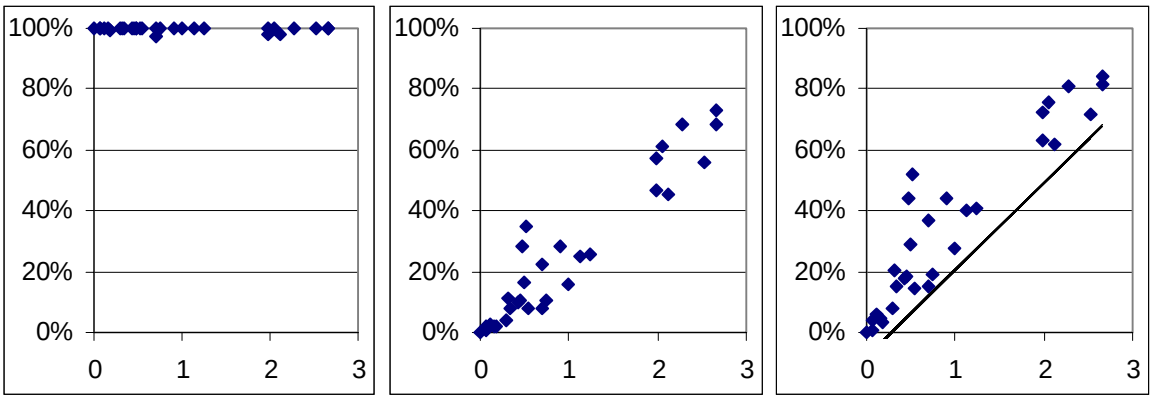


figure 19 : *précision, rappel et F-mesure en fonction de* D_{phrase}

Notons que l'on peut minorer *F* par une fonction linéaire de la densité. Dans le cas présent, par exemple, tous les points se situent au-dessus de la droite représentée figure 19 :

$$F \geq d_{transfuge} / 3{,}5 - 0{,}08$$

Il faudrait conduire une étude empirique plus approfondie pour déterminer la portée d'une telle minoration. En effet, les résultats de cette méthode dépendent de propriétés formelles des textes concernés, comme $d_{transfuge}$, mais aussi de relations traductionnelles (comme le parallélisme ou la proportion d'homographes qui ne sont pas traductions mutuelles) dont on ne peut présumer sans connaissances linguistiques. Il nous semble néanmoins qu'étant donné la multiplicité des contraintes de filtrage que nous imposons, dans le but de limiter le bruit, le facteur déterminant reste $d_{transfuge}$, cette densité permettant d'estimer *a priori* les résultats du préalignement.

II.3.6 Exploitation des cognats

Nous faisons l'hypothèse que l'exploitation des cognats peut être utile dès les premières phases de préalignement, au même titre que les transfuges, le critère déterminant étant la prééminence de la précision sur le rappel.

II.3.6.1 Résultats préliminaires

Dans un premier temps, nous avons mené une série d'observations sur deux textes alignés du corpus Europarl[150], afin de dégager les propriétés formelles du bi-texte sur le plan de la densité des cognats. Ce bi-texte est constitué de 490 x 490 segments alignés manuellement. Pour chaque texte, nous avons regroupé ces segments en 49 blocs de 10 phrases. Un comptage des 4-grammes$^+$ relevés[151] entre chaque bloc anglais - français nous a permis d'extraire une matrice des densités. Pour deux textes T et T' on écrit :

$T = (B_1 B_2 \ldots B_{49})$ $T' = (B'_1 B'_2 \ldots B'_{49})$

n_{ij} = nombre de 4-grammes$^+$ entre B_i et B'_j

l_i et l'_j sont les longueurs respectives (en nombre de mots) des blocs B_i et B'_j. La densité de 4-grammes$^+$ entre i et j s'écrit :

$$d_{ij} = \frac{n_{ij}}{l_i \cdot l'_j} \tag{26}$$

On obtient une première matrice $D = (d_{ij})_{i=1..n,\, j=1..m}$ contenant, pour chaque couple de blocs (B_i, B'_j) la densité de cognats associée.

Si l'on représente sur un graphique l'ensemble des points dont la densité est supérieure à la moyenne, on aboutit à la figure 20. Les lignes verticales et horizontales indiquent un effet marginal important : certains blocs sont très « productifs » en 4-grammes$^+$ communs, et « pèsent » donc plus lourd que d'autres. Ceci peut s'expliquer par la présence de groupes de lettres récurrents entre l'anglais et le français : par exemple *t-i-o-n*, ou *m-e-n-t*.

---

[150] Rapport *a4-0391/96*, sur les droits des personnes handicapées (rapporteur Mme Mary Banotti), en français et en anglais, alignés manuellement par nous au niveau des phrases.

[151] On notera 4-gramme$^+$ une chaîne commune d'au moins 4 caractères consécutifs, et 4-gramme une chaîne commune de 4 caractères exactement. Tous les mots du bloc anglais sont comparés avec tous les mots du bloc français. On compte au maximum un 4-gramme par couple de mots comparés.

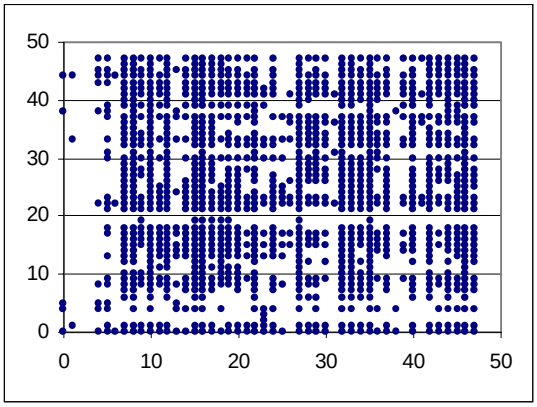


figure 20 : *points* (*i*,*j*) *vérifiant* d_{ij} > *moy*

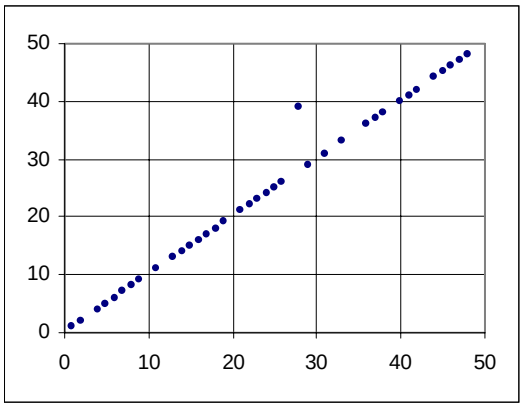


figure 21 : *points* (*i*,*j*) *filtrés*

Pour neutraliser le bruit engendré par ces chaînes récurrentes, nous proposons d'utiliser la mesure C_{ij}, exprimant l'information apportée par n_{ij} par rapport aux marges n_i et n_j :

$$C_{ij} = \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} \tag{27}$$

avec

$$N = \sum_i \sum_j n_{ij} \qquad f_{ij} = \frac{n_{ij}}{N} \qquad f_i = \frac{n_i}{N} \qquad f_j = \frac{n_j}{N}$$

On obtient alors la matrice d'association $C = (C_{ij})_{i=1..n,\, j=1..m}$

Si l'on considère tous les points (*i*,*j*) tels que P_i atteint son association maximum avec P'_j, et P'_j atteint son association maximum avec P_i (condition de réciprocité) :

$$(i,j) \,/\, i = argmax_{\,k=1..n}\,(c_{kj}) \; et \; j = argmax_{\,k=1..m}\,(c_{ik})$$

on obtient les points de la figure 21. Il est notable que cette fois, les points retenus (sauf un) correspondent à l'alignement correct des blocs.

Si l'on applique ensuite les trois critères de filtrage issus du parallélisme, *diagonalité*, *continuité*, et *monotonie*, on arrive à 36 points correctement alignés sur 49, soit une précision de 100 % et un rappel de 73 %.

II.3.6.2 Description de l'algorithme

Il semble donc qu'un filtrage adéquat de la densité de cognats puisse conduire à une série de points d'ancrage très fiables. Un problème se pose néanmoins : le calcul de la matrice d'association précédemment décrite est en $O(n^2)$. Pour un super calculateur, ce n'est pas rédhibitoire, mais pour les moyens limités dont nous disposons[152], les temps de calcul sur de gros textes deviennent vite importants.

Pour diminuer les calculs, il suffit, en suivant les principes heuristiques énoncés, d'utiliser les résultats acquis à l'étape précédente avec les transfuges, et de mesurer la cognation seulement à l'intérieur des îlots de confiance ainsi dégagés. Des matrices d'association différentes sont alors construites entre chaque couple de points d'ancrage successifs : elles sont obtenues à partir des points situés dans une bande de largeur constante autour des points de la sous-diagonale définie par les points d'ancrage.

La figure 22 représente l'espace de recherche qui en résulte.

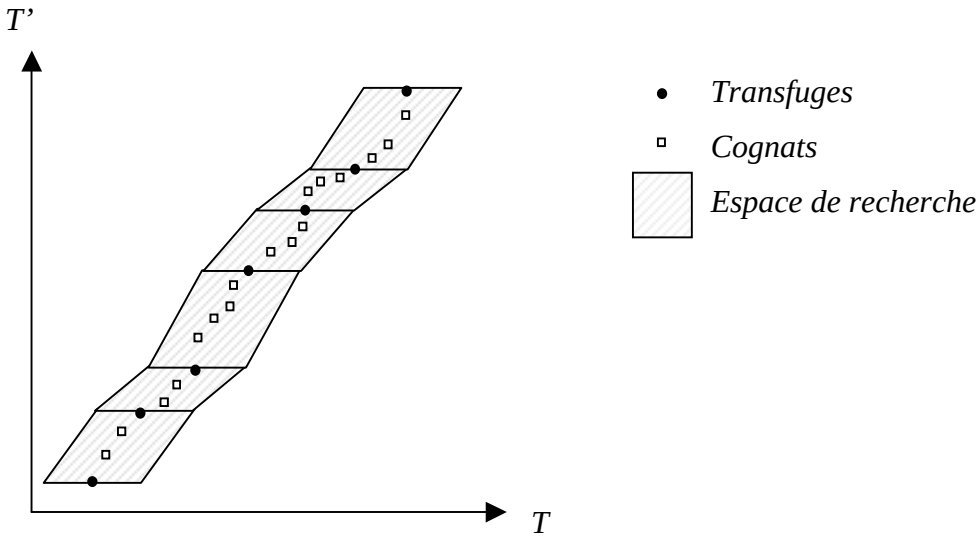


figure 22 : *représentation de l'espace de recherche délimité par des points d'ancrage.*

[152] tous les algorithmes présentés dans ce travail ont été mis en œuvre sur un PC équipé d'un Pentium à 200 Mhz.

Si le rappel du préalignement est borné, supérieur à un certain seuil, par exemple 10 %, l'espace de recherche résultant est en $O(n)$.

Pour que les matrices d'association soient significatives, nous avons imposé qu'elles soient d'une largeur supérieure ou égale à dix phrases : nous n'avons donc retenu du préalignement qu'une trame de points éloignés entre eux d'au moins dix phrases.

– *Algorithme* :

On note :

Préal l'ensemble des points d'ancrage fournis par la phase précédente, A_{cand}, l'ensemble des points d'ancrage avant filtrage, et A l'ensemble des points d'ancrage en cours de construction

(I_0,J_0) le premier point d'ancrage, $(I_{début}, J_{début})$ et $(I_{prochain}, J_{prochain})$ les couples de variables désignant le premier et le second point d'ancrage de l'îlot de confiance courant.

0. Initialisation :
$I_{début} \leftarrow I_0$, $J_{début} \leftarrow J_0$, avec (I_0,J_0)
$A \leftarrow \varnothing$

1. Boucle principale
Tant qu'il existe $(I_{prochain}, J_{prochain}) \in$ *Préal* / $I_{prochain} - I_{début} \geq 10$, $J_{prochain} - J_{début} \geq 10$
Calculer la matrice $(C_{ij})_{I=Idébut…Ifin,J=Jdébut…Jfin}$
$A_{cand} \leftarrow \varnothing$
Pour tout I de $I_{début}$ à $I_{prochain}$
Calculer J_{max} tel que $C_{IJmax}=max_{J=Jdébut…Jprochain}(C_{IJ})$
Calculer I_{max} tel que $C_{ImaxJmax}=max_{I'=Idébut…Iprochain}(C_{I'Jmax})$
Si $I=Imax$ Alors $A_{cand} \leftarrow \{(I,J_{\max})\} \cup A_{cand}$
Pour $(I,J) \in A_{cand}$
Si (I,J) vérifie les critères de *diagonalité*, *continuité* et *monotonie* par rapport à A, alors
$A \leftarrow A \cup \{(I,J)\}$
$I_{début} \leftarrow I_{prochain}$, $J_{début} \leftarrow J_{prochain}$

Retour au début de la boucle ou

2. *Terminaison* : *A contient le résultat*

II.3.6.3 Premiers résultats

Deux modes de comptage des cognats potentiels ont été testés : dans le modèle *symétrique,* les n-grammes entre *P* et *P'* sont dénombrés lors de la comparaison de tous les mots de *P* avec tous les mots de *P'* ; dans le modèle *dissymétrique* on compte tous les mots de *P* ayant au moins un n-gramme commun avec un des mots de *P'*. Cette dernière version est (très) légèrement plus rapide, car dès qu'un mot de *P* a été désigné comme cognat potentiel, on arrête les comparaisons avec ce mot et l'on passe au suivant.

Pour comparer ces deux modèles, nous sommes parti d'une définition simplifiée des candidats cognats : les 3-grammes^{+} et/ou les transfuges.

Les résultats complets ainsi que les paramétrages de ces deux tests sont consignés dans le tableau 86 de l'annexe. Les valeurs moyennes de précision et rappel sont les suivantes :

Modèle	***Précision***		***Rappel***		***F-mesure***	
	Moy.	Min.	Moy.	Min.	Moy.	Min.
Dissymétrique	99,0 %	95,8 %	74,0 %	53,5 %	84,4 %	68,7 %
Symétrique	98,3 %	94,3 %	71,4 %	57,3 %	82,3 %	71,3 %

tableau 15 : *comparaison des modes de comptage des cognats potentiels* (*corpus BAF**)

Pour *Xerox* on a :

Modèle	***Précision***	***Rappel***	***F-mesure***
Dissymétrique	99,1 %	2,8 %	5,4 %
Symétrique	99,1 %	2,9 %	5,7 %

tableau 16 : *comparaison des modes de comptage des cognats potentiels* (*sous-corpus Xerox*)

Notons que les minima sont atteints par la traduction littéraire (*Verne*). Par rapport à la F-mesure du préalignement avec les transfuges, on a gagné 13 points : de 71,4 % à 84,4 % avec une très légère dégradation de la précision (de 99,5 % à 99 %).

Globalement, la différence entre les deux modèles est ténue, mais le modèle dissymétrique correspond sans doute mieux à la réalité de la traduction. En effet, un cognat potentiel, une fois identifié, ne devrait intervenir qu'une fois dans la mesure où il n'est traduit qu'une fois dans une même phrase. Tandis que dans le modèle symétrique, une seule chaîne du type *t-i-o-n* risque d'engendrer un grand nombre de n-grammes, par la multiplication des comparaisons, et de créer un surcroît de bruit.

Toutefois, pour des raisons de symétrie dans le traitement des deux langues, et pour nous rapprocher du modèle de cognation élaboré par Davis *et al*. (1995), nous utiliserons le modèle symétrique dans les expérimentations suivantes.[153]

II.3.6.4 Détermination des cognats

Pour la détermination des cognats, nous nous sommes jusqu'ici contenté de recourir à la méthode commune, basée sur une approximation grossière : tout couple de mots comportant une chaîne commune de plus de 3 caractères constitue un couple de cognats potentiels.

II.3.6.4.1 Précision et rappel des n-grammes

Nous voulons maintenant déterminer dans quelles proportions les appariements ainsi dégagés correspondent à de véritables couples de cognats, ou produisent du bruit. Nous verrons ensuite comment les résultats du filtrage des cognats peuvent être corrélés avec l'alignement produit.

Pour comparer les *cognats candidats* (identifiés automatiquement) avec les *cognats de référence* (les mots apparentés en relation de traduction), il faut dans un premier temps déterminer manuellement ces cognats de référence. Etant donné l'ampleur de la tâche, nous nous sommes restreint à une partie de notre corpus, en l'occurrence les textes du sous-corpus *Cour*. Pour établir cette liste de cognats de référence, nous nous sommes basé sur les critères précédemment énoncés : étymologie commune et équivalence traductionnelle.

---

[153] Globalement, nous avons effectivement vérifié que les différences dans les résultats sont négligeables.

Le deuxième critère repose sur la possibilité de trouver un contexte à l'intérieur duquel les cognats peuvent être considérés comme équivalents. Cette possibilité n'étant pas toujours facile à évaluer, nous contournons cette difficulté par un parti pris restrictif : au sein de notre corpus, nous n'identifions comme cognats que les couples de mots qui sont effectivement en relation d'équivalence, dans le corpus[154]. Ceci peut introduire un léger biais dans nos résultats, avec une précision parfois sous-évaluée. Par exemple, (angl.) *appeal* et (fr.) *appelant* n'apparaissent pas comme traductions mutuelles dans notre corpus, et ne sont donc pas retenus comme couple de cognats, alors qu'en principe ce sont bien des cognats.

Par ailleurs, lorsqu'on confronte les candidats avec les couples de référence, toutes les chaînes homographes sont assimilées, qu'il s'agisse d'homonymes ou d'unités polysémiques utilisées dans des acceptions différentes. Cette assimilation peut conduire à accepter des couples d'unités non-équivalentes, comme *because* et *cause* (dans le sens de « *parle* »), et augmenter artificiellement la précision. Cependant, ces deux biais antagonistes n'affectent le calcul de la précision que de façon marginale et non significative, et celle-ci ne nous intéresse pas pour sa valeur absolue mais pour ses variations, qui sont statistiquement indépendantes de ces biais.

Pour évaluer les résultats de l'identification automatique des cognats, on utilise les mesures habituelles de précision et rappel :

- la précision *Pc* exprime la proportion de couples de cognats corrects par rapport au nombre de candidats extraits :

$$Pc = \frac{\left|CognatsCandidats \cap CognatsDeRéférence\right|}{\left|CognatsCandidats\right|} \tag{28}$$

- le rappel *Rc* exprime la proportion de couples de cognats corrects par rapport nombre de couples cognats de référence :

[154] En d'autres termes, nous avons extrait toutes les correspondances lexicales qui impliquaient des lexies apparentées. Ces correspondances ont été établies manuellement, sur la base de l'identité de désignation des lexies. On a ainsi relevé une liste de 944 couples différents, sans compter les transfuges.

$$Rc = \frac{|CognatsCandidats \cap CognatsDeRéférence|}{|CognatsDeRéférence|} \quad (29)$$

Il existe deux façons de calculer les valeurs de *Pc* et *Rc* : soit on compare toutes les *unités types*[155] de T avec toutes les *unités types* de T' ; soit on compare toutes les *unités occurrences* de T avec toutes les *unités occurrences* de T' , un même couple pouvant intervenir plusieurs fois dans l'évaluation. Nous avons opté pour cette dernière solution, en limitant les comparaisons à l'ensemble des couples de phrases qui font partie de l'espace de recherche de l'algorithme d'alignement : les statistiques sont ainsi directement liées aux cognats réellement utilisés par cet algorithme[156].

Pour chaque comparaison d'un couple d'occurrences (*u*,*u*'), quatre cas de figure peuvent se présenter, qu'on dénombre séparément, comme le montre le tableau 17 :

a+b+c+d = nombre total de couples (*u*,*u*') *comparés*	*couples* (*u*,*u*') *appartenant à la liste des Cognats de référence*	*couples* (*u*,*u*') *n'appartenant pas à la liste des Cognats de référence*
couples (*u*,*u*') *retenus comme candidats*	a	b
couples (*u*,*u*') *non retenus comme candidats*	c	d

tableau 17 : *comparaison des cognats candidats avec les cognats de référence*

On a donc : $Pc = a / (a + b)$ et $Rc = a / (a + c)$

On obtient les statistiques des trois premières colonnes du tableau 18. Nous avons fait figurer, dans la première colonne, les statistiques de précision et rappel liées aux transfuges, afin de servir de base de comparaison. Notons qu'un certain nombre de ces transfuges sont également comptés dans les n-grammes.

[155] Par *types* on entend l'ensemble des unités apparaissant dans chaque texte, chaque unité étant comptée une fois indépendamment de sa fréquence. Par *occurrences* on désigne toutes les occurrences particulières des unités *types*.

[156] On verra que ces statistiques résultent de la comparaison de phrases voisines, à l'intérieur des sections préalignées automatiquement. Entre des phrases quelconques on peut supposer qu'elles

	transfuges			***n-grammes***			***n-grammes + transfuges***			***SCM***			***SCM + transfuges***		
N	***Pc %***	***Rc %***	***Fc %***	***Pc %***	***Rc %***	***Fc %***	***Pc %***	***Rc %***	***Fc %***	***Pc %***	***Rc %***	***Fc %***	***Pc %***	***Rc %***	***Fc %***
≥ 2	100	50	66												
≥ 3	100	29	45	15	75	24	18	95	30	47	55	51	55	76	63
≥ 4	100	21	34	31	48	37	41	76	54	64	45	53	75	74	74
≥ 5	100	16	28	48	38	42	64	71	67	75	39	51	85	72	78
≥ 6	100	15	26	72	26	39	86	61	71	74	30	43	86	65	74
≥ 7	100	13	22	87	19	31	95	56	71	76	21	33	90	58	70
≥ 8	100	7	14	88	10	18	97	52	68	73	13	22	92	55	69
≥ 9	100	6	12	93	8	15	99	52	68	93	10	18	99	53	69

tableau 18 : *résultats comparés des méthodes d'identification des cognats*

La figure 23 représente l'évolution des résultats en fonction de la longueur des n-grammes. L'augmentation de la précision avec le nombre de caractères communs montre que plus un n-gramme est long, plus il est fiable dans la détermination des cognats. Malheureusement les indices qui génèrent le moins de bruit sont aussi les plus rares.

On constate que la prise en compte des transfuges d'au moins 2 caractères[157] donne une meilleure F-mesure qu'avec les n-grammes (la précision est de 100 % car nous avons considéré de manière identique tous les couples homographes). En revanche, les 50 % de rappel obtenu par les transfuges indiquent clairement ce que peuvent apporter les n-grammes, ou d'autres techniques : il reste 50 % de cognats à identifier. On peut sans doute améliorer les résultats globaux en combinant l'identité des transfuges et la ressemblance des *n-grammes*. C'est ce que montre la troisième colonne du tableau 18 (dans cette colonne, les transfuges sont de longueur quelconque, les variations de *n* ne s'appliquant qu'aux n-grammes).

seraient différentes (pour des raisons de continuité thématique), avec une précision et un rappel inférieurs.

[157] Pour les transfuges comptant exactement 2 caractères, nous n'avons tenu compte que des nombres.

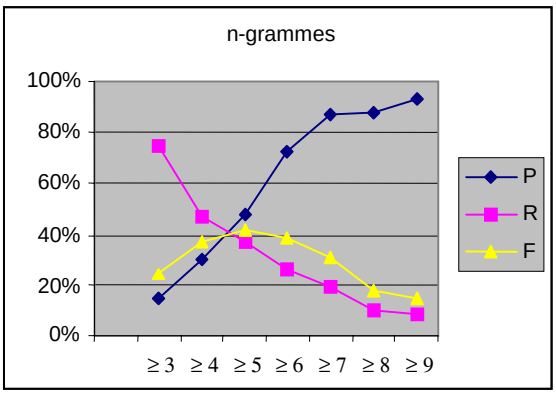


figure 23 : *identification des cognats et longueur des n–grammes*

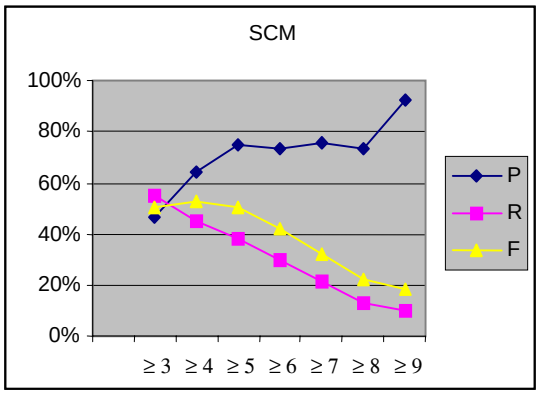


figure 24 : *identification des cognats et longueur des SCM*

II.3.6.4.2 Sous-chaînes maximales

L'identification des cognats par les n-grammes présente deux défauts :

- d'une part, les n-grammes ne permettent pas de reconnaître la « ressemblance » lorsque celle-ci implique des ruptures à l'intérieur des groupes de lettres : par exemple *doc**tor*** (angl.) et *dot**tor**e* (it.) n'ont au plus que 3 caractères consécutifs communs.

- d'autre part, la signification d'un n-gramme dépend étroitement de la taille des mots comparés. Il est clair que 4 caractères consécutifs communs entre (angl.) ***form*** et (fr.) ***form**e* sont plus significatifs qu'un 6-grammes entre (angl.) *explo**ration*** et (fr.) *décla**ration***.

Pour pallier le premier inconvénient, nous proposons de recourir aux sous-chaînes maximales (on notera SCM), à l'instar de Débili et Sammouda (1992) : la plus longue sous-chaîne de caractères commune aux deux mots (en autorisant les sauts). Par exemple, pour (fr.) *docteur* et (it.) *dottore*, la SCM est de longueur 4 : *d-o-t-r*. Mais la combinatoire des SCM est très importante (surtout avec les mots longs), et risque de produire beaucoup de bruit : par exemple (angl.) *pragmatic* est presque totalement inclus dans (fr.)

paradigmatique. Nous en avons donc implémenté une version plus contrainte : les sous-chaînes doivent être *quasiment parallèles*, c'est-à-dire que l'on n'autorise pas plusieurs décrochements du parallélisme (insertion ou omission) en série, et les décrochements ne sont tolérés que lorsqu'ils sont encadrés de caractères identiques.

Ainsi, *p-r-a-g-m-a-t-i* n'est pas une sous-chaîne quasi-parallèle de *pragmatic* et *paradigmatique*, car les caractères *d-i* représentent deux décrochements consécutifs.

Enfin, pour limiter le bruit nous tiendrons compte de la longueur des SCM par rapport à la taille des mots. On calcule le rapport entre la taille du mot le plus long et la longueur de la SCM. Entre deux unités u_1 et u_2 on calcule :

$$r(u_1, u_2) = \frac{l(SCM)}{\max(l(u_1), l(u_2))} \tag{30}$$

Puis on effectue un filtrage en fonction de r. Pour notre corpus nous avons testé différentes valeurs pour ce seuil : les meilleurs résultats ont été obtenus en acceptant les SCM avec $r \geq 2/3$ (cf. tableau 87 de l'annexe).

Les colonnes 4 et 5 du tableau 18 contiennent les résultats avec les SCM seules, puis combinées avec les transfuges. La figure 23 montre les résultats liés aux *SCM* en fonction de leur longueur. On constate que l'évolution est parallèle à celle des *n*-grammes, avec cependant un niveau de précision supérieur pour un rappel similaire dès $n \geq 4$. La valeur maximum de F_C est de 53 %, soit 11 points de plus qu'avec les *n*-grammes.

II.3.6.4.3 Corrélations avec l'alignement

Il reste à étudier l'incidence de cette amélioration sur la mise en œuvre de l'algorithme d'alignement. Nous avons donc cherché à corréler les résultats de la méthode d'identification des cognats (P_c, R_c, F_c) avec les résultats de son exploitation pour l'alignement (P_a, R_a, F_a).

Les résultats obtenus pour différents types de paramétrages figurent dans le tableau 19. Nous avons en outre testé deux mesures combinant n-grammes et SCM (lignes 10 et 11), et nous avons extrait un alignement en utilisant la liste des cognats de référence, déterminés manuellement (dernière ligne).

	P_c %	R_c %	F_c %	P_a %	R_a %	F_a %
3-grammes$^+$	14,6	74,5	24,4	82,2	31,5	45,5
4-grammes$^+$	30,6	47,5	37,2	97,2	62,8	76,3
CMS ≥ 3 + transfuges	54,7	75,7	63,5	99,5	75,5	85,9
CMS ≥ 4 + transfuges	74,6	74,0	74,3	99,8	85,7	92,2
CMS ≥ 5 + transfuges	84,7	71,8	77,7	99,8	85,2	91,9
CMS ≥ 6 + transfuges	85,8	64,7	73,7	99,9	84,0	91,2
CMS ≥ 7 + transfuges	89,7	58,0	70,5	99,9	81,7	89,9
CMS ≥ 8 + transfuges	92,1	55,4	69,2	99,7	79,1	88,3
CMS ≥ 9 + transfuges	98,5	53,4	69,3	99,7	76,1	86,3
Combinaison 1 (**)	67,7	73,3	70,4	99,4	79,9	88,6
Combinaison 2 (***)	75,0	73,3	74,1	99,7	86,2	92,5
Transfuges seuls	100,0	49,6	66,3	99,7	68,4	81,2
Cognats (****)	100,0	100,0	100,0	99,8	74,3	85,2

(**) Combinaison 1 : transfuges, 4-grammes$^+$ (mots de longueur <7), CMS avec N≥4 et r>2/3

(***) Combinaison 2 : transfuges, 3-grammes$^+$ avec r>2/3, CMS avec N≥5 et r=2/3

(****) Cognats : liste de référence obtenue manuellement

tableau 19 : *identification des cognats et alignements résultants (corpus Cour)*

La meilleure F-mesure est obtenue avec la combinaison 2.

Les corrélations linéaires entre P_c, R_c, F_c d'une part et P_a, R_a, F_a d'autre part sont consignées dans le tableau 20 (sans tenir compte de la première ligne du tableau 19) :

	P_a	R_a	F_a
P_c	0,76	0,74	0,75
R_c	0,41	0,72	0,71
F_c	0,85	0,94	0,93

tableau 20 : *corrélations linéaires entre* P_c, R_c, F_c *et* P_a, R_a, F_a

Si l'on coordonne entre elles précisions, rappels et F-mesure on obtient les nuages de points suivants :

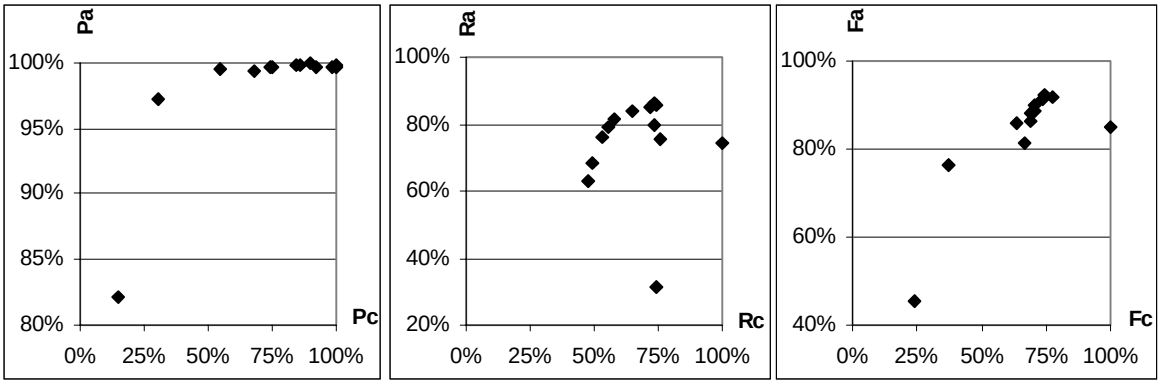


figure 25 : *résultats de l'alignement en fonction des résultats de l'identification des cognats*

On observe que :

– la précision de notre méthode d'alignement est peu sensible au bruit : même pour une précision P_c de 15 %, la précision de l'alignement demeure au-delà de 80 %. Cette robustesse est due à la multiplicité des contraintes de filtrage des points d'ancrage.

– le rappel de l'alignement, très sensible à la qualité du filtrage des cognats, est fortement corrélé au résultat global de ce filtrage : entre F_c et R_a, la corrélation linéaire est de 0,94. Cela confirme l'amélioration des résultats apportée par le recours au SCM. La densité des cognats identifiés entre les phrases des deux textes est donc déterminante.

– un point est marginal : l'alignement obtenu avec les données des cognats de référence (dernière ligne) est légèrement moins bon. Cela pourrait indiquer que trop de cognats (un rappel trop important) peuvent affecter le rappel de l'alignement. Ce peut être lié à la nature de notre méthode : en effet la mesure du lien a tendance à favoriser les appariements avec les phrases courtes. Dès lors, deux phrases « pesant trop lourd » auraient moins de chances d'obtenir un bon score, et donc d'être alignées ensemble. Mais cette hypothèse demanderait une étude plus approfondie pour être confirmée. Quoi qu'il en soit il est raisonnable de

supposer que l'amélioration de F_c peut conduire à des résultats meilleurs pour F_a, moyennant une exploitation différente de la cognation.

II.3.6.5 Résultats généraux

Nous avons appliqué la première combinaison d'indices (transfuges + 4-grammes$^+$ pour l<7, SCM pour n ≥ 4) à l'ensemble du corpus. On obtient une F-mesure moyenne globale de 88,4 % pour le corpus BAF*, et de 5 % pour le sous-corpus *Xerox* (cf. tableau 88 de l'annexe).

On doit pouvoir améliorer les résultats en attribuant des poids différents à chaque cas rencontré, en fonction de sa probabilité de représenter un véritable couple de cognats. On peut calculer cette pondération en fonction des précisions observées des différents cas de figure.

Nous avons testé deux jeux de coefficients : le premier reflétant l'évolution et les irrégularités de la précision observée sur le sous-corpus *Cour*, et le second cherchant à gommer certaines irrégularités[158] en favorisant les SCM les plus longues. Ces pondérations sont appliquées aux fréquences f_{ij} de l'équation (27), en fonction de 9 cas de figures ressortant de la comparaison :

Cas	***transfuges***	***4-gram$^+$ l<7***	***SCM 4***	***SCM 5***	***SCM 6***	***SCM 7***	***SCM 8***	***SCM 9***	***SCM ≥10***
Précision	100 %	48 %	15 %	76 %	62 %	58 %	33 %	83 %	100 %
Pondération 1	10	5	1	8	6	6	3	8	10
Pondération 2	10	6	2	6	7	8	8	9	10

tableau 21 : *pondérations des différents cas de figure enregistrés pour l'identification des cognats*

Les résultats globaux pour chaque jeu de coefficients sont les suivants (cf. tableau 89 de l'annexe) :

[158] Etrangement, on constate que les sous-chaînes de longueur 8 sont peu fiables, et obtiennent une précision plus faible que les sous-chaînes de longueur 4.

Modèle	***Précision***		***Rappel***		***F-mesure***	
	Moy.	*Min.*	*Moy.*	*Min.*	*Moy.*	*Min.*
Sans pondération	99,4 %	97,9 %	80,0 %	60,5 %	88,4 %	74,8 %
Pondération 1	99,6 %	98,6 %	84,5 %	62,0 %	91,2 %	76,1 %
Pondération 2	99,5 %	97,2 %	83,6 %	59,7 %	90,6 %	74,0 %

tableau 22 : *résultats comparés de l'alignement avec les deux pondérations* (*BAF**)

Modèle	***Précision***	***Rappel***	***F-mesure***
Sans pondération	99,0 %	2,6 %	5,0 %
Pondération 1	99,2 %	2,7 %	5,2 %
Pondération 2	98,9 %	2,8 %	5,4 %

tableau 23 : *résultats comparés de l'alignement avec les deux pondérations* (*sous-corpus Xerox*)

Tous les minima du tableau 22 sont atteints avec le sous-corpus littéraire *Verne*. A part pour ce corpus, les pondérations 1 et 2 aboutissent à une amélioration légère de résultats. Globalement, on est passé de 84,4 % de F-mesure moyenne avec l'utilisation simple des 3-grammes^{+}, à 91,2 % avec une caractérisation plus raffinée des cognats. Notamment, le rappel a connu une augmentation très nette d'environ 10 points de 74 % à 84,5 %.

Le jeu de pondération 2 a été testé sur un modèle dissymétrique, mais les différences obtenues sont insignifiantes (F-mesure moyenne = 90,4 %).

Aucune hypothèse linguistique n'ayant été faite, les méthodes basées sur le principe de ressemblance superficielle sont généralisables à d'autres couples de langue : on peut s'attendre à des variations importantes au niveau du rappel, la précision restant élevée (du fait des contraintes de filtrage). Quant aux paramétrages précis qui donnent les résultats optimaux, ils sont vraisemblablement liés aux particularités du corpus étudié et des langues impliquées.

II.3.6.6 Corrélations avec les caractéristiques textuelles

On peut supposer que la réussite de la méthode décrite est fortement conditionnée par la fréquence des cognats observés au sein du corpus. Plus cet indice fournit d'information, plus il devrait permettre un alignement correct et complet.

Pour valider cette hypothèse, nous avons compté, pour chaque texte, le nombre de candidats cognats identifiés. Nous avons ensuite rapporté cette quantité au nombre de couples de mots comparés. On peut alors étudier les corrélations entre la densité de candidats cognats observés et les résultats de l'alignement.

	d_{cognat}
P_a	0,79
R_a	0,82
F_a	0,82

tableau 24 : *corrélation entre les résultats de l'alignement et* d_{cognat} (*BAF**)

Si l'on représente chaque texte du corpus par un point, avec la densité de cognat en abscisse et les résultats de l'alignement en ordonnée, on obtient les nuages suivants (en reprenant les résultats du tableau 22 et du tableau 23, avec la pondération 2, cf. tableau 90 de l'annexe) :

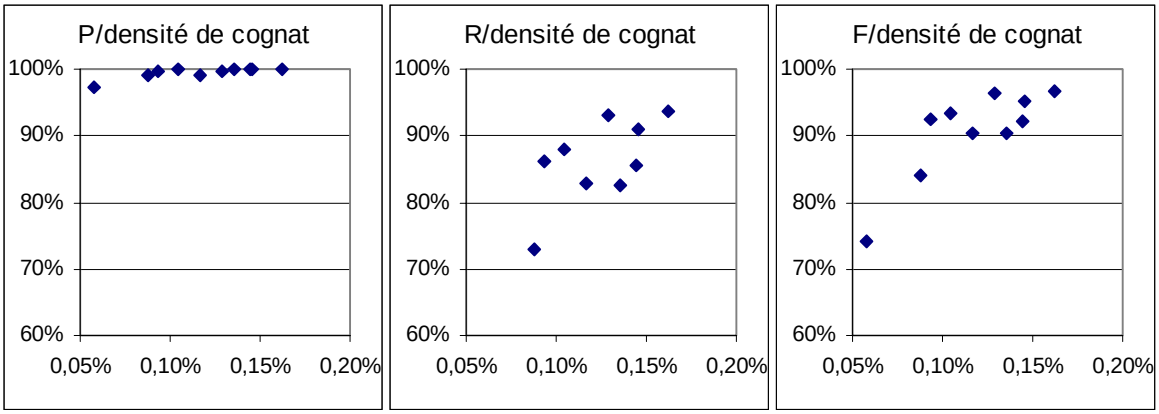


figure 26 : *résultats de l'alignement en fonction de* d_{cognat} (*BAF**)

On constate le même type de corrélation que précédemment : faible sensibilité de la précision, et corrélation forte du rappel de l'alignement avec la densité de cognats

potentiels. Cette corrélation donne un caractère de prévisibilité aux résultats obtenus : pour deux textes dont on sait qu'ils recèlent, dans la comparaison, de nombreux cognats potentiels (on peut estimer cette densité à partir d'un échantillon), on prévoit que le rappel de l'alignement basé sur les cognats (au sens large, incluant les transfuges) sera bon.

Comme avec les transfuges seuls, on peut donner une minoration linéaire, empiriquement vérifiée sur le corpus BAF* :

$$F \geq 500 \cdot d_{cognat}$$

II.3.7 Intégration dans un cadre de programmation dynamique

L'architecture dynamique est employée en dernier ressort, en conformité avec notre heuristique de précision d'abord. En effet, ses résultats sont sensibles aux irrégularités des textes et si l'information vient à manquer, le chemin d'alignement peut se perdre, et entraîner une chute catastrophique des performances. La robustesse étant moindre, il est indiqué d'utiliser cette technique à l'intérieur d'îlots de confiance de taille modeste, fournis par les étapes précédentes. En revanche, c'est à cette étape qu'on peut véritablement maximiser le rappel, dans la mesure où un chemin représente un alignement « complet », à la différence d'un ensemble de points d'ancrage.

L'espace de recherche est déterminé par la suite des points d'ancrages hérités des étapes précédentes. Entre deux points d'ancrage, on considère la *sous-diagonale* formée par la ligne idéale passant par ces deux points : l'espace de recherche est alors défini comme une bande centrée sur cette ligne, un couloir dont la largeur est proportionnelle à la distance séparant les points d'ancrage. Si $P_1 = (X_1,Y_1)$ et $P_2 = (X_2,Y_2)$ sont les coordonnées de deux points d'ancrages successifs, l'espace de calcul est défini par la formule :

$$d(x,y) = \left| \frac{(x - X_1)}{(X_2 - X_1)} - \frac{(y - Y_1)}{(Y_2 - Y_1)} \right| < Seuil \qquad (31)$$

Nous avons fixé la valeur de ce seuil à 0,2, sauf dans le cas où les deux points P_1 et P_2 sont très rapprochés : en effet, les irrégularités de traduction peuvent entraîner des écarts

importants (en proportion) sur un petit intervalle. On a donc utilisé trois valeurs suivant les cas de figure :

$$\begin{array}{llll} si\ L = min(X_2\text{-}X_1, Y_2\text{-}Y_1), & & & \\ Seuil_1 = & 0{,}2 & pour & L > 10 \\ Seuil_2 = & 1 & pour & 10 \geq L > 3 \\ Seuil_3 = & 3 & pour & 3 \geq L \end{array}$$

Soulignons que les points d'ancrage, même s'ils correspondent à des couples de phrases, ne désignent pas nécessairement des binômes de type (1:1). En effet, le couple de phrases correspondant à un point peut faire partie de configurations plus complexes (par exemple (3:1)). Le sous-espace de recherche encadré par (X_1,Y_1) et (X_2,Y_2) doit donc tenir compte des points qui précèdent immédiatement (X_1,Y_1) et / ou qui suivent immédiatement (X_2,Y_2). Nous avons donc élargi chaque intervalle de l'espace de recherche aux points :

$$X_1\text{-}3 \leq x \leq X_2$$
$$Y_1\text{-}3 \leq y \leq Y_2$$

II.3.7.1 Méthode GC

Dans la mise en œuvre de l'algorithme, nous considérons 8 transitions possibles :

T_1=1:1, T_2=2:1, T_3=1:2, T_4=0:1, T_5=1:0, T_6=2:2, T_7=3:1, T_8=1:3

Les fréquences empiriques de ces transitions, calculées sur la totalité du corpus aligné manuellement, sont les suivantes (les fréquences détaillées figurent dans le tableau 91 de l'annexe) :

T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	***Autres***
88,18 %	1,90 %	2,37 %	0,90 %	4,21 %	1,42 %	0,13 %	0,26 %	0,65 %

tableau 25 : *fréquences empiriques des transitions* (*corpus BAF*)

Notons que *Verne* a un profil atypique au sein du corpus, étant donné une grande proportion de contractions et d'omissions. Si l'on calcule les mêmes fréquences sans tenir compte de *Verne*, on obtient des proportions plus représentatives du reste du corpus :

T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	***Autres***
91,73 %	1,69 %	2,15 %	0,94 %	1,44 %	1,19 %	0,11 %	0,21 %	0,54 %

tableau 26 : *fréquences empiriques des transitions* (*corpus BAF sans Verne*)

Dans l'estimation des probabilités de transition *a priori*, nous avons repris les mêmes valeurs que pour les 6 transitions de la méthode GC, servant de référence :

T_1	T_2	T_3	T_4	T_5	T_6
0,89	0,0425	0,0425	0,00495	0,00495	0,011

tableau 27 : *probabilités estimées des transitions employées dans la méthode GC*

Nous avons estimé les probabilités de T_7 et T_8 à environ 1/10 des probabilités des transitions T_2 et T_3. En conservant les mêmes proportions que dans l'estimation de Gale & Church et en normalisant (pour obtenir une somme égale à 1), on obtient les valeurs *a priori* :

T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
0,883	0,0442	0,0442	0,0049	0,0049	0,01	0,0044	0,0044

tableau 28 : *probabilités estimées des transitions étendues*

Afin d'évaluer l'impact de ces deux transitions additives, nous avons testé deux jeux de transitions : T_1-T_6 avec les valeurs de référence de GC, et T_1-T_8 avec les valeurs ainsi normalisées.

En outre, nous avons implémenté l'algorithme avec deux versions différentes de l'indice, en calculant les longueurs des phrases en nombre de mots et en nombre de caractères (nous abrégerons par *NbMots* et *NbCars*).

Les résultats globaux de ces quatre expériences sont consignés dans le tableau 29 (les résultats détaillés figurent dans le tableau 92 et le tableau 93 de l'annexe) :

Modèle	*Précision*		*Rappel*		*F-mesure*	
	Moy.	**Min.**	**Moy.**	**Min.**	**Moy.**	**Min.**
T_1-T_6 (*NbMots*)	93,3 %	65,6 %	92,3 %	64,5 %	92,8 %	65,0 %
T_1-T_8 (*NbMots*)	93,3 %	62,5 %	92,6 %	63,1 %	92,9 %	62,8 %
T_1-T_6 (*NbCars*)	94,6 %	64,8 %	92,8 %	66,0 %	93,7 %	65,4 %
T_1-T_8 (*NbCars*)	94,7 %	63,7 %	93,2 %	65,6 %	94,0 %	64,7 %

tableau 29 : *comparaison des résultats avec 6 ou 8 transitions canoniques et des longueurs exprimées en nombre de mots ou en nombre de caractères* (*BAF **)

Modèle	*Précision*	*Rappel*	*F-mesure*
T_1-T_6 *(NbMots)*	96,5 %	3,9 %	7,5 %
T_1-T_8 *(NbMots)*	96,4 %	3,9 %	7,5 %
T_1-T_6 *(NbCars)*	97,0 %	3,9 %	7,5 %
T_1-T_8 *(NbCars)*	96,6 %	3,9 %	7,5 %

tableau 30 : *comparaison des résultats avec 6 ou 8 transitions canoniques et des longueurs exprimées en nombre de mots ou en nombre de caractères* (*sous-corpus Xerox*)

Il apparaît, comme le remarquent Gale & Church (1991), que la longueur exprimée en caractères est un indice plus fiable. Ceci s'interprète simplement à partir des variations empiriques des rapports des longueurs : on observe que la variance du rapport des longueurs est plus faible lorsque celles-ci sont en nombre de caractères, et que les corrélations entre les longueurs des textes traduits sont plus fortes dans ce cas (cf. tableau 94 de l'annexe) :

	Moyenne ***L(P') / L(P)***	*Variance* ***L(P') / L(P)***	*Corrélation* ***(L(P'), L(P))***
NbCars	1,119	0,123	0,973
NbMots	1,161	0,160	0,962

tableau 31 : *statistiques du rapport des longueurs en nombre de caractères et en nombre de mots*

Le meilleur comportement de *NbCars* est en partie dû aux fluctuations de la ponctuation et des espacements : une suite de dix tirets utilisés pour la présentation aboutit à dix unités, ce qui pèse plus lourd dans une phrase si sa longueur est exprimée en nombre

de mots. Or les variations de mise en forme sont importantes entre les versions françaises et anglaises du corpus.

En ce qui concerne les transitions, d'après les données empiriques, la prise en compte de T_7-T_8 permet de considérer 0,39 % des alignements de référence en plus, laissant 0,65 % des transitions effectives non intégrables par notre algorithme. L'amélioration résultante est modeste : la F-mesure globale augmente de 0,1 % pour *NbMots*, et de 0,3 % pour *NbCars*, passant de 93,7 % à 94 %.

Pour être plus précis, il semble que la progression des résultats soit dépendante de la structure des textes : globalement, l'introduction des deux transitions (3:1) et (1:3) corrige positivement les résultats, sauf pour *Verne* (et dans une moindre mesure *Xerox*) que cette modification fait chuter d'environ 1 point. Ceci est sans doute dû à l'inadéquation de la méthode GC vis-à-vis d'une traduction « libre », comportant de nombreuses contractions et omissions. En utilisant d'autres indices plus adaptés, nous verrons qu'une meilleure couverture des transitions est toujours profitable.

Afin de mieux évaluer l'incidence des probabilités estimées des transitions, nous avons aussi testé un jeu de probabilités basées sur les valeurs empiriques issues des alignements de référence (nous avons pris en compte les probabilités empiriques globales du corpus BAF, excepté *Verne* – les résultats détaillés sont dans le tableau 95 de l'annexe).

Modèle	***Précision***		***Rappel***		***F-mesure***	
Probabilités	Moy.	Min.	Moy.	Min.	Moy.	Min.
empiriques de T_1-T_8	94,6 %	64,8 %	92,9 %	65,9 %	93,7 %	65,3 %

tableau 32 : *résultats obtenus avec l'utilisation des probabilités empiriques des transitions*

Une légère dégradation des résultats montre les limites d'un modèle s'appuyant sur le seul indice des longueurs de phrase.

II.3.7.2 Intégration des cognats

Pour compenser les insuffisances de l'indice, il faut intégrer une source d'information complémentaire : la cognation peut à nouveau jouer un rôle intéressant.

II.3.7.2.1 Calcul de la distance

Comme le suggèrent Simard *et al*. ainsi que Davis *et al*. (1995), il est possible d'utiliser la cognation dans le cadre dynamique, en élaborant une mesure de distance basée sur la densité des cognats. Nous reprenons la précédente mesure de la densité, en tenant compte de la pondération 2 (du tableau 23) :

$$d_{ij} = \frac{C_{ij}}{l_i \cdot l_j} \tag{32}$$

où C_{ij} est la somme pondérée des cognats observés entre P_i et P'_j.

En calculant cette densité pour tous les couples de phrases équivalentes, d'une part, et pour des phrases quelconques, d'autre part, on détermine les probabilités empiriques d'obtenir d_{ij} inférieure à une valeur donnée. Nous avons effectué ces relevés sur les sous-corpus *Cour* et *Verne*, dont les densités de cognats sont assez différentes, respectivement d_{cognat} = 0,093 et d_{cognat} = 0,057. On observe les distributions suivantes[159] :

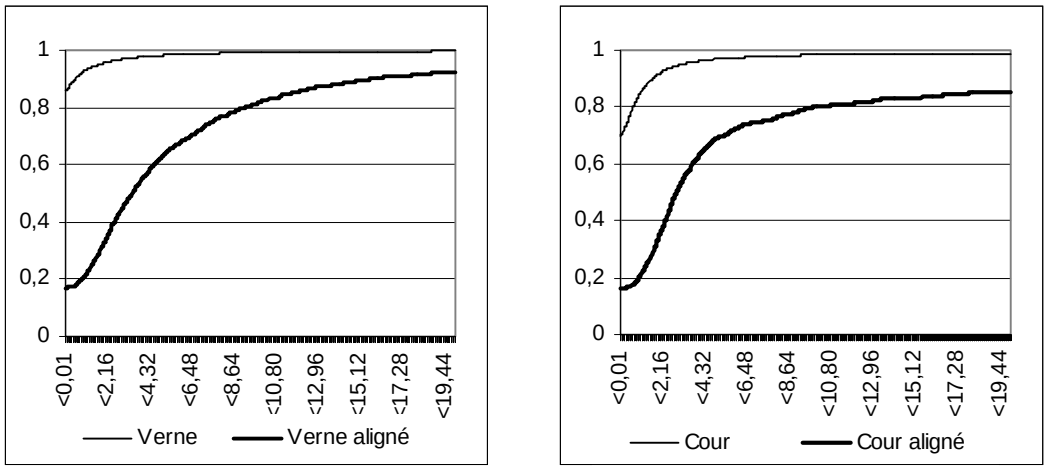


figure 27 : $p(d_{ij} \leq x)$ *pour des couples de phrases équivalentes et des couples quelconques*

[159] à un facteur multiplicatif près : c'est d_{ij} x 100 qui est représenté.

Pour chaque corpus, on peut mesurer la valeur discriminante de l'indice à la différence de surface entre les courbes pour les phrases équivalentes et les phrases quelconques (les courbes en gras correspondent aux premières). Par exemple, la probabilité d'obtenir $d_{ij} < 0{,}01$ est $p_e = 0{,}2$ pour deux phrases équivalentes et $p = 0{,}85$ entre deux phrases quelconques.

On constate que les structures de ces distributions, pour *Verne* et *Cour*, sont similaires. Dès lors, on peut supposer que ces distributions empiriques sont généralisables et peuvent servir de base à une estimation grossière, *a priori*, pour d'autres corpus. Nous avons choisi d'utiliser les distributions issues de *Cour*, dont la traduction est plus littérale, pour implémenter une mesure de distance sur la totalité du corpus BAF[160].

En reprenant les notations introduites précédemment, on calcule la probabilité *a priori* d'aligner deux segments dans un binôme B_i :

$$P(B_i / d) = \frac{p(d / B_i)\, p(B_i)}{p(d)} \tag{33}$$

d'où l'on déduit la mesure de distance suivante :

$$\mathrm{D}_{\mathrm{cognat1}} = -\log P(B_i / d) = -(\log p(d / B_i) - \log p(d) + \log p(\mathit{transition})) \tag{34}$$

où $p(d/B_i)$ et $p(d)$ correspondent aux distributions empiriques de la figure 27 correspondant à *Cour*.

[160] On pourrait déceler ici un cercle vicieux, car on utilise des données issues d'un corpus aligné manuellement pour en tirer par la suite un alignement automatique : ce dernier étant dépendant de l'alignement manuel il devient totalement inutile. Nous parlons cependant d'estimation *a priori* car nous faisons l'hypothèse que la courbe obtenue empiriquement peut être employée pour n'importe quel corpus, et c'est précisément ce que nous montrons.
En outre, il faut noter que le corpus *Cour* ne représente que 8 % de la masse totale des corpus que nous traitons. Un éventuel biais ne peut donc concerner que les résultats de *Cour* : or, si l'on néglige ceux-ci, nos résultats globaux restent pratiquement inchangés.

En effectuant la même approximation que Gale & Church, on peut estimer que les variations de $p(d)$ sont négligeables : $-log\ p(d_C)$ étant assimilé à une constante positive, on peut ignorer ce terme dans le calcul. On aboutit alors à une distance simplifiée :

$$D_{cognat2} = -\log P(B_i / d) = -(\log p(d / B_i) + \log p(transition)) \quad (35)$$

II.3.7.2.2 Résultats

Nous avons étudié $D_{cognat1}$ et $D_{cognat2}$ combinées à la distance D_{GC} établie par Gale & Church[161]. Différentes pondérations ont été évaluées.

En appliquant la formule :

$$D_{combinée} = (1- k_{co})\ D_{GC} + k_{co}\ D_{cognat} \quad (36)$$

on observe l'évolution suivante de la F-mesure globale pour les différentes valeurs de k_{co} (pour le détail des valeurs de *P*, *R* et *F* cf. tableau 96 de l'annexe) :

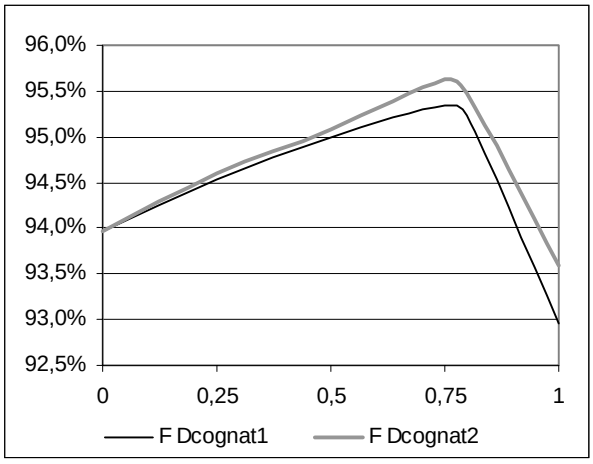


figure 28 : *évolution de F en fonction des pondérations* k_{co} *de la distance mixte* (*après lissage de la courbe établie d'après 5 points*)

[161] cf. équation (6) p. 251, calculée sur la base du nombre de caractères, avec les transitions T_1-T_8.

Il apparaît que la distance $D_{cognat2}$ donne des résultats légèrement meilleurs. L'interprétation de ce phénomène est délicate. $D_{cognat2}$ étant le résultat d'une approximation supplémentaire, cette mesure aurait dû en toute logique fonctionner moins bien. Par rapport à $D_{cognat1}$, $D_{cognat2}$ accorde une importance accrue au terme $p(d/cognat)$, qui n'est pas compensée par la soustraction de $p(d)$. Le troisième terme $p(transition)$ y intervient donc dans une moindre mesure. Or ce terme est lui même issu d'une approximation générale, puisque les probabilités de transition sont estimées *a priori*, identiques pour tous les textes du corpus. On peut alors faire l'hypothèse que le bruit engendré par ces valeurs approchées est minimisé dans $D_{cognat2}$.

Par ailleurs, la structure des deux courbes montre une amélioration constante jusqu'à $k_{co} = 0{,}75$ suivie d'une chute assez rapide au-delà. Par rapport aux deux points extrêmes, pour $k_{co} = 0$ et $k_{co} = 1$, les courbes se situent toujours au-dessus (courbes convexes) : cela confirme le caractère additif des deux sources d'information. Chaque indice pallie les insuffisances de l'autre indice : lorsque les longueurs de phrase ne peuvent permettre de discriminer, les cognats sont susceptibles de prendre le relais, et réciproquement. L'hypothèse de Simard *et al* est donc confirmée : les deux indices ne se contrarient pas, mais fonctionnent en corrélation et de façon complémentaire.

Les meilleurs résultats sont obtenus avec $D_{cognat2}$ pour k_{co} =0,75 (cf. tableau 97 de l'annexe) :

	Précision		***Rappel***		***F-mesure***	
	Moy.	**Min.**	**Moy.**	**Min.**	**Moy.**	**Min.**
*BAF **	96,5 %	70,6 %	94,8 %	72,7 %	95,6 %	71,7 %
Xerox	97,0 %		3,9 %		7,6 %	

tableau 33 : *résultats de la distance combinée avec* $k_{co} = 0{,}75$

Par rapport aux résultats précédents, rappel et précision ont été améliorés conjointement d'environ 1,6 %. L'amélioration la plus sensible concerne *Verne* : pour ce texte, *F* a augmenté de 64,7 % à 71,7 %.

Cette combinaison d'indices est donc plus fiable et produit moins de bruit, surtout vis-à-vis d'une traduction contractée comme *Verne*. On peut donc raisonnablement penser qu'une plus grande adéquation des probabilités de transition doit avoir des répercussions plus nettes et plus décisives qu'avec le seul indice GC. En appliquant les fréquences empiriques des transitions T_1-T_8, on obtient les résultats ci-dessous (cf. tableau 98 de l'annexe) :

	Précision		***Rappel***		***F-mesure***	
	Moy.	**Min.**	**Moy.**	**Min.**	**Moy.**	**Min.**
*BAF **	97,2 %	77,0 %	94,9 %	75,2 %	96,0 %	76,1 %
Xerox	97,0 %		3,9 %		7,6 %	

tableau 34 : *résultats de la distance combinée avec* k_{co} *= 0,75 et les probabilités de transitions empiriques*

Bien sûr, dans la pratique, on ne dispose pas de ces jeux de probabilités *avant* d'avoir aligné automatiquement. Mais cette légère amélioration confirme qu'une modélisation plus fine des transitions permettrait d'améliorer les résultats de façon significative : on pourrait par exemple moduler les probabilités des transitions T_2-T_n en fonction de la différence des nombres de phrases de chaque version.

Comme on pouvait s'y attendre, le corpus le plus sensible aux probabilités de transition empiriques est *Verne*, dont le profil est atypique :

Verne **(k_{co} *= 0,75*)**	***Précision***	***Rappel***	***F-mesure***
T_1-T_8 *a priori*	70,6 %	72,7 %	71,7 %
T_1-T_8 *empiriques*	77,0 %	75,2 %	76,1 %

tableau 35 : *influence des probabilités de transition empiriques sur le sous-corpus Verne*

On peut comparer les résultats obtenus jusqu'à présent[162] avec ceux du système APA (CTT/KTH, Stockholm & LIA, Avignon), qui a obtenu le meilleur classement au cours de la deuxième campagne d'évaluation du projet ARCADE. On constate que nos propres résultats sont voisins, sauf pour le corpus Verne :

[162] sans « tricher », i.e. sans les probabilités de transitions empiriques, cf. tableau 33.

Corpus	*Système APA*			*Résultats obtenus*		
	Précision	***Rappel***	***F-mesure***	***Précision***	***Rappel***	***F-mesure***
cour	99,2 %	97,0 %	98,1 %	99,8 %	97,4 %	98,6 %
hans	98,8 %	97,8 %	98,2 %	98,6 %	97,0 %	97,8 %
ilo	98,7 %	96,8 %	97,8 %	99,9 %	98,7 %	99,3 %
onu	99,4 %	99,5 %	99,5 %	99,6 %	98,9 %	99,2 %
tao1	99,7 %	98,1 %	98,9 %	99,3 %	98,1 %	98,7 %
tao2	99,4 %	99,1 %	99,2 %	99,6 %	98,8 %	99,2 %
tao3	99,6 %	98,0 %	98,8 %	99,6 %	96,0 %	97,8 %
citi1	99,6 %	94,9 %	97,2 %	98,9 %	93,5 %	96,1 %
citi2	99,1 %	94,5 %	96,7 %	99,4 %	96,4 %	97,9 %
verne	90,4 %	93,0 %	91,7 %	70,6 %	72,7 %	71,7 %
xerox	95,6 %	3,9 %	7,5 %	97,0 %	3,9 %	7,6 %
Moyenne	98,1 %	88,4 %	89,4 %	96,6 %	86,5 %	87,6 %
Moyenne (/Verne)	98,9 %	88,0 %	89,2 %	99,2 %	87,9 %	89,2 %

tableau 36 : *comparaison des résultats avec ceux du système APA*

La spécificité du corpus Verne, ainsi que l'importante marge de progression, nous invitent à lui réserver un traitement particulier.

II.3.7.2.3 Le corpus Verne

Etant donné les irrégularités du chemin d'alignement propre à ce corpus, il est possible que l'espace de recherche précédemment défini (p. 320) ne soit pas assez large pour intégrer toutes les sinuosités du chemin. Nous avons donc testé la combinaison d'indices avec le paramètre $Seuil_1 = 0{,}5$ au lieu de $Seuil_1 = 0{,}2$, ce qui autorise une plus forte déviation.

Verne (k_{co} ***= 0,75***)	***Précision***	***Rappel***	***F-mesure***	***dF***
T_1-T_8 *a priori*	71,1 %	73,0 %	72,1 %	+0,5 %
T_1-T_8 *a posteriori*	78,7 %	76,8 %	77,7 %	+1,6 %

tableau 37 : *résultats de la distance combinée 2 sur le corpus Verne avec un élargissement de l'espace de calcul*

L'amélioration, même légère, des résultats confirme notre hypothèse. Avec $Seuil_1 = 0{,}5$ le chemin correct est, probablement, presque entièrement compris dans

l'espace de recherche. Notons qu'aucune autre amélioration n'est constatée pour des valeurs plus élevées de $Seuil_1$. Par ailleurs, nous avons cherché à développer un nouvel indice plus adapté à une traduction du type *Verne*. Notre idée est d'évaluer, entre deux phrases, la moyenne des similitudes entre les distributions des mots en langue source et cible. Pour une unité *u* en langue source, et une unité *u*' en langue cible, nous avons mesuré la similitude de leurs distributions en calculant l'information apportée par l'une sur l'autre. Mais plutôt que de recourir à l'information mutuelle, nous avons préféré utiliser une valeur toujours positive (car deux phrases en relation de traduction peuvent contenir des mots qui ont un comportement mutuellement inhibiteur). Nous avons donc calculé le lien statistique (déjà évoqué dans l'exploitation des cognats) :

$$sim(u_i, u_j') = \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} \tag{37}$$

où f_i est la fréquence de u_i, f_j la fréquence de u_j', f_{ij} la fréquence de cooccurrence de u_i avec u_j'. entre deux phrases *P* et *P*', on calcule donc la moyenne des similitudes pour tous les couples de mots comparés :

$$sim(P, P') = \frac{\sum_{i \in P_1, j \in P_2} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j}}{l \cdot l'} \tag{38}$$

où *l* et *l*' sont les longueurs respectives de *P*, *P*' en nombre de mots.

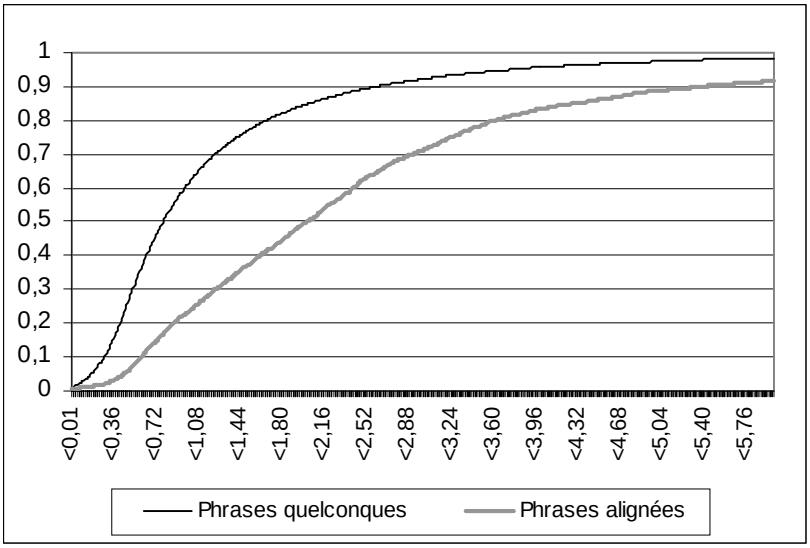


figure 29 : *distributions comparées de l'indice distributionnel entre des phrases quelconques et des phrases équivalentes* (*corpus Verne*)

Le pouvoir discriminant de cet indice peut-être mesuré par la différence de surface sous les deux courbes de la figure 29, représentant respectivement les distributions de l'indice pour des phrases quelconques, et pour des phrases équivalentes.

Comme précédemment nous avons testé différentes pondérations pour les indices, en conservant la pondération relative des deux premiers (à savoir 0,25 - 0,75).

$$D_{combinée2} = (1- k_{co}-k_{dis})\ DGC + k_{co}\ D_{cognat} + k_{dis}\ D_{distrib}$$

Les meilleurs résultats ont été obtenus avec (cf. tableau 38) :

$1- k_{co}-k_{dis} = 0,17$ $\quad k_{co} = 0,5$ $\quad k_{dis}= 0,33$

	Précision		***Rappel***		***F-mesure***	
T_1-T_8 *a priori*	**Moy.**	**Min.**	**Moy.**	**Min.**	**Moy.**	**Min.**
	96,4 %	70,1 %	95,0 %	76,2 %	95,7 %	73,1 %

tableau 38 : *résultats avec une distance combinée intégrant un indice distributionnel* (*BAF **)

L'amélioration (cf. tableau 33) est mineure au regard du coût des calculs supplémentaires : on constate néanmoins une légère progression de 1 % au niveau du corpus *Verne* (qui correspond aux minima du tableau 38).

Le bruit inhérent à cet indice ne permet sans doute pas d'aller plus loin. Nous pensons néanmoins que l'information extraite à partir des distributions lexicales peut enrichir l'alignement de manière déterminante : la simple cooccurrence de toutes les unités d'une phrase avec toutes les unités d'une autre étant une donnée trop peu significative, il nous faudra vraisemblablement filtrer cette information afin d'obtenir de véritables correspondances lexicales. A partir de ces correspondances, nous verrons comment créer un indice plus précis (cf. infra, § III.3.10)

II.4 Problèmes et perspectives

Ces premiers résultats expérimentaux nous confirment que le parallélisme est bien une propriété formelle des textes en relation d'équivalence traductionnelle : des indices aussi superficiels que l'occurrence de chaînes de caractères similaires, ou les longueurs de phrases, peuvent fournir une information assez riche pour aboutir, dans de nombreux cas, à l'alignement d'unités textuelles aussi fines que les phrases.

Comme nous l'avons montré, la prise en compte de ces indices doit suivre un certain ordre : l'alignement final étant obtenu par raffinements successifs, il faut d'abord exploiter les points d'ancrage les plus fiables, susceptibles de dégager de larges îlots de confiance. Pour garantir cette fiabilité, nous avons énuméré un faisceau de *contraintes* : certaines sont basées sur le renforcement mutuel des points (redondance, monotonie, continuité), d'autres sur des présomptions quant au passage du chemin (diagonalité). Le principe de précision d'abord impose une utilisation régressive de ces conditions : les premiers points d'ancrage, qui mettent en correspondance des zones de grandes dimensions, doivent présenter le maximum de garantie, et par conséquent remplir toutes les conditions requises avec la plus grande rigueur ; mais pour les points suivants, il est possible de relâcher progressivement ces contraintes en assouplissant les paramètres (seuil de déviation, redondance, etc.).

Fiabilité décroissante : + → -

Indices	***Contraintes***	***Complexité***
Transfuges 1. numériques 2. initiale en majuscule 3. quelconque	redondance diagonalité continuité monotonie	$O(n\ log(n))$
Densité de cognat	diagonalité continuité monotonie	$O(n^2)$
Longueur des phrases + densité de cognats	diagonalité monotonie	$O(n^2)$

Coût croissant : - → +

tableau 39 : *hiérarchie des indices, des contraintes et des coûts*

Parallèlement, à mesure que l'espace de recherche se rétrécit, il devient possible de mettre en œuvre des méthodes plus coûteuses en calcul et plus aléatoires, comme les modèles de variation du rapport des longueurs. Le tableau 39 récapitule la hiérarchie des indices, des contraintes et du coût des méthodes.

En appliquant cet ordre de priorité, nous sommes parvenu à des résultats satisfaisants sur tous les textes du corpus BAF. Pour la plupart des alignements obtenus, dont la F-mesure dépasse 97 %, il est difficile de progresser encore, puisqu'on atteint pratiquement le niveau de l'alignement manuel, qui en outre n'est pas une référence absolue (les deux annotateurs optaient parfois pour des solutions différentes). Il faut par ailleurs tenir compte des erreurs dues à la segmentation (qui a été estimée avec une précision et un rappel voisins de 97,5 %, cf. J. Véronis & P. Langlais, 2000) et des cas de transition utilisés dans l'alignement de référence mais non pris en compte par nos algorithmes (p. ex. (2:3), etc.).

II.4.1 Limites

Restent deux sous-corpus marginaux : *Xerox* et *Verne*. Pour ce dernier, nous examinerons plus loin (cf. p. 518) des méthodes d'extraction de correspondances lexicales permettant plus de finesse dans la détermination des transitions.

Pour Xerox, le problème est très clair : toutes nos méthodes présupposent une définition étroite du parallélisme, imposant la monotonie des appariements. Or, ce texte contient un glossaire dont les deux versions sont ordonnées alphabétiquement, sans respect de la monotonie. Pour traiter cette difficulté il faut donc se doter de méthodes alternatives, ne s'appuyant pas sur la séquentialité. Cela implique la résolution de deux sous problèmes :

- détecter les zones non monotones, afin de les traiter séparément des portions parallèles.
- développer des méthodes d'alignement pour ces zones, sans recourir à l'ordre des segments.

II.4.1.1 Détection des zones problématiques

Pour la détection automatique des zones non monotones, il faut mettre à profit les résultats obtenus dès les premières phases du processus d'alignement : comme le remarque I. D. Melamed (1996 : 7), toute zone non monotone se manifeste par un « trou », une lacune dans la continuité des points d'ancrage fournis par les méthodes de préalignement (avant l'extraction du chemin complet).

Ce type de lacune s'observe assez clairement avec le sous-corpus *Xerox*, comme le montrent les points d'ancrage issus de la cognation : ces points sont très denses dans tout le texte, mais se raréfient cependant dans la zone correspondant au glossaire (zone encadrée).

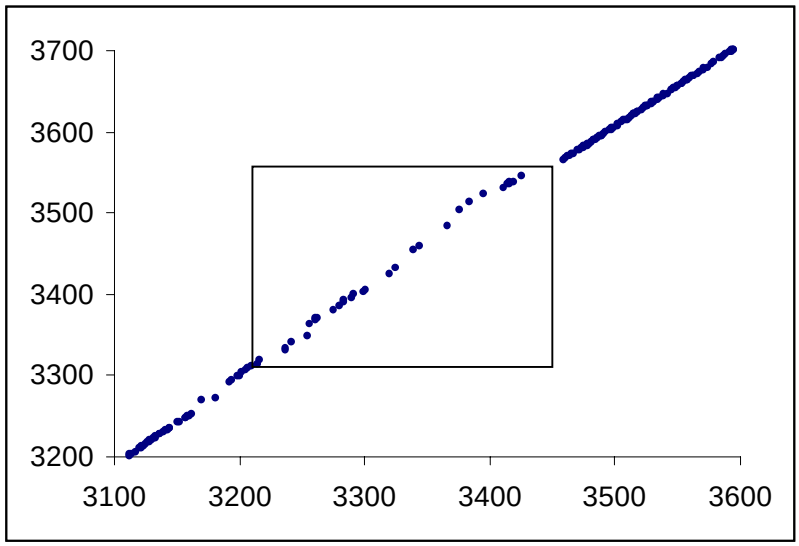


figure 30 : *dispersion des points d'ancrage dans la zone de non-monotonie* (*sous-corpus Xerox*)

En fonction du nombre total de points d'ancrage obtenus, on peut fixer une densité minimale pour des groupes de points consécutifs : lorsqu'un groupe de points est trop dispersé, le rectangle encadrant ce groupe peut être traité à part comme zone de non-monotonie.

Par exemple, pour la totalité du corpus *Xerox*, la moyenne des écarts entre deux points consécutifs, en abscisse et en ordonnée, est de $Ecart_{Xerox}$ = 1,749 phrases environ.

Calculée pour la seule zone de non-monotonie, cette moyenne passe à *Ecart* = 6,569, soit près de 4 fois plus.

Pour k points (X_1,Y_1), (X_2,Y_2),..,(X_k,Y_k) la moyenne des écarts s'écrit :

$$Ecart = \frac{1}{k-1}\sum_{i=2..k}\frac{(X_i - X_{i-1}) + (Y_i - Y_{i-1})}{2} \qquad (39)$$

Cet indicateur peut donc servir de base à la détection des zones posant problème : pour chaque point d'ancrage, nous avons calculé la moyenne des écarts à l'intérieur d'une fenêtre de f points consécutifs à partir de ce point ; lorsque cette valeur dépasse un certain seuil s, le point marque le début d'une zone, qui se termine lorsque au moins p points consécutifs atteignent à nouveau une valeur inférieure au seuil fixé.

En appliquant cette méthode au corpus, avec $f = 10$, $p = 4$ et $s = 2{,}5 \cdot Ecart_{Xerox}$, on obtient les zones marquées en clair sur la figure 31 :

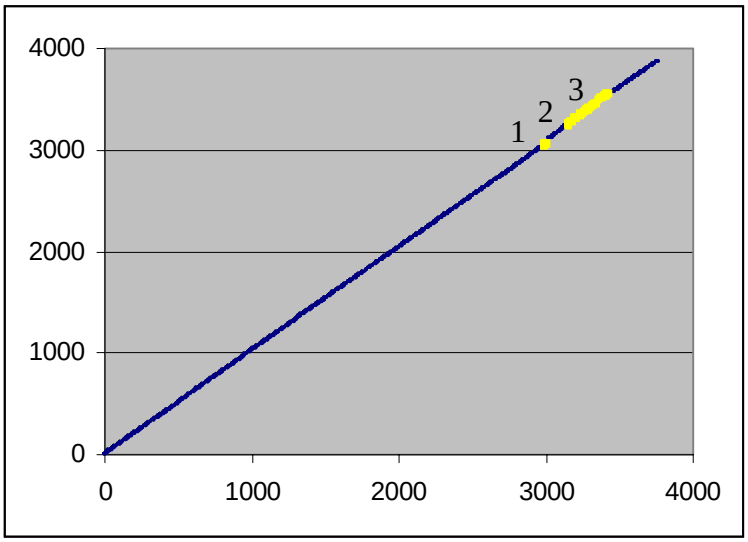


figure 31 : *zones détectées supposées problématiques*

Trois zones rapprochées ont été détectées. Elles correspondent respectivement à :

1. l'insertion en français des phrases de 3 048 à 3 079, sans correspondants en anglais ;
2. l'insertion en français des phrases de 3 254 à 3 263, sans correspondants en anglais ;
3. la zone de non monotonie correspondant au glossaire.

Notons que pour la détection automatique des insertions (ou omissions), une fenêtre de largeur 2 suffirait.

Globalement, le succès d'une telle méthode dépend du choix des paramètres :

– plus le seuil *s* est bas, et plus la détection devient sensible. Par exemple avec $s = 2 \cdot Ecart_{Xerox}$ on passe de 50 à 82 points suspects, répartis sur 5 zones supplémentaires ; avec $s = 4 \cdot Ecart_{Xerox}$ on ne détecte plus que 12 points suspects, tous situés dans le glossaire.

– plus la largeur *f* de la fenêtre est grande, moins la détection devient sensible. Mais combinée à un abaissement du seuil *s*, l'augmentation de *f* a tendance à « diluer » les zones détectées avant et après les écarts les plus importants.

– l'augmentation de *p* permet de renforcer la continuité des zones détectées, en évitant leur fragmentation en plusieurs petites zones.

Le niveau de sensibilité de la détection dépend de la largeur des zones obtenues et du coût des techniques complémentaires qui peuvent s'affranchir de la condition de monotonie. Il n'est pas utile de définir tous ces paramètres *a priori*, puisqu'ils dépendent des contraintes spécifiques des corpus et de la précision recherchée, mais on peut concevoir un algorithme itératif affinant ces paramètres d'une itération à l'autre, afin d'obtenir des zones problématiques de taille raisonnable pour la mise en œuvre de techniques plus coûteuses.

Enfin, si la rupture de monotonie est due à l'interversion de deux zones contiguës, on obtient une configuration originale facilement repérable (Melamed, 1996 : 7), représentée figure 32.

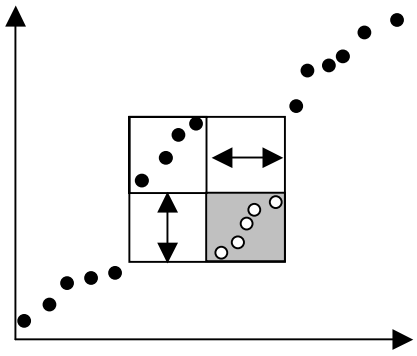


figure 32 : *interversion de zones contiguës*

Dans ce cas, la zone requérant un traitement spécifique n'est pas délimitée par *une* rupture de continuité, mais par *deux* ruptures consécutives, concernant successivement les deux axes, de manière séparée : la rupture est d'abord verticale (texte cible), puis horizontale (texte source). Avec une telle configuration, les points appartenant à la zone intervertie *après* (points en blanc) sont automatiquement éliminés par l'application des contraintes de monotonie. Pour récupérer ces points, il faut simplement examiner la zone définie à l'intersection des deux bandes ignorées (zone grise).

II.4.1.2 Méthodes d'alignement non monotones

Fluhr, Bisson & Elkateb (in Véronis, 2000 §9) proposent d'appliquer les méthodes de recherche d'information multilingue (en angl. CLIR, *Cross Language Information Retrieval*) à l'alignement : pour les auteurs, ces deux tâches constituent des problèmes similaires puisqu'« elles demandent toutes deux le calcul d'une valeur de similarité entre deux textes écrits dans des langues différentes »[163] Une phrase à aligner est donc traitée comme une requête : sur la base d'un dictionnaire bilingue, on examine toutes les traductions possibles des mots qui la constituent, après quoi on recherche la phrase cible

[163] "both require the computation of a proximity value between two texts that are in different languages"

qui réalise la meilleure adéquation avec la liste de mots obtenus – de la même manière qu'on chercherait le document qui coïncide le mieux avec la liste des concepts exprimés par la requête. Dans le système mis en œuvre les relations de dépendance entre les mots sont prises en compte, ce qui permet de diminuer le bruit. La seule différence entre les deux approches est, d'après les auteurs, que « si les dictionnaires bilingues sont exhaustifs, et que deux phrases sont des traductions mutuelles, l'intersection sera égale à chacune des phrases »[164] – tandis qu'en recherche d'information, l'intersection entre la requête et les documents pertinents ne peut concerner qu'une partie du document. En fait, nous pensons que sauf dans le cas très particulier d'une traduction mot à mot, il est normal que l'intersection des phrases soit partielle, quelle que soit la taille du dictionnaire.

Dans le système d'alignement ainsi mis en œuvre, on cherche d'abord les appariements (1:1) des phrases séparément, puis les alignements (1:2) et (2:1) sont recherchés dans le voisinage des couples obtenus. Les résultats de cette méthode sont bons : la F-mesure moyenne est d'environ 87 % sur le BAF et de 93 % sur le corpus Xerox (y compris le glossaire). Ils prouvent que le recours à un dictionnaire de transfert assez complet permet de résoudre les configurations d'alignement qui échappent aux méthodes superficielles : il faut donc y voir une approche complémentaire des méthodes étudiées jusqu'à présent.

II.4.2 Perspectives

II.4.2.1 Alignements multilingues

Lorsque les indices disponibles sont insuffisants pour obtenir un alignement satisfaisant, il est nécessaire de recourir à une source d'information extérieure. Pour pallier le déficit de similarité entre deux textes, Simard (in Véronis, 2000 §3) met en œuvre une ingénieuse méthode d'alignement trilingue, dans le cas où une autre traduction, dans une autre langue, serait disponible. L'auteur remarque que ce type d'alignement n'est pas une fin en soi, car la plupart des utilisations des corpus multilingues parallèles concernent des

[164] "if the bilingual dictionaries are exhaustive, and the two sentences are translations of each other the intersection will be equal to each of the sentences."

bi-textes : dans le contexte de l'aide à la traduction, par exemple, on manipule rarement plus de deux langues en même temps. Mais il peut être néanmoins intéressant d'aligner simultanément des corpus multilingues, dans la mesure où ils apportent un surcroît d'information, qui peut être utile à la constitution des bi-textes. L'idée est que plusieurs traductions contiennent toujours plus d'information qu'une seule, et que cette information peut être partagée et redistribuée sur tous les couples de langues pris deux à deux. Il s'agit simplement d'appliquer une méthode de *triangulation* permettant le renforcement des hypothèses établies sur chaque bi-texte indépendamment.

L'auteur en apporte l'illustration sur un corpus trilingue constitué d'une version anglaise, française et espagnole de l'évangile selon Jean. Le principe de la méthode est le suivant :

1. Les trois textes *A*, *B* et *C* sont d'abord alignés deux à deux en utilisant une technique classique combinant cognation et longueur des phrases.
2. Les alignements X_{AB}, X_{BC} et X_{AC} obtenus sont segmentés en sous-sections correspondantes, sur la base de leur concordance. Cette étape se base sur le principe de transitivité : un alignement de (A,C) peut être déduit des alignements X_{AB} et X_{BC}, en appliquant une simple règle de saturation transitive (« *transitive closure* ») : si P_A est alignée avec P_B et P_B avec P_C alors, P_A est alignée avec P_C. Les trois alignements X_{AB} et X_{BC} et X_{AC} sont concordants sur l'alignement des trois phrases (P_A,P_B,P_C) si la saturation transitive des trois alignements donne le même résultat que les alignements pris deux à deux. A la concordance, on rajoute deux autres conditions, afin de s'assurer de la validité des points d'ancrage ainsi dégagés : ne sont retenus que les triplets (P_A, P_B, P_C) correspondant à une transition 1:1 pour les trois alignements, et pour lesquels la mesure de distance ne dépasse pas un certain seuil.
3. Pour chaque groupe de sous-sections correspondantes (A_i, B_i, C_i), on identifie le couple de langues pour lequel l'alignement deux à deux a obtenu le meilleur score (i.e. la distance minimale) : cet alignement est probablement plus fiable que les deux autres. La sous-section restante (A_i, B_i ou C_i) est ensuite alignée avec les couples issus de cet alignement (respectivement X_{BiCi}, X_{AiCi}, ou X_{AiBi}). Cette

dernière opération nécessite une légère adaptation de l'algorithme d'alignement : il ne s'agit plus d'examiner des appariements entre segments, mais entre un couple de segments déjà alignés et un segment (p. ex. (S_A,S_B) avec S_C). Pour ce faire, la mesure de distance est calculée comme la somme des distances du segment avec chaque membre du couple (p. ex. distance((S_A,S_C) + distance(S_B,S_C)). On obtient finalement un alignement trilingue (une suite de triplets (S_A, S_B, S_C)).

4. A chaque fois qu'un tel triplet met en jeu des segments incluant plusieurs phrases, on évalue la possibilité de scinder ces segments afin d'obtenir un alignement plus fin ; enfin, on supprime les phrases dont l'élimination aboutit à une augmentation du score de l'appariement entre deux langues.

Simard justifie les étapes de cette technique par les hypothèses et constats suivants :

– On pourrait tenter d'implémenter directement l'alignement simultané des trois textes, en généralisant l'algorithme de programmation dynamique utilisé dans le cas bilingue. Mais l'espace de recherche serait en O(n^3), ce qui s'avère vite rédhibitoire. Dans la méthode présentée, on n'effectue que des alignements deux à deux, la complexité est donc seulement multipliée par une constante.

– On fait l'hypothèse que les carences d'informations (longueurs, cognats, etc.) ne se situent pas aux même endroits pour les trois couples de textes. Lorsque les trois alignements séparés convergent vers les mêmes points d'ancrage, et que ceux-ci présentent certaines garanties, on peut en tirer un pré-découpage très fiable en sous-sections correspondantes : c'est le principe de l'étape 2.

– Par suite, au niveau de chaque sous-section, on élit le couple de langues L_1L_2 qui présente l'alignement *a priori* le plus sûr : on suppose que celui-ci a mieux profité de l'information disponible. En alignant la sous-section de la troisième langue L_3 avec cet alignement, présumé plus fiable, on espère faire bénéficier le nouvel alignement (qui additionne le « vote » des appariements de L_1L_3 et L_2L_3) de cette meilleure information.

– Dans ce type d'appariement triangulaire, Simard constate que la transitivité aboutit parfois à des regroupements trop larges. Par exemple, si une phrase en L_1 est alignée avec trois phrases en L_2, la ou les phrases de L_3 alignées avec ce couple seront automatiquement alignées avec les trois phrases de L_2, alors que certaines sont à considérer comme des insertions (car il se peut que la version de L_3 soit plus « compacte »). D'où une diminution de la précision. L'étape 4 de la méthode vise à réduire ces correspondances qui débordent (la transitivité des appariements n'est alors plus respectée).

La mise en œuvre de cette méthode permet une augmentation d'environ 1 % des résultats globaux, par rapport aux résultats obtenus en alignant deux par deux. La triangulation semble donc effectivement apporter une meilleure assise aux hypothèses portant sur les transitions. Comme le souligne l'auteur, ces expérimentations demandent à être étendues à des corpus plus grands, aux caractéristiques variées, et concernant plus de trois langues.

En outre, ces travaux dégagent des pistes prometteuses concernant l'alignement multilingue au niveau des unités lexicales, tâche considérée plus difficile et pour laquelle des améliorations sont encore à attendre.

II.4.2.2 Cas des langues non-apparentées

La plupart des méthodes présentées reposent sur l'observation des ressemblances entre les deux versions parallèles : on peut se demander s'il est possible de généraliser les méthodes étudiées à des couples de langues où la cognation, voire les longueurs ne sont pas utilisables.

Dans une certaine mesure, les travaux de Y. Choueka, E. Conley & I. Dagan (in Véronis 2000 §4) montrent qu'on peut répondre par l'affirmative : l'alignement est envisageable même entre des langues non apparentées présentant des difficultés spécifiques comme l'hébreu et l'anglais.

En effet, un certain nombre d'hypothèses ne tiennent plus dans le couple anglais / hébreu : par exemple, d'après les auteurs, on ne peut plus faire l'hypothèse que les

frontières de phrase sont détectables de manière fiable et que le découpage des phrases est similaire entre les deux langues. Par ailleurs les alphabets sont différents, et on ne peut plus s'appuyer sur les cognats sans le recours à un système de translitération *ad hoc*. En outre l'hébreu implique des difficultés spécifiques : il n'y a pas de différence majuscule / minuscule, et pas de vocalisation, d'où de très nombreuses ambiguïtés liées à l'homographie ; la morphologie de l'hébreu est très riche : les 35 000 lemmes constituant un dictionnaire tel que le *Rav-Milim* aboutissent à un nombre estimé à 70 millions de formes fléchies ; enfin la syntaxe est libre. D'où la nécessité d'opérer un pré-traitement spécifique afin de désambiguïser et de lemmatiser les formes.

Malgré ces obstacles, les auteurs montrent que des techniques basées sur les distributions peuvent suffire : l'application de la méthode DKvec, qui a par ailleurs fait ses preuves sur un corpus anglais / chinois (Fung & Mc Keown, 1994, cf. infra, p. 407) permet d'obtenir un alignement satisfaisant[165] en appariant des mots sur la base de leur vecteur distributionnel. Par suite en appliquant un algorithme inspiré du modèle 2 de Brown *et al*. (1993, cf. § III.2.2.3.3), les auteurs obtiennent des correspondances lexicales assez satisfaisantes (52 % de précision pour un rappel d'environ 35 %).

[165] Un échantillon de points de référence ayant été établi manuellement, 90 % de ces points sont situés à une distance de ± 50 unités de l'alignement obtenu.

II.5 Conclusion de la deuxième partie

L'étude de l'alignement a démontré un fait important : que la relation traductionnelle entre les deux textes n'implique pas toujours la correspondance d'unité à unité, qu'il s'agisse des phrases, du nombre des mots qu'elles contiennent, ou des unités susceptibles de « résister » à la traduction, comme les nombres et les noms propres.

Et pourtant, même si on ne peut jamais présager du sort réservé à telle ou telle unité, on peut rassembler de nombreux faisceaux d'indices qui dessinent globalement, au-delà du flou des écarts locaux, la *forme* de la relation bi-textuelle : dans la très grande majorité des cas, ces indices permettent d'inférer la relation d'équivalence au niveau des phrases (ou des groupes de phrases).

Dans la partie suivante nous tenterons d'aller un peu plus loin : par un nouveau réglage de l'outil statistique, nous essaierons d'ajuster un peu mieux notre focale, afin d'obtenir une image plus nette ; par la mise en œuvre d'autres techniques, centrées sur le comportement des mots, nous chercherons à diminuer le grain de nos observations au niveau lexical : ces réglages permettront peut-être la mise en évidence objective de phénomènes de transcodage généraux, à travers la masse de traductions particulières.

Partie III

Les correspondances lexicales

« Le plus grand chef-d'œuvre de la littérature n'est jamais qu'un dictionnaire en désordre. »

Jean Cocteau, *Le Potomak*.

« Le langage est chimie pour le sens et physique pour les formes. Il est chimie, car il se crée, à partir d'un nombre restreint d'éléments linguistiques, un nombre infini de combinaisons à signification nouvelles ; cependant les éléments qui entrent en combinaison pour donner une signification nouvelle ne perdent pas leur identité formelle comme c'est le cas des éléments d'un composé chimique, et la forme du langage est donc pour l'essentiel physique. »

Danica Seleskovitch, *Langue*, *Langage et Mémoire*, *1975* : 49-50

III L'extraction de correspondances lexicales

Dans une belle métaphore, Seleskovitch compare l'élaboration du sens à une transformation chimique : c'est que, dit-elle, les éléments combinés perdent « leur identité formelle » en se mêlant dans la substance résultante. Dans cette opération de synthèse, le sens n'est pas réductible à la somme de ses constituants, et puisque la traduction opère au niveau du sens, des constituants différents dans chaque langue peuvent fusionner en un alliage équivalent.

Les observations de Seleskovitch sur la prise de note en traduction consécutive révèlent deux types de comportements lexicaux : certains mots fusionnent et perdent leur identité au sein du produit final, d'autres subsistent et gardent leur identité formelle (Seleskovitch compare ces derniers à des raisins dans une brioche, qui résistent à la cuisson) :

> « En étudiant non seulement l'interprétation proposée par ses collègues mais également les notes de consécutive qu'ils avaient prises, Seleskovitch constate que certains mots du discours original sont notés et traduits par les participants. Ce sont les chiffres, les appellations, les énumérations et les termes techniques. Par contre d'autres mots, qui possèdent ce qu'elle avait appelé dans *L'interprète dans les conférences internationales* des équivalents conventionnels dans l'autre langue, n'avaient été ni notés ni traduits tels quels. Fondus dans l'opération chimie du sens, ils avaient fait l'objet d'une réexpression. »(Laplace, 1994 : 239)

Ces « raisins dans la brioche », que nous appellerons des correspondances lexicales, ne sont pas sans rappeler les points d'ancrage utilisés dans les techniques d'alignement, à la différence près que la conservation de l'identité formelle n'implique pas nécessairement la ressemblance superficielle, mais une identité de comportement sémantique. Ce qui semble définir ces associations de mots, c'est leur indépendance par rapport aux contingences d'une situation particulière, comme le remarque Seleskovitch (citée par Laplace, 1994 : 239) :

> « Ce n'est donc pas l'existence d'une équivalence dans une autre langue qui amène l'interprète à prendre le mot en notes, mais la conscience que le mot entendu a une personnalité propre, indépendante du message ; sa signification pertinente se dégage grâce au contexte mais, à part cela, il passe du niveau de la langue à celui de la parole, sans prendre de sens autre que celui que lui confère le code. »

La persistance de certaines équivalences lexicales est bien entendu le reflet des équivalences au niveau de la désignation. Pergnier (1993 : 113) remarque que les mots dont la traduction est invariable en fonction du contexte sont monosémiques :

> « Les seuls mots de la langue qui échappent à ce traitement sont ceux qui, par leur nature ou leur usage, sont parfaitement monosémiques et qui, par conséquent, ne sont pas soumis à la structuration par le système de signification. Parmi ceux-là, on peut citer les nombres, les noms propres et les mots appartenant exclusivement à un domaine technique précis. Leur monosémie en fait de purs symboles, pour lesquels désignation et signification se confondent »

Mais rien ne prouve que les mots dont on retrouve les équivalents dans le produit final de la traduction soient toujours les mêmes : en d'autres termes, il serait hâtif de distinguer deux classes de mots ayant des propriétés distinctes, comme on opposerait le liant de la pâte et les raisins dans la brioche, à partir d'oppositions telles que polysémiques / monosémiques, mots vides / mots pleins, ou syncatégorématiques / catégorématiques, etc.

Ce qui se dessine à travers les correspondances lexicales, pour une traduction donnée, c'est l'émergence de points fixes, de pivots, de mots assumant un rôle central dans l'économie de la relation d'équivalence traductionnelle en cours de construction. Ces mots constituent en quelque sorte la charpente stable de la relation d'équivalence, puisqu'ils sont équivalents au niveau d'une traduction particulière, mais *aussi* au niveau des codes.

Nous avons vu, en étudiant les techniques d'alignement, qu'il existait des méthodes permettant d'extraire automatiquement ces correspondances lexicales (Kay & Röscheisen, 1988, 1993 ; Fung & Church, 1994). De nombreux travaux se sont concentrés sur ce type de tâche (Church & Hovy, 1993 ; Dagan, Church & Gale, 1993 ; Fung & Wu, 1994 ; Wu & Xia, 1994 ; Gaussier & Langé, 1995 ; Smajda, Mc Keown & Hatzivassiloglou ; Véronis, 1997 ; Resnik & Melamed, 1997 ; Melamed, 1998a) : dans cette dernière partie nous nous proposons d'explorer le détail de ces techniques, et d'examiner jusqu'à quel niveau de granularité il est possible de développer le concept de parallélisme.

Cependant, il serait illusoire de vouloir limiter le problème de l'extraction des correspondances lexicales à des questions de statistiques et d'algorithmes. La notion de

correspondance lexicale soulève une nouvelle fois le problème de l'opposition entre traduction et transcodage. Prenons l'exemple suivant, cité par Kay (in Véronis, 2000)

angl. : *Gravity is a pervasive force in the world…* (*Scientific American*)
fr. : *La pesanteur s'exerce partout sur la terre…* (*Pour la science*)

Certes, dans ces deux versions, on peut lier sémantiquement l'anglais *pervasive* avec *partout*. Mais la question se pose : ces deux unités sont-elles équivalentes au niveau des codes, y a-t-il entre elles l'« identité formelle » des raisins de Seleskovitch ? Kay (in Jean Véronis, 2000 : xiv) pose le problème en ces termes :

> « Pour un chercheur intéressé par la traduction de grande qualité, un programme d'alignement qui apparierait *pervasive force*, ou seulement *pervasive*, avec *partout*, pourrait ouvrir d'intéressantes perspectives, mais comme source d'entrée potentielle dans un dictionnaire bilingue, cela pourrait constituer une source de frustration. »[166]

Du point de vue du dictionnaire bilingue, censé enregistrer des équivalences au niveau des *codes*, une telle correspondance serait peu pertinente. L'exemple donné par Kay montre qu'avant d'envisager la possibilité d'automatiser, il est essentiel de circonscrire la notion même de correspondance lexicale.

Les problèmes soulevés sont nombreux :

- Toutes les unités du texte source ont-elles une correspondance dans la cible, conformément au principe de quasi-bijection ?

- Sur quel plan faut-il situer la relation d'équivalence ? est-il légitime de se limiter aux seules correspondances susceptibles de figurer dans un dictionnaire bilingue ? Ou bien faut-il élargir la notion à des correspondances issues d'un contexte particulier, comme dans le cas de *pervasive* et *partout*.

[166] "For a researcher interested in high-quality translation, an alignment program that paired pervasive force, or at least pervasive, with partout (everywhere) might stimulate important insights, but as a source of potential entries in a bilingual dictionary, it might constitute a source of frustration."

– Quelles sont les unités concernées par ce type de relation : les mots, les composés, les syntagmes ? L'épineux problème de la définition linguistique de l'unité lexicale n'est-il pas rendu plus aigu par la prise en compte simultanée de deux systèmes linguistiques différents ?

Si certaines propriétés formelles du bi-texte rendent possible le recours à des outils d'extraction automatique, l'éclaircissement de ces problèmes est un préalable indispensable à l'interprétation de ces propriétés.

III.1 Le concept de correspondance

Traditionnellement (P. Brown *et al*. 1990 ; Langé & Gaussier 1995 : 71 ; Véronis 1997 : 193), la notion de correspondance lexicale est présentée comme une forme d'alignement au niveau lexical : ce genre de correspondance est défini par le lien sous-jacent à des mots ou groupes de mots en relation de traduction mutuelle à l'intérieur de deux segments alignés. P. Brown *et al* (1993 : 267) donnent l'exemple suivant, tirés de leur corpus :

angl. : *The poor don't have any money*
fr. : *Les pauvres sont démunis*

A l'intérieur de ces deux phrases les auteurs cherchent à établir des correspondances au niveau des mots (« *word level* »). Ils en extraient l'alignement suivant :

A={(*The* ; *Les*) (*poor* ; *pauvres*) (*don't have any money* ; *sont démunis*)}

don't have any money et *sont démunis* étant alignés en tant que groupe de mots.

Cet exemple soulève le problème des unités de segmentation : il n'est pas toujours possible d'aligner au niveau désiré, en l'occurrence les mots, car on n'obtient pas toujours de correspondance terme à terme.

Au cours du projet ARCADE (Véronis, 1997), visant à définir un cadre rigoureux pour l'évaluation des correspondances lexicales extraites automatiquement, un certain nombre de problèmes ont été énumérés. L'extraction d'alignements lexicaux complets

ayant été jugée hors de portée des systèmes actuels, l'évaluation s'est orientée vers un sous-problème considéré plus réaliste : le repérage de traduction (« *translation spotting* »). « Etant donné un mot ou une expression particulière dans le texte source, cela consiste à détecter sa traduction dans le texte cible »[167] (Véronis & Langlais, in Véronis, 2000 § 19).

Pour la mise en œuvre d'une telle évaluation, il faut établir manuellement l'alignement de référence d'un certain nombre de mots-test destinés à fournir un étalon de référence. Il apparaît que même le sous-problème du « *translation spotting* » soulève des difficultés. Certaines divergences traductionnelles imposent d'adopter une définition élargie de l'unité lexicale, car la relation d'équivalence engage très souvent des expressions polylexicales complexes. Dans un guide destiné aux annotateurs du corpus, une typologie de ces divergences traductionnelles a été établie, afin d'harmoniser les solutions choisies dans l'alignement manuel (les unités alignées sont en gras) :

– Phraséologie (« *Phrasal correspondences* »)

fr. : *les étudiants **qui ont de petits moyens***
angl.: ***less well off** students*

– Omissions

fr. : *le **petit** lac Prespa*
angl.: *the lake of Prespa*

– Changement de partie du discours (« *Change in part-of-speech* »)

fr. : *La Commission a fait la **proposition** suivante*
angl.: *The Commission **proposed** the following*

– Changement morphologique (« *Change in number, tense, mood* »)

fr. : *La Communauté européenne **apporte** une aide*
angl.: *The European Community **has been providing** assistance*

– Expression discontinue (« *Discontinuity* »)

[167] "Given a particular word or expression in the source text, it consists in detecting its translation in the target text"

fr. : *Quel soutien la Commission* ***apporte****-t-elle*
angl. : *What support* ***is*** *the Commission* ***giving***

– Anaphore (« *Referring expressions* »)

fr. : *les* ***cartes*** *sont en libre circulation*
angl. : *they are freely available*

– Dissymétrie dans la parataxe (« *Non parallel conjuncts* »)

fr. : ***cartes de crédit*** *ou* ***de paiement***
angl. : ***credit-card*** *or* ***pay-card***

– Divergence traductionnelle (« *Divergent translations* »)

fr. : (…) ***apporte*** *des informations importantes*
angl. : (…) *is a piece of information vital to* (…)

On retrouve les phénomènes de divergence étudiés dans la première partie. La présente typologie ne nous satisfait pas complètement, dans la mesure où différents plans y sont imbriqués. Nous préférons systématiser ces problèmes de façon plus synthétique, en distinguant deux niveaux spécifiques : les problèmes de segmentation et les problèmes de divergence sémantique.

III.1.1 Problèmes de segmentation

Dans la pratique, on peut dire que les alignements lexicaux entrelacent les niveaux : une unité peut correspondre à un groupe d'unités, une phrase entière, voire plusieurs phrases. Qu'on soit confronté à des unités composées ou discontinues, le problème est le même : il s'agit d'identifier des unités possédant un certain degré *d'autonomie*. Or cette autonomie n'est jamais que relative. Dans toute construction linguistique, de même que le tout est déterminé par ses parties, le tout détermine ses parties. Comme le décrit Rastier, l'interprétation d'un texte réalise un va-et-vient entre des structures globales, « macrosémantiques » (isotopies, molécules sémiques, anaphores) et contenus locaux « microsémantiques » (sèmes inhérents), suivant un jeu dialectique d'assimilation et de dissimilation. A tout moment l'autonomie d'une lexie peut s'altérer dans la cristallisation d'une expression complexe portant un sens nouveau. Toute unité peut signifier en tant

qu'elle-même ou bien en tant que partie plus ou moins dépendante d'une construction plus large. Comme on l'a vu précédemment (§ I.1.3.2), le degré d'autonomie d'une unité par rapport aux unités englobantes (syntagmes, phrases, etc.) décrit un véritable continuum.

Dans le cas des correspondances, le problème de la segmentation interne de chaque texte est compliqué par le besoin d'isomorphisme entre ces deux segmentations. Supposons que l'on donne une autre traduction de l'exemple précédent ; on aboutit alors à une segmentation différente :

> *angl.* : *The poor don't have any money*
> fr. : *Les pauvres n'ont pas d'argent*
> A={(*The* ; *Les*) (*poor* ; *pauvres*) (*don't have* ; *n'ont pas*) (*any* ; *d'*) (*money* ; *argent*)}

A l'intérieur d'une même langue, l'identification d'unités composées dépendait d'un faisceau de contraintes multiples : linguistiques, idiomatiques ou pragmatiques (par exemple dans le cas des pluritermes). Mais dans le rapport entre deux textes, les unités de segmentation dépendent d'un autre ordre de contrainte, se croisant avec le précédent : il s'agit de la *compositionnalité traductionnelle* des deux portions de texte alignées. Cette fois la segmentation n'est plus déterminée par l'autonomie syntactico-sémantique des unités alignées, mais par l'autonomie relative, vis-à-vis de la traduction, de ces unités. Pour désigner la segmentation qui en découle, nous parlerons de *segmentation traductionnelle*. D'après Véronis (in Véronis, 2000 §1), le traitement modulaire et autonome des tâches de segmentation et d'appariement est théoriquement possible, pour un couple de langues donné :

> « L'alignement ou l'extraction de lexiques peut, au moins en principe, être décomposé en deux phases : (1) détecter les mots et les expressions dans les textes sources et cibles, et (2) apparier les uns avec les autres. En pratique, ces deux tâches ne peuvent être rendues complètement modulaires, parce que les unités mises en œuvre dans la langue source dépendent de la langue cible (par exemple, le français *demande de brevet* devrait être considéré comme un tout

> dans un alignement avec l'allemand *Patentanmeldung*, tandis que les mots peuvent être alignés un par un avec l'italien *domanda di brevetto*). »[168]

Ce que l'auteur met ici en évidence, en se situant au niveau des codes, ce sont les *unités de traduction*. Or ce type de segmentation n'est pas équivalent à la segmentation traductionnelle, qui dépend du degré de parallélisme d'une traduction *particulière*. Ce dernier niveau de segmentation fournit des syntagmes non reconnaissables comme unités autonomes dans une langue donnée, comme le syntagme « *don't have any money* » de l'exemple précédent. Puisque la segmentation traductionnelle est étroitement liée à la compositionnalité, elle est soumise, de manière contingente, aux caractéristiques particulières de la traduction liées aux choix de traduction, aux habitudes et à l'idiosyncrasie du traducteur, etc.

Réciproquement, la segmentation traductionnelle peut aboutir à la décomposition de certains syntagmes, alors qu'ils constituent des entités lexicales autonomes au niveau des codes. Ainsi, dans l'exemple suivant, les expressions figées peuvent être alignées de la sorte :

angl. : *To be the very devil*
fr. : *Avoir le diable au corps*
it. : *Avere il diavolo in corpo*
Français / Italien : A ={(*Avere* ; *Avoir*) (*il* ; *le*) (*diavolo* ; *diable*) (*in*; *au*) (*corpo* ; *corps*)}
Anglais / Français : A = {(*To be the very devil* ; *Avoir le diable au corps*)}

L'unité du phrasème, conservée entre l'anglais et le français, est brisée entre l'italien et le français, dans la mesure où pour des raisons historiques les deux expressions sont composées de la même manière dans les deux langues.

---

[168] "The alignment or extraction of lexicons can be broken down into two phases, at least theoretically : (1) detect the words and expressions in the source and target texts, and (2) map them to each other. In practice, these two tasks cannot be fully modularised because the units to use in the source language are dependent upon the target language (for example, the French *demande de brevet* should be taken as a single chunk in an alignment with the German *Patentanmeldung*, whereas the words can be aligned one by one with Italian *domanda di brevetto*)."

III.1.2 Problèmes de divergences sémantiques

Outre les distorsions syntaxiques précédemment évoquées, le concept de correspondance se heurte au problème des divergences sémantiques. Considérons ces deux phrases issues de notre corpus :

> fr. : (…) *les différentes politiques mises en œuvre pour permettre l'accès des personnes handicapées à l'emploi*.
> angl. : (…) *the various policies for access to employment for disabled people.*

La tournure anglaise est beaucoup plus elliptique qu'en français. On a le choix entre deux alignements possibles :

(*for* ; *mises en œuvre pour permettre*)

ou bien, en considérant deux insertions en français :

(*for* ; *pour*) (∅ ; *mises en œuvre*) (∅ ;*permettre*)

On est confronté au même problème avec les phénomènes d'implicitation et d'explicitation déjà évoqués :

> fr. : *TGV*
> angl. : *TGV* (*French high speed train*)

Ces divergences découlent des différentes manières d'aborder un même arrière plan extralinguistique. Prenons l'exemple suivant, issu de notre corpus :

> fr. : (…) *6400 signatures, sur l'émission de billets de banque identifiables par les aveugles et par les personnes à vision réduite.*
> angl.: (…) *6400 signatures, on the marking of banknote for the benefit of the blind and partially sighted*.

La même réalité est présentée suivant deux points de vue légèrement différents. Sur le plan linguistique, on ne peut dire qu'il y ait identité de contenu sémantique. Comme on l'a vu dans la première partie, le traducteur peut choisir de privilégier différents niveaux d'équivalence (cf. § I.1.2.3) : équivalence dynamique, dénotative, connotative, etc. Dans la mesure où la traduction fait toujours une traversée dans le contexte extralinguistique, ce genre de distorsion sémantique est courant, même en domaine spécialisé. Les niveaux

conceptuels et référentiels étant primordiaux dans les domaines technico-scientifiques, une reformulation est toujours possible de la part d'un traducteur s'attachant à conserver les concepts au détriment de la forme linguistique.

Là encore les possibilités de distorsion sémantique entre un énoncé et sa traduction décrivent un continuum allant de l'identité à la différence totale. Jacqueline Henry (in Lederer & Israël, 1991 : 115) illustre cette possibilité de recréation avec la traduction suivante d'un slogan publicitaire relatif au golf :

fr. : *Pour faire putt de velours*
angl.: *It will make your greens come true*

Ici le contenu sémantique a disparu au profit de la conservation du procédé (en l'occurrence un calembour).

Dès lors qu'on admet divers degrés de divergence, un nouveau problème surgit : à partir de quel éloignement sémantique doit-on décider de ne pas faire correspondre deux segments ?

Ajoutons que les distorsions sémantiques se combinent avec les difficultés de segmentation précédemment décrites, en les aggravant, car la segmentation s'effectue sur une base sémantique, puisqu'elle dépend de la compositionnalité traductionnelle. Considérons la traduction suivante, issue du corpus JOC :

fr. : *Pour la bonne tenue de ces registres, l'évaluation des cas de mortalité constatés par les autorités apporte des informations importantes*.
angl. : *The assessment of the official cause of death is a piece of information vital to these registers*.

Doit-on aligner *vital* avec *importantes* (…) *pour la bonne tenue de ces registres* ?

Faut-il privilégier les alignements suivants :

(*cause of death* ; *cas de mortalité*) et (*official* ; *constatés par les autorités*)

ou bien l'alignement moins fin mais plus respectueux du sémantisme global ? :

(*official cause of death* ; *cas de mortalité constatés par les autorités*)

On constate que les deux types de problèmes dégagés, l'identification des segments et leurs fluctuations sémantiques, sont étroitement imbriqués. La complexité et l'extrême intrication des phénomènes mis en jeu expliquent sans doute pourquoi, dans la littérature, la nature problématique de la notion de correspondance est le plus souvent passée sous silence.

Par exemple, Julian Kupiec (1993 : 17) présente les correspondances comme un lien de « représentation » entre séquences : « Une séquence de mot en Ei est ici définie par la correspondance d'une autre séquence en Fi, si on considère que les mots de celle là représentent les mots celle-ci. »[169] Dans d'autres modèles, on définit les correspondances en termes génératifs : « Brown *et al*. (1990) introduisent l'idée d'un alignement entre paires de phrases comme un objet indiquant, pour chaque mot de la phrase française, le mot anglais duquel il provient. » (P. Brown *et al*., 1993 : 266)[170] Ce processus de génération à partir du texte source ne concerne pas les mots séparément, mais par groupe : « Ici, les quatre mots anglais *don't have any money* fonctionnent ensemble pour générer les deux mots français *sont démunis*. » (ibid.)[171] La description formelle du processus de génération est clairement définie dans le modèle proposé par les auteurs, mais les problèmes soulevés par le concept de correspondance n'y sont pas abordés. La raison en est simple : quels que soient les formalismes employés, la plupart des travaux se situent dans le paradigme du transcodage – or l'extraction des correspondances concerne des corpus de messages, dont l'équivalence transcende les déterminations linguistiques.

[169] "A word sequence in Ei is defined here as the correspondence of another sequence in Fi if the words of one sequence are considered to represent the words in the other"
[170] "Brown et al. (1990) introduce the idea of an alignment between a pair of strings as an object indicating for each word in the French string that word in the English string from which it arose"
[171] "Here, the four English words don't have any money work together to generate the two French words *sont démunis*."

III.1.3 Correspondances *vs* alignement maximal

Puisque l'équivalence traductionnelle ne se réduit pas à l'équivalence linguistique, rien n'autorise à supposer que la compositionnalité traductionnelle s'étende au niveau des unités lexicales[172].

Comme le montrent les exemples précédents, et ainsi que le notent Langé & Gaussier (1995 : 76), « quasi-bijection et quasi-synchronisation, ne sont pas vérifiées lorsqu'on arrive au niveau des syntagmes ou des mots ». Il y aurait donc contradiction à parler d'alignement lexical. Fathi Débili (1997 : 200), qui préfère employer le terme d'« appariement » lexical, remarque à son tour qu'« il ne peut être que partiel. Il n'est ni biunivoque, ni séquentiel, ni compact. Les correspondances sont floues et contextuelles. » Mais Débili emploie également le terme d'appariement pour le niveau phrastique, au risque de maintenir une certaine ambiguïté. Dans la mesure où l'alignement est basé sur le principe de la compositionnalité traductionnelle, nous ferons désormais la distinction entre *correspondances lexicales* et *alignement*.

III.1.3.1 Correspondances lexicales

Il nous faut donc trouver d'autres critères que la compositionnalité pour redéfinir le concept de correspondance lexicale. En première approximation, on peut ramener le problème à des données simples, à partir des deux critères suivants :

1. les unités appariées sont des unités lexicales ;
2. ces unités doivent être en relation d'équivalence traductionnelle.

Le premier critère implique que le problème de segmentation soit cantonné au plan des codes linguistiques : la segmentation n'est plus une conséquence de la compositionnalité traductionnelle, mais elle doit être décidée en amont, à partir de critères

[172] De même, rien ne nous permet d'affirmer *a priori* qu'il y a compositionnalité traductionnelle au niveau des phrases : mais pour nous la phrase n'est qu'un étalon de segmentation, sans consistance linguistique, à partir duquel on peut effectuer toute sorte de regroupements pour satisfaire à la compositionnalité : en fait, il serait plus correct d'employer l'expression « alignement au niveau des phrases » plutôt que « alignement des phrases ».

monolingues (et contrastifs si on étend la notion d'unité à celle d'unité de traduction). Le second critère était déjà présent dans la définition de l'alignement.

Ces critères laissent encore une grande part d'indétermination, mais il est tout à fait possible de les détailler, suivant les élargissements ou les restrictions qu'on apporte à la définition des définitions des unités lexicales ou de l'équivalence. Ces spécifications concernent principalement les trois axes suivants :

– *Types d'unités* : *lexies, unités de traduction*

On peut s'intéresser aux unités lexicales dans un sens restreint, c'est-à-dire entendues comme des lexèmes de la langue, ou dans un sens élargi, englobant l'ensemble des lexies au sens de Mel'čuk, Clas & Polguère (1995 : 57). Dans ce dernier cas, les unités lexicales peuvent recouvrir en outre les unités phraséologiques ou « phrasèmes complets », les collocations ou « semi-phrasèmes », voire les unités à caractère idiomatique (dans le sens défini dans la première partie) ou « quasi-phrasèmes » (1995 : 46).

Même ramené à un problème monolingue, la définition de l'unité lexicale est loin d'être évidente, tant il est vrai qu'en langue les unités ne sont jamais données, mais déduites du système dans son entier. Ainsi que nous le rappelle Ducrot (1968 : 50) « la segmentation n'est justifiable que si l'expression peut être classée à l'intérieur de différents groupes, et chaque unité de l'expression doit son individualité au seul fait qu'elle est le représentant d'un de ces groupes. » C'est une des grandes intuitions de Saussure, et toute la linguistique structurale repose sur la recherche des unités en tant qu'elles sont déductibles du système.

Comme nous l'avons montré dans le chapitre dédié à l'étude des phénomènes contrastifs, il peut être intéressant d'intégrer aussi les unités de traduction qui découlent des divergences entre les langues, comme les idiotismes. Au problème des unités lexicales s'ajoute donc un deuxième niveau de structuration, faisant intervenir deux codes et non plus un seul.

Ainsi, notre définition ne résout pas, loin s'en faut, le problème de la segmentation : la détermination des unités lexicales, et plus généralement des unités de traduction, constitue un champ encore ouvert et problématique. Mais elle permet de détacher le

problème de la segmentation des contingences de la compositionnalité sous-jacente à deux textes parallèles, car ces deux ordres de fait demeurent hétérogènes l'un à l'autre.

Par delà les lexies, il est aussi possible d'étendre les correspondances lexicales aux pluritermes, dont l'unité n'est pas définie linguistiquement mais sous l'effet de contraintes extralinguistiques. Dans la mesure où les unités sont définies de manière cohérente *en dehors* de la compositionnalité, rien n'interdit ensuite de les mettre en correspondance lorsqu'elles réalisent la condition d'équivalence.

– *Niveau d'équivalence* : *dynamique, conceptuel, référentiel, expressif, sémantique*

Nous avons dégagé précédemment la possibilité de situer l'équivalence traductionnelle à différents niveaux interprétatifs : niveau dynamique, dénotatif et connotatif. Lorsque l'équivalence est valide hors situation, au niveau des idiomes, nous parlons d'équivalence linguistique.

Ces niveaux, non exclusifs et souvent entrelacés, forment en quelque sorte un *continuum* dans l'identification du sens, allant des déterminations pragmatiques et extralinguistiques au plan linguistique strict.

Si l'on reste dans le cadre élargi de l'équivalence traductionnelle, les correspondances lexicales peuvent donc se situer à chacun de ces plans. Lorsque le traducteur de *l'Isola del giorno prima* de Eco choisit de traduire *polipi soriani* par *polypes ocellés*, l'identification de *soriani* (en français *tigré*) avec *ocellé* n'est justifiée que sur le plan du style, l'accumulation des qualificatifs rares visant à créer une figure d'hypotypose. En se plaçant sur ce plan, on peut considérer ces deux unités comme équivalentes.

Mais le problème s'en trouve légèrement déplacé : de telles correspondances ont-elles un quelconque sens une fois sorties de leur contexte ? Rappelons que dans la perspective de l'aide à la traduction, l'extraction des correspondances lexicales n'a d'intérêt que dans leur possible réutilisation. Il paraît donc pertinent de s'intéresser à leur degré de généralité : c'est là un nouveau *continuum* qui se dessine, corrélé aux niveaux d'équivalence, mais non réductible à celui-ci.

– *L'axe de la dépendance contextuelle*

Il est clair que l'équivalence sémantique confère un grand degré de généralité aux correspondances lexicales, dans la mesure où celle-ci est définie au niveau des deux codes linguistiques. A l'inverse, les équivalences référentielles, stylistiques et dynamiques sont rarement consistantes à l'extérieur d'un contexte très spécifique, comme dans l'exemple précédent.

On pourrait penser que les niveaux d'équivalence et l'axe de la dépendance contextuelle sont étroitement corrélés. Mais il n'en est rien : d'une part la polysémie confère aux équivalences sémantiques (codées en langue) une certaine dépendance contextuelle, les recouvrements sémantiques étant parfois très partiels ; et d'autre part on peut trouver des équivalences à portée générale pour tous les types de correspondance :

– conceptuelle :

fr. : *réalité humaine* (traduction de Heidegger par Henri Corbin)
all. : *dasein*

– référentielle :

fr. : *le vainqueur d'Austerlitz*
angl. : *the defeated of Waterloo*

– stylistique :

fr. : *trois petites truites cuites…*
it : *trentatré trentini entrarono in Trenta trottolando*

– dynamique :

fr. : *qu'est-ce que tu as eu à Noël* ?
it. : *che cosa ti ha portato la befana ?*

fr. : *alexandrin*
it. : *endecasillabo*

Signalant les difficultés inhérentes aux effets contextuels, Débili (1997 : 203) propose de distinguer deux types de correspondances : les « correspondances lexicales », pouvant être attestées par un dictionnaire bilingue, et les « correspondances contextuelles » dépendantes d'une « recomposition locale et contextuelle, fondée sur une " compréhension

humaine " des deux phrases ». Mais cette distinction nous paraît quelque peu binaire et simplificatrice, tout d'abord parce qu'elle assimile les niveaux d'équivalence à la dépendance contextuelle, en opposant le niveau sémantique aux liens contextuels, et ensuite parce que l'attestation d'un dictionnaire reste malgré tout un critère arbitraire si l'on ne précise pas *comment* doit être construit ce dictionnaire (comme dans le *DEC* de Mel'čuk, Clas & Polguère, 1995). D'ailleurs, si l'on considère l'extraction des correspondances comme un outil d'investigation empirique destiné aux lexicographes, le fait de se baser sur un dictionnaire préétabli peut sembler quelque peu circulaire, à moins d'admettre que ce dernier ne fournirait que des informations partielles à enrichir.

III.1.3.1.1 Caractéristiques formelles

L'hypothèse de quasi-bijection étant caduque au niveau lexical, nous récusons donc le terme *d'alignement lexical* (en anglais « *word alignment* ») couramment usitée dans la littérature (Brown, 1993 ; Kupiec, 1993 ; Dagan, Church & Gale 1993), et dont l'application serait limitée à des traductions mot à mot. A la différence d'un alignement, un ensemble de correspondances lexicales peut être *fragmentaire* : suivant le degré de compositionnalité des portions de textes concernées, un certain nombre d'unités lexicales peuvent se retrouver sans correspondances.

Par exemple, si l'on considère la traduction suivante, déjà citée :

> angl.: (…) *the marking of banknote for the benefit of the blind and partially sighted*
> fr. : (…) *l'émission de billets de banque identifiables par les aveugles et par les personnes à vision réduite*

on peut tirer trois correspondances lexicales (en négligeant les mots grammaticaux tels que « de », « par », etc.) :

> (*banknote* ; *billets de banque*) (*blind* ; *aveugles*) (*partially sighted* ; *personnes à vision réduite*)

Les unités non appariées doivent être considérées comme un résidu découlant des divergences structurelles imputables à la traduction, et non comme des ajouts ou des

omissions à valeur secondaire (comme ce le serait dans le cas d'un alignement). En fait, ce résidu n'est pas un simple bruit, car il est révélateur des phénomènes contrastifs intéressants, comme le note Diana Santos (in Véronis, 2000 §8) :

> « Plutôt que de rechercher des règles fiables ou des correspondances, de considérer les données qui s'éloignent de ces normes comme du bruit résiduel (Church & Gale, 1991), ou encore de rejeter la création stylistique hors du champ de la sémantique (Dyvi, 1998), je pense que toutes les paires de traductions – incluant les simples erreurs et les réécritures complètes – mettent en lumière les systèmes des deux langues. En effet, le plus souvent, les erreurs de traduction sont liées aux difficultés mêmes qui découlent des différences entre les langues (…) »[173]

Mais au-delà des phénomènes contrastifs, le résidu est la conséquence naturelle de la relation d'équivalence, dont le lieu est situé en dehors des textes source et cible.

Toujours sur le plan formel des correspondances, notons qu'une même lexie peut rentrer dans plusieurs appariements simultanément, comme dans l'exemple suivant[174] :

> fr. : (…) *concernant le règlement relatif aux nouveaux aliments.*
> angl.: (…) *in connection with the regulation on novel foods and novel food ingredients*.

Ici, le texte anglais en dit plus que le texte français. Si l'on considère *food ingredients* comme un terme, *aliments* peut être inséré dans deux correspondances lexicales simultanées :

(*foods* ; *aliments*) (*food ingredients* ; *aliments*)

Le tableau 40 résume les caractéristiques distinctives des concepts de correspondance et d'alignement :

---

[173] "Instead of looking for reliable rules or correspondences and considering data which deviates from that norm to be residual noise (Church & Gale, 1991) or rejecting creative language as not relevant for semantics (Dyvi, 1998), I believe that every translation pair – including plain errors et complete rewriting – sheds light on the systems of the two languages. In fact, more often than not, translation mistakes are correlated with actual difficulties stemming from language differences (''garden path translations'')"

Correspondances lexicales	***Alignement***
Critère de segmentation	
L'unité découle de critères généraux, monolingues et contrastifs : vis-à-vis de ses équivalents possibles dans la langue d'arrivée, le segment source doit se comporter comme une seule unité.	La granularité de la segmentation est variable, et elle est limitée par le grain de la compositionnalité traductionnelle : elle dépend des caractéristiques singulières de la traduction.
Caractéristiques formelles	
Relations de plusieurs à plusieurs (une même unité peut entrer dans plusieurs correspondances). Toutes les unités n'ont pas de correspondant.	Quasi-bijection. Quasi-monotonie au-delà de la phrase.
Nature des unités appariées	
Unités de traduction : lexies au sens large (lexèmes, phrasèmes, tournures idiomatiques, collocations), pluritermes.	Pas de pertinence syntaxique précise : mot, groupe de mots, syntagme, phrase, groupe de phrase.

tableau 40 : *correspondance* vs *alignement*

III.1.3.2 Alignement maximal

Il nous semble que la notion de correspondance lexicale est souvent assimilée à ce qu'on pourrait appeler un *alignement à résolution maximale*, c'est-à-dire situé au niveau des plus petites unités satisfaisant à la compositionnalité traductionnelle. Pour un tel alignement, on peut supposer l'hypothèse de quasi-monotonie inapplicable pour le niveau subphrastique, et valide au-delà. L'hypothèse de quasi-bijection reste valide.

On peut illustrer la notion d'alignement maximal à partir de l'exemple suivant:

[174] Source Europarl, réf. A4-006, introduction.

fr. : *Récolte de données à caractère personnel par les services secrets d ' un Etat membre sur les candidats aux concours organisés par les institutions européennes*
angl. : *Confidential secret service information on applicants for European civil service post*

Comme alignement maximal de ces deux phrases, nous proposons :

A={(*Confidential* ; *à caractère personnel*) (*secret* ; *secrets*) (*service* : *par les services*) (∅ ; *d'un Etat membre*) (*information* ; *Récolte de données*) (*on* ; *sur*) (*applicants* ; *les candidats*) (*for European civil service post* ; *aux concours organisés par les institutions européennes*)}

Tandis qu'un ensemble de correspondances lexicales se limite à :

C={(*confidential* ; *personnel*) (*secret service* ; *services secrets*) (*information* ; *données*) (*on* ; *sur*) (*applicant* ; *candidat*) (*European* ; *européennes*)}

La *maximalité* impose que les segments ne doivent pas être décomposables en segments plus petits, alignables à leur tour. Mais ces deux critères, compositionnalité et maximalité n'éliminent pas, et de loin, toutes les indéterminations.

Par exemple dans le cas de *vining peas* et de *petits pois* [175], il est difficile de trancher entre les deux alternatives suivantes :

– soit on considère que ces deux lexies ne peuvent être décomposées, dans la mesure où les parties n'en sont pas alignables, car *vining* ne correspond pas à *petits* ;

– Soit on aligne *pois* à *peas,* en estimant que *petits* et *vining* sont alignés avec l'ensemble vide, puisque le flou de la notion de *quasi-* bijection le permet.

Considérons l'exemple déjà cité p. 362. Une application de notre définition nous amène à proposer l'alignement maximal suivant :

(*the marking of banknote for the benefit of*; *l'émission de billets de banque identifiables par*) (*the blind*; *les aveugles*) (*and* ; *et par*) (*partially sighted* ; *les personnes à vision réduite*)

[175] exemple tiré des discussions relatives au projet ARCADE.

Cette décomposition est artificielle et discutable à plusieurs niveaux : on pourrait segmenter encore au niveau des déterminants (*les* ↔ *the*) ; on pourrait aligner *billets de banque* avec *banknote*, en laissant des « trous » dans les deux premiers segments (les segments peuvent être discontinus). Mais à mesure qu'on décompose, l'autonomie du sens des segments obtenus a tendance à s'étioler, du fait des effets contextuels.

Ainsi, comme on l'a vu précédemment, les divergences, tant sur le plan structurel que sémantique, connaissent une continuité de degrés incompatible avec le caractère discret des découpages et des décompositions. Les effets de seuil sont inévitables. De surcroît au niveau subphrastique, il est à craindre que ce genre d'indétermination se multiplie de façon drastique, dans la mesure où tous les découpages sont permis.

Le découpage en phrase est certes relativement arbitraire, mais il permet de donner un grain assez large et constant pour limiter les fluctuations syntaxiques. Si le seul critère permettant de déduire une décomposition est sémantique (dans le cas où l'équivalence recherchée serait à ce niveau), les effets de seuil risquent de devenir totalement incontrôlables : comment décider où commence l'identité et où se termine la différence ?

III.1.3.3 Test de commutation et méthode de Nida

Une solution à ce problème a été esquissée par M.-D. Mahimon (1999 : 34). Citant Catford (1965 : 28) : « une traduction textuelle est […] cette partie d'un TS [texte source] qui subit une modification si et seulement si une partie donnée du TC [texte cible] est elle-même modifiée. », Mahimon propose de généraliser ce principe pour établir un test de commutation entre unités lexicales des deux langues. La même idée est à l'œuvre dans la méthode proposée par Malavazos *et al*. (2000, cf. l'exemple p. 204), qui repose sur l'observation suivante :

> « L'idée principale est basée sur le constat qu'étant donné une paire de phrases source et cible, toute modification de la phrase source aboutira probablement en un ou plusieurs changements dans la phrase cible, et qu'il est en outre probable que les unités constantes et variables de la phrase source correspondent respectivement aux unités constantes et variables de la phrase cible. » [176]

[176] "The main idea is based on the observation that given any source and target language sentence pair, any alteration of the source sentence will most likely result in one or more changes in the respective target sentence, while it is also highly likely that constant and variable units of the source sentence correspond to constant and variable target units respectively."

Dans le test de commutation classique, la forme linguistique est déduite de la commutation parallèle des unités sur le plan du contenu et sur le plan de l'expression : on s'intéresse à la différence sémantique engendrée par une différence sur le plan de l'expression.

Ici, la commutation est redoublée et parcourt les deux sens :

1. expression → contenu : en faisant commuter une unité linguistique du texte source on fait commuter des éléments de sens ;
2. contenu → expression : partant de cette différence sémantique, on essaye de rétablir l'équivalence traductionnelle en faisant commuter une (ou plusieurs) unité(s) du texte cible.

Les unités du texte source et du texte cible qui ont commuté parallèlement, lors de ces deux phases, sont alors mises en correspondance. Notons que les transformations 1 et 2 présupposent que l'équivalence traductionnelle soit située au niveau sémantique.

Mahimon (ibid. :37) donne un premier exemple de commutation :

fr. : *Ce projet de loi prévoira un système de déclaration des maladies infectieuses*
angl. : *This bill will provide for an infectious disease notification system*

Si l'on fait commuter *Ce* avec *Chaque*, on peut rétablir l'équivalence en faisant commuter *This* avec *Each* :

fr. : *Chaque projet de loi prévoira un système de déclaration des maladies infectieuses*
angl. : *Each bill will provide for an infectious disease notification system*

Ainsi, on peut établir la correspondance de *Ce* avec *This*.

Pour désigner une telle commutation, nous noterons : *Ce* || *This*

III.1.3.3.1 Principes formels de la commutation

Afin de préciser les différentes configurations de la commutation, Mahimon énonce un certain nombre de principes correspondant aux deux critères de *maximalité* et de *compositionnalité* déjà cités. Nous systématisons ces principes de la manière suivante :

- *Principe de commutation minimale* : « quand cela est possible, la commutation doit affecter un seul mot (ou morphème) à la fois. » (1999 : 36) En d'autres termes, les parties qui commutent doivent être aussi petites que possible, dans le respect du principe suivant.

- *Principe de transitivité*. Les unités qui commutent ensemble d'un même côté doivent former des classes d'équivalence. On peut donc définir une seconde relation, noté ≡ et définie par :

$$\exists\, U' \,/\, U_1 \parallel U' \text{ et } U_2 \parallel U' \;\Leftrightarrow U_1 \equiv U_2$$
$$\exists\, U \,/\, U \parallel U_1' \text{ et } U \parallel U_2' \;\Leftrightarrow U'_1 \equiv U'_2$$

La saturation des classes d'équivalence est obtenue par l'application de la transitivité :

$\forall\, (U_1, U_2, U_3) \in S^3$ (ou S'^3) si $U_1 \equiv U_2$ et $U_2 \equiv U_3$ alors $U_1 \equiv U_3$

De la sorte, la relation ≡ vérifie bien les trois propriétés définissant une relation d'équivalence (au sens mathématique) : réflexivité, commutativité et transitivité. Notons que cette relation se situe au niveau du découpage des unités dans chaque langue et non à celui de l'équivalence traductionnelle, comme chez Simard (in Véronis, 2000 §3).

Pour illustrer ces principes (systématisés de manière légèrement différente), Mahimon donne les illustrations suivantes, sur la base de l'exemple déjà fourni :

> fr. : *Ce projet de loi **prévoira** / **entérinera** un système de déclaration des maladies infectieuses*
> angl. : *This bill will **provide for** / **confirm** an infectious disease notification system*

d'où : *prévoira* || *provide for* (1)

On a aussi :

fr. : *Ce projet de loi* ***prévoira*** / ***prévoit*** *un système de déclaration des maladies infectieuses*
angl. : *This bill* ***will provide*** / ***provides*** *for an infectious disease notification system*

par conséquent : *prévoira* || *will provide* (2)

L'application du principe de transitivité nous donne :

(1) + (2) ⇒ *prévoira* || *will provide for*

Autre exemple emprunté à Mahimon (1999 : 43) :

fr. : […] *Les membres de nos* ***services*** / ***écoles*** *de police* […]
angl. : […] *members of our police* ***forces*** / ***academy*** […]
⇒ *services* || *forces* (*1*)

[…] *Les membres de nos services de* ***police*** / ***surveillance*** […]
[…] *members of our* ***police forces*** / ***surveillance personnel*** […]
⇒ *police* || *police forces* (*2*)

d'où (1) + (2) ⇒ *services* (*de*) *police* || *police forces*

La commutation apparaît comme une technique adaptée à la problématique bi-textuelle : les unités dégagées correspondent exactement à la notion de compositionnalité. En outre la commutation a l'avantage d'être d'une grande simplicité formelle. Mahimon note avec justesse que les « unités de commutation » (1999 : 49) ainsi dégagées permettent d'éviter les tests linguistiques complexes (tels que ceux de Gross, 1996) destinés à identifier les « unités polylexicales » dans le cas monolingue : « la prise en compte des unités polylexicales dans une perspective bi-linguistique peut être grandement simplifiée » (1999 : 52)

III.1.3.3.2 Limites de l'application du test

Cependant l'auteur pose quelques limites à l'application du test de commutation « pour des raisons d'ordre syntaxique » et dans le cas de « traduction libre » (1999 : 53).

Elle donne l'exemple de certaines prépositions, non commutables au sein de certaines structures syntaxiques :

fr. : *Les pétitionnaires demandent au parlement* ***d'****établir* (…)
angl. : *The Petitioners are asking* ***to*** *establish* (…)

De même « le pronom relatif, parce qu'il amalgame un morphème de subordination et un morphème anaphorique reprenant le contenu sémantique du SN qui précède, ne peut commuter indépendamment. » (1999 : 68)

La raison de ces limites se trouve dans le principe même du test de commutation bilingue. En effet, ce qu'on cherche à mettre en correspondance, ce sont des *unités de traduction* : par voie de conséquence, les différences sémantiques engendrées par la commutation ne doivent dépendre que de ces unités, et *non d'éventuels changements syntaxiques*. En d'autres termes, la commutation doit être « *isosyntaxique* », c'est-à-dire que les relations syntaxiques des unités commutées avec le reste de la phrase doivent être rigoureusement les mêmes.

Par exemple, dans les deux phrases déjà citées, on peut formellement faire commuter la préposition à condition de faire commuter le verbe simultanément :

fr. : *Les pétitionnaires* ***demandent*** **/** ***vont*** *au parlement* ***d'*** **/** ***pour*** *établir* (…)
angl. : *The Petitioners are* ***asking*** / ***coming*** *to establish* (…)

On en déduirait, après application de la transitivité, la relation :

demandent de || *are asking*

Mais ces commutations, formellement correctes, ne sont pas licites car elles ne sont pas *isosyntaxiques*. En effet, le changement d'actance découlant de la commutation de *demandent* avec *vont* transforme la fonction du syntagme deuxième actant *d'établir…* en un circonstant : *pour établir…* Par ailleurs le rôle syntactico-sémantique de la préposition *au* s'en trouve modifié, même si l'on n'observe pas de changement en surface.

De la même manière, la commutation ne doit pas altérer le sens des unités voisines lorsque celles-ci sont polysémiques. Le choix d'une acception dépendant des relations contextuelles, la commutation d'une unité peut avoir des effets de bord et entraîner une réinterprétation des formes avoisinantes.

fr. : […] *la base bruxelloise du mouvement qui* ***mène des*** / ***parcourt les*** *campagnes* […]
angl. : […] *its Brussels centre , which* ***runs campaigns*** / ***travels all over countries*** […]

III.1.3.3.3 Délimitation du champ commutationnel

Nous avons précisé les caractéristiques d'une commutation légitime vis-à-vis de la compositionnalité. Reste à éclaircir un point : toutes les commutations permises, à l'intérieur de ces limites, convergent-elles nécessairement vers un même découpage des unités de commutation ? Car le linguiste ou le traducteur qui effectue les commutations possède une marge de liberté qui s'exerce à deux reprises : il choisit d'une part une forme ou un groupe de formes de substitution, et il établit d'autre part la traduction qui lui semble convenir le mieux. Si différentes commutations aboutissaient à différents découpages, cette liberté de choix introduirait une part d'arbitraire dans le résultat de l'opération. Qui plus est, ayant décidé *a priori* des formes à mettre en correspondance, il est peut être possible, à rebours, de trouver un groupe de commutation qui satisfasse cette décision. Auquel cas, la commutation deviendrait caduque en tant que procédure de découverte et de contrôle : les unités qui en découlent seraient identifiées en amont.

Reprenons l'exemple précédent, en le modifiant légèrement :

fr. : […] *Les membres de nos services de sécurité* […]
angl. : […] *members of our security services* […]

fr. : […] *Les membres de nos services de* ***sécurité*** / ***renseignement*** […]
angl. : […] *members of our* ***security*** / ***intelligence*** *services* […]

fr. : […] *Les membres de nos* ***services*** / ***unités*** *de sécurité* […]
angl. : […] *members of our security* ***services*** / ***unit*** […]

d'où l'on déduit les correspondances indépendamment :

services || *services* et *sécurité* || *security*

Mais si l'on effectue une autre commutation, on aboutit à d'autres unités :

fr. : […] *Les membres de nos services de* ***sécurité*** / ***entretien*** […]
angl. : […] *members of our* ***security services*** / ***maintenance department*** […]

et l'on obtient cette fois :

services de sécurité || *security services*

En fait, la commutation peut être l'occasion d'introduire artificiellement une expression figée, qui a un impact global sur la traduction, et aboutit à une segmentation plus large. Par exemple, supposons que les phrases à aligner soient :

fr. : […] *Les membres de nos services parisiens* […]
angl. : […] *members of our Parisian services* […]

Les expressions *services parisiens* et *Parisian services* sont des combinaisons libres. Considérons les commutations suivantes :

fr. : […] *Les membres de nos services* ***parisiens*** / ***culturels*** […]
angl. : […] *members of our* ***Parisian services*** / ***cultural department*** […]

fr. : […] *Les membres de nos* ***services*** / ***bureaux*** *parisiens* […]
angl. : […] *members of our Parisian* ***services*** / ***offices***[…]

on en déduit : *services parisiens* || *Parisian services*

Le fait d'avoir introduit artificiellement l'expression figée *cultural department* aboutit à traiter *service parisien* comme une unité à part entière.

En résumé, le choix des commutations n'est pas innocent : suivant les cas de figure (et la combinatoire de la langue), ils peuvent aboutir à la conservation ou à la fragmentation des expressions figées ou semi-figées telles que *services de sécurité*, au même titre que des expressions libres, du type *services parisiens*. On est à nouveau confronté au problème de la non-consistance des unités de segmentation évoqué dans le précédent chapitre.

Pour réduire cette forme d'arbitraire, il faut donc préciser encore un peu plus les modalités d'application du test. En effet, ces modalités ne peuvent se restreindre aux propriétés formelles de la commutation, telles que les principes de transitivité et de minimalité précédemment dégagés. Ce qui manque à notre caractérisation, c'est la définition des paradigmes dont découlent les unités de commutation. En d'autres termes, il faut préciser, pour chaque unité de la phrase source, un *champ commutationnel*. Il existe deux façons de borner ce champ commutationnel, de l'extérieur et de l'intérieur, en précisant respectivement :

1. les commutations illicites ;
2. l'ensemble minimal des commutations licites permettant de dégager une segmentation stable.

En visant un certain degré de généralité, nous proposons les critères suivants :

– *Les unités commutées, pour être licites, doivent satisfaire aux principes suivants* :

  – *Principe de conservation syntaxique*. Les unités ne doivent pas altérer les relations syntaxiques sous-jacentes au reste de la phrase (cf. supra).

  – *Principe de liberté combinatoire*. Les unités ne doivent pas introduire des constructions polylexicales nouvelles en se composant avec des unités voisines qui étaient en combinaison libre. Par exemple, on ne commutera pas *service* ***parisien*** / ***militaire.***

  – *Principe de conservation isotopique.* Quand les unités sont en combinaison libre avec les unités voisines, elles ne doivent pas en altérer l'interprétation sémantique. Par exemple, on ne commutera pas ***mener*** / ***parcourir*** *la campagne* ni *le sol était jonché de feuilles* ***vertes*** / ***A4***

  – *Principe de compatibilité sémantique*. Les unités ne doivent pas aboutir à des phrases sémantiquement « anormales ». Par exemple, on ne commutera pas *les membres de nos* ***écoles*** / ***maternelles*** *de police*

– *Pour que la commutation soit significative, le champ commutationnel doit avoir un minimum d'amplitude* :

  – *Principe de différence sémantique*. Les unités commutées doivent présenter une différence sémantique minimale, afin que la traduction manifeste la variation. Par exemple, on ne commutera pas *accepté* ***comme*** / ***en tant qu'*** *acquis communautaire*.

  – *Principe de dissymétrie*. Lorsque l'unité commutée fait partie d'une construction plus large (mot composé, locution, collocation), il faut chercher à la substituer par une unité qui rompe la symétrie du point de vue de la traduction. On cherchera ainsi à affecter globalement la traduction du phrasème, de sorte qu'il soit conservé comme unité de commutation complète. Ainsi, comme on l'a vu supra, on ne commutera pas *services de* ***sécurité*** / ***renseignements*** mais plutôt *services de* ***sécurité*** / ***entretien***, afin dégager *service de sécurité* comme unité à part entière.

Les unités les plus favorables pour la commutation sont celles qui partagent le même taxème, c'est-à-dire la même classe minimale de substitution. Ces unités permettent généralement d'apporter une différence en respectant les principes de conservation syntaxique et isotopique. La relation d'antonymie peut aussi jouer un rôle intéressant, lorsque l'antonyme est sémantiquement compatible.

Il ressort de ce qui précède que la commutation n'est pas une opération aussi simple qu'il n'y paraît. En outre, dans l'application de nos critères, la décision de traiter des unités séparément ou ensemble *précède* l'application de la commutation. On ne peut donc conclure que les unités polylexicales dérivent simplement de la commutation : celle-ci ne fait qu'étayer les hypothèses émises en amont.

Malgré ces réserves, il reste que la commutation est sans doute le test le plus intéressant pour guider la détermination des unités qui découlent de la compositionnalité.

III.1.3.3.4 Problème des unités sans correspondance

Un autre problème est sous-jacent à la commutation : implicitement, pour que le test soit envisageable, les phrases source et cible doivent avoir le même *contenu sémantique*. En effet, la commutation des unités est censée suivre les deux phases déjà décrites : création d'une différence dans la source et rétablissement de l'identité sémantique par création de la même différence dans la cible. Mais lorsqu'il n'y a pas exactement identité sémantique au départ la possibilité de la double commutation devient caduque. En effet, même si l'on rétablit l'identité sémantique en commutant, les unités de départ resteront réfractaires aux correspondances déduites.

Reprenons un exemple déjà donné, et cherchons à appliquer la commutation, en prenant soin, à chaque fois, de rétablir au mieux l'identité sémantique.

> fr. : (…) *l'émission de billets de banque identifiables par les aveugles et par les personnes à vision réduite*
> angl.: (…) *the marking of banknote for the benefit of the blind and partially sighted*

> fr. : (…) ***l'émission*** / ***la destruction*** *de*/***s*** *billets de banque identifiables par les aveugles et par les personnes à vision réduite*
> angl.: (…) *the* ***marking*** / ***destruction*** *of banknote* ***for the benefit of*** / ***that are identifiable by*** *the blind and partially sighted*

> fr. : (…) *l'émission de billets de banque* ***identifiables*** / ***inutilisables*** *par les aveugles et par les personnes à vision réduite*
> angl.: (…) *the* ***marking*** / ***issue*** *of banknote* ***for the benefit of*** / ***useless for*** *the blind and partially sighted*

On obtient, par l'application de la transitivité :

l'émission …identifiables || marking … for the benefit of

Que signifie cette correspondance ? en dehors du contexte précis de ces deux phrases, rien. Le problème est le suivant : *émission*, tout comme *identifiables* n'ont pas d'équivalent précis dans la phrase anglaise, de même que *marking* et *for the benefit of* dans l'autre sens.

Cette absence de correspondant claire peut être caractérisée par les possibilités importantes de commutation sans contrepartie : *émission* peut commuter avec *création*, *fabrication*, *impression*, *tirage*, *production*, *diffusion*, *introduction* sans que la relation d'équivalence avec la phrase anglaise n'en soit altérée. De même *identifiables* peut commuter avec *utilisables*, *reconnaissables*, *lisibles*, *manipulables*, *déchiffrables*, etc. Ces possibilités de commutation « à vide » dénotent le lien très lâche de ces unités avec la phrase cible.

On peut se risquer à donner une explication géométrique de notre problème. Supposons que la construction du sens de la phrase, dans un contexte interprétatif donné, et pour une construction syntaxique donnée, soit une fonction de son lexique : ce genre de fonction (qu'on pourra nommer *fonction interprétative*) fera vraisemblablement alterner, avec la variation du lexique le long d'un continuum sémantique, des périodes de progression continue avec des ruptures « catastrophiques » (au sens de René Thom). Ces ruptures correspondent aux passages brutaux d'une interprétation à une autre dans l'espace des concepts, espace englobant (et non réductible au sémantique) puisque fonction des représentations encyclopédiques du monde. Dès lors la notion de « petite variation »[177] dans la phrase source, compensée par une « petite variation » dans la phrase cible, n'est tenable qu'à deux conditions :

- que ces variations opèrent sur des portions *continues* des deux fonctions interprétatives. Par exemple, lorsqu'une commutation crée l'émergence d'une construction polylexicale plus large, il y a forcément rupture de la continuité sémantique.

- que les deux phrases construisent le sens de manière *analogique*, suivant les mêmes rapports lexicaux, condition nécessaire au parallélisme des variations, garantissant la possible mise en correspondance des unités commutées. Lorsque deux phrases ne construisent pas le sens de manière analogique, il est probable qu'une petite variation, sur une unité, affecte plusieurs unités simultanément dans la phrase cible. Mais les unités sources et cibles ainsi connectées, parce qu'elles

[177] On peut raisonnablement y voir une métaphore du calcul infinitésimal.

« varient » ensemble, ne seront pas pour autant sémantiquement identiques : c'est le cas dans *l'émission …identifiables* || *marking … for the benefit of.*

Enfin, dans son étude, Mahimon (1999 : 59) propose une autre solution que la commutation, en s'inspirant d'une méthode d'analyse des traductions développée par Nida (1964 : 184-192). Le principe en est simple : partant du texte source, on peut appliquer mécaniquement ce que Nida appelle le *transfert minimal*, en effectuant une traduction mot à mot, sur laquelle on réalise les adaptations minimales requises par la grammaire de la langue cible.

	Texte Source		***Transfert minimal***		***Texte cible***
1	*Les*	1	*The*	1	*The*
2	*pétitionnaires*	2	*petitioners*	2	*petitioners*
3	*estiment*	3	*consider*	3	*believe*
4	*que*	4	*that*	4	*that*
5	*l'*	5	*the*	5	*the*
6	*usage*	6	*use*	6	*use*
7	*de*	7	*of*	7	*of*
8	*la*	8		8	
9	*marijuana*	9	*marijuana*	9	*marijuana*
10	*est*	10	*is*	10	
11	*cause*	11	*cause*	11	*causes*
12	*de*	12	*of*	12	
13	*problèmes*	14	*physical*	14	*physical*
14	*physiques*	15	*psychological*	15	*psychological*
15	*psychologiques*	16	*and*	16	*and*
16	*et*	17	*financial*	17	*financial*
17	*financiers*	13	*problems*	13	*problems*
18	*qui*	18	*which*	18	
19	*mènent*	19	*are leading*	19	*leading*
20	*à*	20	*to*	20	*to*
21	*une*	21	*an*	21	*an*
22	*recrudescence*	22	*increase*	22	*increase*
23	*de*	23	*in*	23	*in*
24	*la*	24		24	
25	*criminalité*	25	*crime*	25	*crime*

tableau 41 : *appariement basé sur la construction d'un transfert minimal*

Bien sûr, Nida n'assimile pas ce type de transfert à la traduction proprement dite, qu'il nomme *transfert littéraire*. Le transfert minimal est en quelque sorte la projection la plus proche du texte source dans le système d'arrivée : ce n'est même pas ce que nous appelons une *traduction littérale*, qui implique un transcodage complet intégrant toutes les composantes de l'idiome d'arrivée, et pas seulement les contraintes grammaticales ou lexicales.

Le transfert minimal proposé par Nida peut ensuite servir de base de comparaison pour l'analyse de traductions réelles : la comparaison s'effectuant à l'intérieur d'une même langue, elle est censée devenir plus facile. Mahimon propose d'utiliser ce type de comparaison pour établir des correspondances, puisqu'il est possible de relier les unités cibles de la traduction avec les unités issues du transfert minimal (cf. l'exemple du tableau 41, donné par Mahimon, 1999 : 56, où les relations traductionnelles sont indiquées par des numéros).

Mais cette méthode apparemment simple pose peut être plus de problèmes qu'elle n'en résout :

- La traduction mot à mot n'est pas une opération si « mécanique » que cela puisqu'elle impose de choisir, pour chaque mot, parmi toutes les traductions possibles. L'idée que le transfert minimal puisse fournir une base de comparaison objective servant de référence unique pour la comparaison des traductions paraît illusoire : le transfert minimal n'est pas donné, mais construit, et implique des choix plus ou moins arbitraires.

- Rien ne permet d'affirmer que la comparaison du transfert minimal avec une traduction donnée soit plus facile que la comparaison directe de la source avec la cible : comme nous l'avons montré, les possibilités de paraphrase à l'intérieur d'une même langue sont aussi vastes que les possibilités de traductions entre deux langues. Si la comparaison est basée sur l'*identité* des unités du transfert minimal avec les unités du texte cible, on risque de ne trouver qu'une partie des correspondances ; si la comparaison est basée sur la similitude, i.e. l'équivalence sémantique des unités cibles, la méthode est inutile puisqu'une telle comparaison

> peut être effectuée directement, sans passer par une traduction mot à mot. L'introduction du transfert minimal ne fait selon nous qu'introduire un intermédiaire supplémentaire qui complique la comparaison de la source avec la cible.

A la différence du test de commutation, cette méthode ne fournit pas de critères formels susceptibles de résoudre les cas litigieux, car elle repose sur l'intuition et la compétence du locuteur de la même manière qu'une comparaison directe de l'original avec sa traduction.

III.1.4 Mise au point d'un corpus de référence

Afin de poursuivre notre tâche d'évaluation des méthodes destinées à l'extraction automatique des correspondances, il nous faut maintenant élaborer un corpus de référence, contenant des correspondances établies manuellement. Ces correspondances nous serviront par la suite d'étalon destiné à la comparaison des résultats des méthodes étudiées. Elles doivent donc découler, de manière aussi rigoureuse et systématique que possible, de la définition que nous avons établie précédemment, tant sur le plan de la segmentation que sur celui de l'appariement.

Au bout du compte, cette partie de notre travail s'est révélée beaucoup plus longue et plus ardue que nous l'avions prévu. Nous nous sommes heurté à de très nombreux problèmes de cas limite et de frontières floues, face à des questions du genre : faut-il rejeter ou accepter tel couple d'unités présentant une différence sémantique importante ? cette tournure forme-t-elle une unité à part entière ou bien faut-il la décomposer ?

Pour une réalisation systématique et rigoureuse, il aurait fallu développer des critères *ad hoc* intégrant un nombre toujours plus grand de cas particuliers. Même si cette étape avait pu être réalisée de façon pleinement satisfaisante, il aurait été intéressant de confier la mise en œuvre de ces critères à plusieurs personnes, comme dans les projets ARCADE et Blinker (Melamed, 1998d).

Dans ce dernier projet, 250 versets de la Bible ont été annotés manuellement, afin que chaque mot soit relié à zéro, un ou plusieurs mots correspondants dans le segment

aligné. Pour assurer une certaine cohérence, Melamed se base sur le consensus intersubjectif : la tâche d'annotation est effectuée cinq fois, par cinq annotateurs différents. Un guide d'annotation est mis au point, en concertation avec les annotateurs, afin de les contraindre à adopter des stratégies communes face aux problèmes les plus courants. Par suite, l'application des critères définis dans ce guide dépend du jugement intuitif des annotateurs. Afin de limiter les erreurs et de faciliter la tâche des annotateurs, un outil d'annotation *ad hoc*, disposant d'une interface adaptée, a été développé. Enfin, l'accord entre les annotateurs peut être mesuré quantitativement, par des mesures telles que précision et rappel.

Ainsi, Melamed a pu évaluer la part de variation subjective susceptible de compromettre la validité des annotations. L'accord moyen, entre les annotateurs pris deux à deux, est d'environ 82 % (ce pourcentage dérive de la F-mesure). En ne tenant compte que des mots pleins (« *content words* »), l'accord est meilleur, aux alentours de 92 %. Melamed en conclut que le corpus annoté est « raisonnablement fiable » (« *reasonably reliable* ») comme étalon de référence pour l'évaluation (« *gold standard* »).

En ce qui concerne notre propre évaluation, nous n'avions ni le temps ni les moyens de nous lancer dans une entreprise similaire. Lors de la constitution de notre corpus de référence, nous avons donc dû réviser sérieusement nos ambitions et trouver un compromis : pour que l'évaluation soit représentative, il nous fallait une quantité minimale de correspondances – il était donc hors de question d'appliquer des méthodes aussi coûteuses que la commutation. Mais par ailleurs, les correspondances extraites devaient présenter des garanties de consistance et de cohérence, pour ne pas compromettre la signification de nos résultats. Partant de principes généraux destinés à guider nos choix, nous avons opéré de manière intuitive, en explicitant en cours de route certains critères plus précis. Malgré nos efforts, nous devons admettre qu'une grande part d'appréciation intuitive et de subjectivité sont intervenus dans nos choix : au final le corpus de référence n'est pas exempt d'incohérences et d'imperfections diverses. Cependant, comme le montreront les résultats ultérieurs, le « bruit » issu de ces imperfections n'altère en rien certaines conclusions, dans l'évaluation des méthodes. Même grossier, cet étalon peut jouer pleinement son rôle de comparateur.

Pour élaborer le corpus de référence, nous avons choisi d'utiliser le corpus JOC, déjà employé dans le projet ARCADE pour évaluer une tâche voisine de l'extraction des correspondances, le « *translation spotting* » (cf. p. 351). Le corpus JOC est constitué de questions écrites posées par des membres du Parlement européen, suivie des réponses données par la Commission. Ces questions concernent des sujets variés : agriculture, économie, environnement, institutions, droits de l'homme, transports, etc. Elles ont été publiées en 1993, dans les Séries C du Journal officiel de la Communauté européenne, et collectées dans le cadre du projet MLCC-MULTEXT.

Pour le corpus de référence nous avons retenu un échantillon du corpus JOC, par tirage aléatoire. Sur les 69 160 binômes générés par l'alignement automatique de la phase précédente, nous en avons tiré 1 000. Parmi ces 1 000 couples de phrases, un certain nombre ne contenaient que des appariements de cellule de tableau du type (3$; 3$). En supprimant ces binômes (sans intérêt pour l'évaluation), on aboutit à 767 couples de phrases, qui se répartissent comme suit[178] :

	Corpus JOC entier		***Corpus de référence (échantillon du JOC)***	
	Anglais	*Français*	*Anglais*	*Français*
Couples de phrases	69 160		767	
Occurrences de formes simples	1 070 630	1 251 279	14 852	17 962
Types de formes simples	29 779	36 003	3 571	4 082

tableau 42 : *constitution de l'échantillon*

Les termes *type* et *occurrence* recouvrent la distinction usuelle vocable / mot effectuée par C. Müller (1968) : les types sont les unités du vocabulaire attaché au corpus, les occurrences sont les instanciations de ces types dans le texte.

[178] Cette fois, dans les comptages de formes simples, nous n'avons pas utilisé les statistiques de *Word*, mais nous avons tenu compte des unités effectivement traitées par nos algorithmes. On peut noter quelques variantes : dans la mesure où nous avons négligé les signes de ponctuation (point, virgule, point virgule, deux points, apostrophes, points d'exclamation et d'interrogation, parenthèses, tiret, slash, guillemet) nous comptions parfois deux mots quand *Word* n'en comptait qu'un, notamment dans le cas des mots liés par un tiret ou une apostrophe.

III.1.4.1 Principes de segmentation

Au chapitre I.1.3.5, nous avons énuméré les principales catégories d'unités de traduction polylexicales susceptibles d'intéresser les traducteurs : les composés figés, les idiotismes, les collocations, les pluritermes… Au cours de la segmentation manuelle, nous avons tenté de prendre en compte toutes ces catégories d'unités, afin qu'elles soient considérées comme un tout dans l'extraction des correspondances.

Selon les principes dégagés précédemment, l'identification des unités a été effectuée séparément pour chaque partie du bi-texte, indépendamment de sa traduction. Cependant, certaines unités de traduction ne sont apparues qu'après l'établissement des correspondances : c'est le cas des périphrases, en combinaison libre, qui présentent un certain degré de généralité, et correspondent à une seule unité lexicale de l'autre côté du bi-texte.

Par exemple :

fr. : *nomadisme*
angl. : *gypsy way of life*

Ignorer ces unités polylexicales « libres » nous aurait conduit à rejeter de très nombreuses correspondances valides hors contexte, et donc susceptibles de s'intégrer dans une mémoire de traduction. En outre, le procédé périphrastique étant très courant dans la traduction de lexies n'ayant pas d'équivalent standard dans la langue d'arrivée, il serait dommage de ne pas inscrire, dans une base de correspondances lexicales, les solutions trouvées dans les corpus traductions.

III.1.4.2 Comparaison avec les consignes du projet ARCADE

Il peut être intéressant de comparer nos propres choix aux critères établis pour la constitution du corpus de référence d'ARCADE. Le « *translation spotting* » s'appuie sur trois règles de base :

– *Règle 1* : *Enregistrer autant de mots que nécessaire de chaque côté*[179]

Etant donné un mot dans le texte source, le repérage de sa traduction dans la cible nécessite d'intégrer autant de mots que nécessaire dans la cible, puis dans la source, afin d'assurer une « équivalence réciproque ».

Par exemple :

Incorrect : fr. : *une **carte** de paiement* ↔ angl. : *a **pay-card***
Correct : fr. : *une **carte de paiement*** ↔ angl. : *a **pay-card***

Nous avons une conception différente de l'équivalence visée par la règle 1. Le guide d'annotation donne l'exemple suivant :

fr. : *en faveur des **petits** redevables*
angl. : *persons **with low turnovers***

Cette équivalence respecte le niveau fonctionnel : adjectif ↔ groupe adjectival. Mais sur le plan sémantique, l'équivalence est plus complète au niveau des syntagmes nominaux :

fr. : *petits redevables* ↔ angl. : *persons with low turnovers*

Si l'on s'en tient au strict plan des unités lexicales, il y a un résidu important, et l'on doit se contenter de :

fr. : *petits* ↔ angl. : *low*

Ces exemples soulèvent le problème du parallélisme grammatical : les structures étant différentes il peut être vain de rechercher une identité fonctionnelle dans l'application de la règle 1. Par exemple le guide ARCADE suggère d'intégrer les pronoms relatifs dans les exemples suivants :

[179] "Rule 1 : Mark as many words as necessary on each side"

Incorrect :
fr. : *C'est le cas, a priori, de la "Lion Mark" qui **apporte** des informations supplémentaires*
angl. *The Lion Mark is one such mark, **providing** the purchaser with additional information*

Correct :
fr. : *C'est le cas, a priori, de la "Lion Mark" **qui apporte** des informations supplémentaires*
angl. *The Lion Mark is one such mark, **providing** the purchaser with additional information*

La prise en compte du pronom relatif est commandée par l'identité fonctionnelle (au sens grammatical, et non pragmatique). Dans notre perspective, l'identité fonctionnelle n'est pas requise, puisque nous négligeons les contenus grammaticaux dans l'établissement de l'équivalence. Les unités étant déterminées indépendamment de la compositionnalité, nous acceptons les correspondances entre unités de nature différente :

Là où le guide ARCADE préconise :

fr. : *pour les étudiants **qui ont de petits moyens***
angl.: *for **less well off** students*

nous préférons :

fr. : *pour les étudiants qui **ont de petits moyens***
angl. *for **less well off** students*

Notons que la possibilité d'extraire cette correspondance de son contexte est discutable : elle est néanmoins intéressante d'un point de vue idiomatique, et montre que la notion d'équivalence peut parfois s'appliquer à des unités antonymes.

Le guide reconnaît par ailleurs la possibilité d'apparier des unités assumant des fonctions différentes. Par exemple, il propose :

fr. : *La Commission a fait la **proposition** suivante*
angl. : *The Commission **proposed** the following*

Mais dans ce cas notre choix aurait été différent, pour une autre raison : *faire* + *proposition* peut être considérée comme une unité intégrant la collocation d'un substantif avec son verbe support.

– *Règle 2* : *Enregistrer aussi peu de mots que possible chaque côté*[180]

Dans les limites d'application de la règle 1, il faut rechercher le granularité la plus fine. Par exemple :

Incorrect : fr. : *une* ***carte de paiement*** ↔ angl. : a ***pay card***
Correct : fr. : *une* ***carte*** *de paiement* ↔ angl. : a ***pay*** *card*

Mais pour décomposer ainsi une unité en sous parties et apparier ces sous parties séparément, la compositionnalité doit être complète : « Si certaines sous parties se correspondent indépendamment dans d'autres contextes, mais pas toutes, l'expression doit être considérée comme un tout. »[181]

Incorrect : fr. : *des petits* ***pois*** ↔ angl. : *vining* ***peas***
Correct : fr. : *des* ***petits pois*** ↔ angl. : ***vining peas***

Par rapport à cette règle, notre approche est divergente, puisqu'elle se base sur une vision maximaliste des unités : lorsqu'une unité polylexicale est traduisible mot à mot, nous gardons néanmoins cette unité comme un tout. Dans le cas de *carte de paiement*, le parallélisme de l'expression anglaise *pay card* ne nous permet pas de décomposer, si *carte de paiement* est reconnue comme une unité (par exemple en tant que pluriterme).

Par exemple, le guide ARCADE propose :

fr. : *La Communauté européenne* ***apporte*** *une aide aux …*
angl. : *The European Community* ***has been providing*** *assistance to …*

[180] "Rule 2 : Mark as few words as possible on each side"
[181] "If some subpieces match independently in other contexts, but not all, mark the expression as a whole."

Pour nous, *apporter une aide à* et *to provide assistance to* sont des unités de traduction (collocations). Même si la traduction est compositionnelle, nous préférons les conserver telles quelles :

fr. : *La Communauté européenne* ***apporte une aide aux*** …
angl. : *The European Community* ***has been providing assistance to*** …

Nos unités de traduction doivent être d'un rang supérieur ou égal à celui des unités lexicales : elles ne peuvent par exemple être incluses dans une unité figée, même si la langue cible autorise le calque. Ce choix est motivé par la recherche d'une meilleure adéquation entre unité de traduction et unité de pensée, ou concept.

C'est pourquoi nous souscrivons à l'application de la troisième règle :

– *Règle 3* : *Enregistrer les mots entiers*[182]

Notamment, les composés séparés par des tirets ne doivent pas être décomposés.

Incorrect :	fr. :	*une* ***carte*** *de paiement*	↔	angl. :	*a pay****card***
Incorrect :	fr. :	*une* ***carte*** *de paiement*	↔	angl. :	*a pay-****card***
Correct :	fr. :	*une* ***carte de paiement***	↔	angl. :	*a* ***paycard***
Correct :	fr. :	*une* ***carte de paiement***	↔	angl. :	*a* ***pay-card***

Enfin, le guide d'annotation comporte un certain nombre de recommandations précises relatives à des cas de figure particuliers. Examinons ces remarques point par point :

– *Noyau verbal*

Le guide d'annotation remarque que le passage à la traduction altère parfois les valeurs morphosyntaxiques comme le nombre, le temps, le mode, etc. Ces différences étant négligées dans l'extraction des correspondances, les unités appariées peuvent porter des marques différentes :

fr. : *Quelle contribution la Communauté **apporte**-t-elle aux …*
angl. : *What funding **is provided** by the EC for*

fr. : *La Communauté européenne **apporte** une aide aux*
angl. : *The European Community **has been providing** assistance to*

Les auxiliaires doivent être intégrés avec le verbe, sauf s'ils ont une contrepartie dans la traduction (par application de la règle 2) :

Incorrect :

fr. : *La Communauté européenne **apporte** une aide aux …*
angl. : *The European Community has been **providing** assistance to …*

fr. : *La Commission a toujours **financé** de manière soutenue …*
angl. : *The Commission **has provided** steady **funding** for …*

Correct :

fr. : *La Communauté européenne **apporte** une aide aux …*
angl. : *The European Community **has been providing** assistance to …*

fr. : *La Commission a toujours **financé** de manière soutenue …*
angl. : *The Commission has **provided** steady **funding** for…*

Par ailleurs l'auxiliaire anglais *do* est négligé lorsqu'il apparaît dans les formes négatives, dans les interrogatives et dans les tournures d'insistance.

Nous avons appliqué ces mêmes règles, en considérant les auxiliaires comme faisant partie de la morphologie verbale, sauf quand il y avait parallélisme, afin de pouvoir apparier les auxiliaires indépendamment le cas échéant (la correspondance des auxiliaires pouvant constituer une information intéressante).

En ce qui concerne les particules verbales, le guide suggère de les intégrer au noyau verbal dans la mesure où elles sont indissociables de son contenu sémantique :

[182] "Rule 3: Mark entire words"

Incorrect :

fr. : *son intention de **reprendre** au KYDEP les entrepôts...*
angl. : *to **take** away from KYDEP the warehouses...*

Correct :

fr. : *son intention de **reprendre** au KYDEP les entrepôts...*
angl. : *to **take away** from KYDEP the warehouses...*

Nous avons également appliqué cette règle, qui découle directement de l'identification des unités polylexicales figées.

– *Prépositions*

Dans la perspective du guide, les prépositions qui suivent les verbes ne doivent pas leur être rattachées, sauf quand elles font partie d'une expression plus large :

Incorrect :

fr. : *la Commission **apporte son soutien à** un certain nombre d'initiatives*
angl.: *The Commission **promotes** a number of initiatives*

Correct :

fr. : *la Commission **apporte son soutien** à un certain nombre d'initiatives*
angl. *The Commission **promotes** a number of initiatives*

fr. : *Les notices ont été **postées**...*
angl. : *The notices have been **sent by post**...*

Nos critères nous conduisent à une solution opposée : nous avons montré que dans certains cas la préposition n'est pas en combinaison libre mais dépend du verbe : il est alors plus cohérent de considérer l'ensemble *verbe + prép*. comme une unité.

– *Conjonctions non parallèles*

Il peut arriver que les termes d'une conjonction ne soient pas distribués de façon parallèle dans les deux langues. Le guide donne les exemples suivants :

fr. : *permet des économies substantielles et **apporte** de nombreuses possibilités*
angl. : *does **permit** significant saving and many additional opportunities*

fr. : *émetteurs de **cartes de crédit** ou **de paiement**.*
angl. : ***credit-card** or **pay-card issuers**.*

Ces exemples montrent qu'une même unité peut entrer dans plusieurs appariements. Dans le cas du « *translation spotting* » on peut faire apparaître un seul de ces appariements, comme dans le cas de (*apporte* ; *permit*), ou bien les deux appariements simultanément comme dans le cas de (*cartes de crédit … de paiement* **;** *credit-card … pay-card*), suivant l'unité dont on cherche la correspondance.

Dans la constitution de notre corpus de référence, nous pouvons faire figurer plusieurs appariements pour chaque couple de phrases : on peut donc réutiliser plusieurs fois le même mot, comme *carte* dans : (*cartes de crédit* ; *credit-card*) *et* (*carte de paiement* **;** *pay-card*). Ainsi *télévision par satellite et par câble* a donné lieu à deux unités : *télévision par satellite* et *télévision par câble*.

– *Déterminants*

Le guide suggère de ne pas inclure les déterminants (articles, possessifs, adjectifs indéfinis) et les prépositions, à moins qu'elles fassent partie d'un composé nominal.

Incorrect :
fr. : *Considérant le rôle important que jouent **les taxis** …*
angl. : *In view of the important role played by **taxis** …*

fr. : *La Commission n'**apporte aucun soutien** à ces initiatives.*
angl. : *The Commission does not **support** these initiatives.*

Correct :
fr. : *Considérant le rôle important que jouent les **taxis**…*
angl. : *In view of the important role played by **taxis**…*

fr. : *La Commission n'**apporte** aucun **soutien** à ces initiatives.*
angl. : *The Commission does not **support** these initiatives.*

fr. : *Ce rapport **apporte la preuve**…*
angl. : *The report **demonstrates**…*

Ces règles sont conformes à nos propres principes. Dans certaines expressions comme *le cas échéant*, il est clair que l'article n'est pas en combinaison libre, et nous l'avons donc intégré à l'expression.

– *Intégration du génitif pour une meilleure adéquation sémantique* :

Le guide propose également d'ignorer la marque du génitif sauf s'il est sémantiquement présent dans la partie correspondante :

Incorrect :

fr. : *Ces gaz sont responsables du changement climatique dans le* ***monde***.
angl.: *These gases are responsible for the change in the* ***world's*** *climate*.

Correct :

fr. : *Ces gaz sont responsables du changement climatique dans le* ***monde***.
angl.: *These gases are responsible for the change in the* ***world****'s climate*.

fr. : *Ces gaz sont responsables du changement climatique* ***mondial***.
angl.: *These gases are responsible for the change in the* ***world's*** *climate*.

Ce genre de distinction ne peut être mis en œuvre lorsqu'on identifie les unités séparément de part et d'autre. Le génitif étant une marque morphologique comme une autre, on pourrait décider de l'intégrer au mot dans tous les cas, ou au contraire de la négliger systématiquement. Nous avons opté pour cette dernière solution.

– *Passif*

Dans le cas de construction diathétique, le guide d'annotation suggère d'intégrer les auxiliaires :

Incorrect :
fr. : *La Communauté* ***apporte*** *en Grèce*
angl. : *Greece has been* ***granted***

Correct :
fr. : *La Communauté* ***apporte*** *en Grèce*
angl. : *Greece* ***has been granted***

Le recours à la diathèse pouvant être révélateur de préférences idiomatiques, et la passivation affectant l'orientation sémantique du verbe, nous avons également opté pour l'intégration de l'auxiliaire.

Les convergences et les divergences entre nos propres critères et ceux du projet ARCADE illustrent la complexité du problème et la subtilité de certains choix : globalement, ils sont dus à des optiques différentes. Dans le cas du « *translation spotting* », on recherche un alignement de grain minimal, incluant un mot présélectionné : ce qui explique que la recherche d'équivalence sémantique induise l'intégration d'éléments grammaticaux comme les pronoms relatifs ou la marque du génitif, et que certains éléments comme les prépositions ne soient pas pris en compte s'ils sont traduits de manière compositionnelle : dans le cas du « *translation spotting* », la compositionnalité traductionnelle est le critère déterminant. Dans notre évaluation, c'est la consistance des unités qui prime, indépendamment de la traduction particulière qui en est donnée.

III.1.4.3 Exemples tirés du corpus

Afin d'illustrer nos principes de segmentation, nous donnons (cf. tableau 43) quelques exemples des unités de traduction extraites de notre corpus de référence. Ces exemples n'ont rien de systématique ni d'exhaustif, mais permettent de donner une idée de la variété des cas de figure et de la nature des nombreux problèmes qui se posent : cas limites d'équivalence sémantique, problème des articles contractés qui incluent une préposition et un déterminant, problème des unités discontinues, formules elliptiques, etc.

Pour chaque couple d'unités, nous avons fait figurer le numéro du segment d'où l'exemple est tiré (cf. annexe A-V). Nous mettons entre parenthèses les segments qui n'appartiennent pas à la rubrique indiquée, et ne figurent qu'en tant que traduction. Nous avons classé les segments en fonction des différents cas de figure. Le classement s'adresse soit à la version anglaise, soit la version française, soit aux deux. Enfin, le symbole ¤ indique une place vide (par exemple un argument) dans les unités discontinues.

N° du segment	*Version anglaise*	*Version Française*
	Expressions figées	
15698	*ad hoc*	*ad hoc*
4116	*acquis communautaire*	(*acquis communautaire*)
23359	*little egrets*	*aigrettes garzettes*
47864	*background notes*	*notes de synthèse*
15472	*budget heading*	*ligne budgétaire*
61718	(*rail*)	*chemins de fer*
32024	(*complications*)	*effets secondaires*
8072	(*above*)	*ci-dessus*
15652	(*closely*)	*de près*
50032	(*completely*)	*tout à fait*
10786	*At the moment*	*À l'heure actuelle*
56851	(*also*)	*en outre*
12837	(*and*)	*aussi bien ¤ que*
36486	*about to*	*en voie de*
37889	*in order to*	(*ainsi*)
2211	*with a view to*	*afin que*
66947	*on behalf of*	*au nom de*
56022	(*for*)	*à des fins d'*
	Intégration des prépositions « indicateurs d'argument » et du nucleus verbal (particules, auxiliaires temporels et modaux, pronoms réfléchis)	
56902	*change from ¤ to*	*remplacer ¤ par*
36867	*earmarked ¤ for*	*avait affecté ¤ au*
26777	*urge*	*intervenir auprès des*
10999	(*to*)	*accordée à*
11099	*put forward*	*avait présenté*
8072	(*communicated*)	*a communiqué*
24410	*resulted in*	*entraîné*
31647	*is proceeding with*	*a entrepris*
20903	*have been*	*a été*
21198	(*was*)	*a été*
14429	(*arise*)	*se manifestant*
53439	*intend to*	*a l'intention de*
47262	*to be able to*	*aient la possibilité de*
	Collocations	
17020	*adding its support*	*apporter son appui*
50386	*adverse effects*	*incidences négatives*

Idiotismes

58777	*capital injection*	*apport de capital*
27577	*race riots*	*violences racistes*
64550	*award-winning*	*récompensés*
15941	*employment*	*activité professionnelle*
25524	*measures ¤ for its protection*	*mesure de protection*
42113	*response preparedness*	*la faculté de faire face à*
37302	*are well aware of*	*n'ignorent pas*
65615	*be circumvented*	*échapper à*
15063	*be left unemployed*	*perdre leur emploi*
48088	*contradicts*	*est en contradiction avec*
65540	*cooperate in*	*unir leurs efforts pour*
15063	*building a nuclear arsenal*	*se doter de l'arme nucléaire*
50032	*have reached maturity*	*sont arrivés à maturité*
9545	*have taken place with the support of*	*ont bénéficié du concours de*
20399	*is sending out negative signals*	*transmet une image négative*
41979	*in late*	*à la fin de l'année*
54754	(*for*)	*à des fins*
65791	*in recent months*	*au cours des derniers mois*
30594	*in the Community*	*au niveau communautaire*
50123	*this is clear from the fact that*	*ainsi qu'il ressort de*
11065	*there is good reason to believe that*	*il semble justifié de*
18819	*of little assistance*	*peu adéquats*
30286	*least well off*	*plus démunis*
21198	*years of higher education*	*BAC +*
47326	*the question arises*	*la question se pose*

Noms propres, toponymes, noms d'institutions, pluritermes

57504	*Baltic Sea*	*mer Baltique*
12098	*Mr José Valverde López*	*M. José Valverde López*
36142	*olive oil*	*huile d'olive*
8072	*Board of Governors*	*Conseil des gouverneurs*
44891	*ACP*	*Afrique des Caraïbes et du Pacifique ACP*
66141	*Eastern European countries*	*pays d'Europe de l'Est*
37095	*Community Member States*	*Etats membres de la Communauté*
30900	*Community Support Frameworks*	*Cadres communautaires d'appui CCA*
23281	*assessment exercises*	*bilans préventifs*
36211	*Community law*	*droit communautaire*
55821	*anti-fraud work*	*lutte contre la fraude*
52030	*blood products*	*dérivés sanguins*
26862	*breakdown services*	*assistance routière*
52980	*social security benefits*	*prestations de sécurité sociale*
20053	*non-profitmaking companies*	*associations à but non lucratif*
18750	*financial year*	*exercice budgétaire*
22119	*financial year*	*exercice*

tableau 43 : *exemples d'unités de traduction identifiées manuellement*

III.1.4.4 Niveau d'équivalence retenu

Une fois les unités dégagées, nous pouvons établir des appariements sur la base de l'équivalence traductionnelle. Mais un certain nombre de ces appariements impliquent des distorsions sémantiques importantes et ne sont recevables que dans un contexte précis. Afin de les éliminer, nous nous sommes donné une définition restrictive de l'équivalence. Comme nous l'avons déjà remarqué, les équivalences peuvent décrire un continuum le long de l'axe de la dépendance contextuelle : nous avons choisi de ne retenir que les appariements ayant une certaine généralité, et détachables du contexte précis dont on les extrait. L'équivalence peut donc se situer à différents niveaux : sémantique, conceptuel, culturel, dynamique, etc., pourvu que l'appariement soit réutilisable dans des contextes divers. Ce critère de sélection nous paraît en effet le plus adapté dans le cadre de la constitution d'une mémoire de traduction.

Examinons l'exemple ci-dessous :

angl. : *If the Community is to make progress a common mining policy is essential*
fr. : *Le processus d'adoption d'une politique minière communautaire est à l'évidence bien avancé*

Nous retenons les correspondances suivantes :

C={(*Community* ; *communautaire*) (*common* ; *communautaire*) (*mining* ;*minière*) (*policy* ;*politique*)}

Du fait d'une certaine divergence de point de vue, certaines unités demeurent sans correspondances : *if the*, *essential*, etc.

D'autres entretiennent une relation d'équivalence étroitement liée au contexte : (*make progress* ; *adoption d*') – un schème conceptuel similaire, impliquant un processus évolutif, se retrouve dans chacun des deux membres. Mais cette équivalence est étroitement liée au contexte de la traduction, et dépend dans une large mesure du prédicat : « faire progresser une politique commune » ≈ « adopter une politique commune ». Nous avons donc rejeté une telle correspondance.

Pour les mêmes raisons, dans l'exemple suivant :

angl. : *Illegal transactions involving the heritage*
fr. : *Transactions illégales aux dépens du patrimoine*

nous n'avons pas conservé la correspondance (*involving* ; *aux dépens du*).

En revanche, comme le montrent les exemples de correspondances donnés dans le précédent tableau, nous avons autorisé une certaine variabilité sémantique, tant qu'elle n'affectait pas l'indépendance contextuelle : (*cooperate in* ; *unir leurs efforts pour*).

Enfin, le critère d'indépendance contextuelle nous a conduit à éliminer des correspondances basées sur une équivalence référentielle trop spécifiquement localisée dans l'espace ou dans le temps, comme :

(*Mr Delors* ; *le Président de la Commission*)

De même, toutes les références anaphoriques ont été éliminées :

Dans

angl. : *at the pre-Council briefings*
fr. : *lors des briefings susmentionnés*

nous n'avons pas retenu : (*pre-Council* ; *susmentionnés*)

Notons enfin que nous avons négligé, dans l'établissement de ces correspondances, toutes les valences sémantiques liées au système grammatical : actance, parties du discours, temps, aspect, nombre, etc. Ceci nous a conduit à accepter l'équivalence d'unités très diverses, pourvu qu'elles portent le même sens.

Par exemple, on a pu extraire :

(*According to* ; *fait valoir qu'*)

en se basant sur l'équivalence entre « *According to X, Y* » et « *X fait valoir qu'Y* », fréquente dans notre corpus.

Le seuil d'équivalence fixé aboutit donc au rejet d'un certain nombre de formes résiduelles, n'entrant dans aucune correspondance. Le résidu semble être légèrement plus important en français, ce qui dénote une tendance à l'explicitation plus marquée dans la partie française du bi-texte. Au final, environ 20 % des formes simples sont rejetées dans le

résidu : globalement, nous avons extrait 9 727 correspondances, concernant une proportion d'environ 80 % de formes simples :

	Anglais		***Français***	
Lexies simples	8 031	83 %	7 631	78 %
Lexies polylexicales	1 696	17 %	2 096	22 %

tableau 44 : *répartition des 9 727 couples manuellement extraits*

	Anglais		***Français***	
Nombre total de formes simples dans l'échantillon	14 852		17 962	
Nombre de formes simples dans les correspondances	12 384	83 %	13 689	76 %
Nombre de formes simples résiduelles	2 468	17 %	4 273	24 %

tableau 45 : *formes retenues et formes résiduelles*

III.1.4.5 Lemmatisation partielle

Pour certaines langues à la morphologie riche, la lemmatisation est une opération incontournable si l'on veut faire émerger des régularités au niveau des cooccurrences. Comme le notent Choueka, Conley & Dagan (in Véronis, 2000 §4), c'est le cas de l'hébreu :

> « Un texte anglais peut contenir le nom *computer* ou le verbe *to supervise* des dizaines de fois, tandis que les équivalents hébreux peuvent apparaître sous la forme de dizaines de variantes qui n'apparaissent qu'une ou deux fois chacune, ce qui fausse les statistiques pertinentes. »[183]

Même si le français et l'anglais sont loin d'avoir une morphologie aussi riche que celle de l'hébreu, nous avons, pour chaque couple retenu, manuellement ramené les unités à leur lemme. On peut ainsi s'attendre à une légère amélioration des résultats, du fait d'une plus grande régularité dans la distribution des cooccurrences.

Pour des raisons pratiques, nous n'avons effectué qu'une lemmatisation partielle. En effet, la lemmatisation a été faite unité par unité, à l'intérieur des couples extraits, hors de

[183] "An English test may contain the noun computer or the verb to supervise dozens of times, while the Hebrew counterparts would appear in dozens of formally different variants, occurring once or twice each, skewing by this the relevant statistics."

tout contexte syntaxique : dans certains cas d'ambiguïté, il ne nous était donc pas possible de trancher (p. ex. *prises* peut être rattachée au substantif *prise* comme au verbe *prendre*, *centrale* peut être un substantif ou un adjectif, etc.). Notons qu'une partie de ces ambiguïtés ont pu être levées grâce aux unités correspondantes : par exemple, dans le couple (*among*, *entre*), *entre* n'est pas ambiguë. Les plus nombreux cas d'ambiguïté irrésolus concernent les verbes du premier groupe en français : *contrôle* peut être un substantif ou bien la forme fléchie d'un verbe. Mais globalement, le nombre d'occurrences non lemmatisées reste marginal.

De toutes façons, le but de la lemmatisation est d'éliminer une certaine forme de variabilité superficielle autour de noyaux lexicaux portant un sens jugé constant : dans cette perspective, c'est une opération qui peut admettre différents degrés, de la simple élimination des marques de pluriel pour les substantifs, au regroupement des allomorphes de toute une famille dérivationnelle (p. ex. *informer*, *informateur*, *information*, etc.). Fung (in Véronis, 2000 §11) remarque par exemple que « les traductions chinoises de *beauty*, *beautiful*, *beautifully* sont souvent les mêmes. »[184] Dans l'étude de la traduction, où c'est la constance sémantique qui importe, plus que les considérations morphologiques, on pourrait ainsi envisager d'effectuer une lemmatisation profonde. Mais comme le remarquent Choueka, Conley & Dagan (in Véronis, 2000 §4), il n'est pas évident qu'une telle réduction soit profitable.

Nous avons choisi de conserver les frontières entre les parties du discours, afin de pouvoir mettre à jour, ultérieurement, des différences de comportement au niveau de ces catégories (nous verrons comment traiter les unités ambiguës dans des classes *ad hoc*, cf. infra, p. 489). Le fait que tous les regroupements ne soient pas poussés à leur maximum ne pose pas de problème : il suffit que ces regroupements soient suffisants pour faire apparaître des régularités traductionnelles de façon plus marquée (si notre hypothèse est avérée).

Pour cette lemmatisation, nous avons donc suivi les principes suivants :

– les substantifs sont ramenés au singulier, si cela n'altère pas leur signification (p. ex. on a conservé *services secrets*) ;

[184] "the Chinese translations for beauty, beautiful and beautifully are often the same."

- les adjectifs en français sont ramenés au masculin singulier ;
- les verbes sont ramenés à l'infinitif ;
- dans les expressions polylexicales, on lemmatise la partie variable (p. ex. *a lancé un appel aux* donne *lancer un appel à*) ;
- les articles, les pronoms sont de même ramenés au masculin singulier. Dans le cas des articles contractés on utilise la forme analytique (*des*, *du* donnent *de le*).

Par ailleurs, sur le plan formel, nous avons éliminé certaines variations :

- les majuscules de début de phrase sont supprimées (sauf pour les noms propres, les sigles, etc.).
- les fautes de frappe et d'orthographe sont corrigées.

Il en résulte, pour les lemmes, une réduction du nombre de types.

	Anglais	***Français***
Occurrences de lexies / lemmes	12 195	14 000
Types de lexies	3 981	4 514
Types de Lemmes	3 354	3 665

tableau 46 : *réduction des types consécutive à la lemmatisation*

III.2 Modèles d'appariement pour l'extraction des correspondances

L'extraction automatique de correspondances lexicales présuppose deux types de tâches, consécutives ou simultanées : une étape de segmentation qui aboutit à la délimitation des unités, et une étape consacrée à l'appariement de ces unités. Le plus souvent, les deux problèmes sont traités indépendamment.

Dans le présent travail, nous avons laissé de côté le problème de la segmentation automatique. De nombreuses techniques ont été développées, la plupart combinant 1/ l'extraction, à partir d'analyse locale, de syntagmes correspondant à des patrons syntaxiques jugés typiques des unités polylexicales (p. ex. les composés nominaux de la forme *sub*. *de sub*., etc.), 2/ un filtrage statistique permettant de ne retenir que les unités les

plus significatives du point de vue de la cooccurrence de leurs constituants (Christian Jacquemin, 1991 ; Didier Bourigault, 1992 ; Dunning, 1993 ; Smajda 1993 ; Daille, 1994). Ces méthodes ont été spécifiquement appliquées à l'extraction de correspondances lexicales et terminologiques (Daille, Gaussier & Langé, 1994 ; Smajda, Mc Keown & Hatzivassiloglou, 1996 ; Mc Enery, Langé, Oakes & Véronis, 1997 ; Gaussier, Hull & Aït-Mokhtar, in Véronis, 2000 §13).

Certaines techniques, purement statistiques, se basent même sur la bi-textualité pour l'étude de la polylexicalité, le critère de non-compositionnalité traductionnelle fournissant un indice pour l'extraction des unités polylexicales qui ne sont pas traduites mot à mot (D. Melamed, 1997d).

Dans la suite de notre étude, nous développerons l'autre aspect de l'extraction des correspondances lexicales, que nous jugeons central dans l'explication des phénomènes bi-textuels : l'appariement automatique des unités équivalentes.

III.2.1 Indices d'association

Il existe toute une batterie d'indices quantitatifs permettant de mesurer la « force d'association » liant deux unités. Mais la signification de ces quantités dépend beaucoup des critères employés pour définir la notion *d'association* : on peut associer des unités sur la base de leur ressemblance graphique (cf. la notion de cognat), d'une description sémantique sous la forme de réseau ou de graphe conceptuel, ou encore sur leurs similarités combinatoires lorsqu'elles apparaissent dans les mêmes co-textes. Bien sûr, ce genre d'association requiert d'expliciter à leur tour les notions de ressemblance, de combinatoire, de co-texte, etc.

Il apparaît dès lors que les quantités, loin d'être « méthodologiquement » neutres, ne représentent pas de simples *données*. Elles sont le fruit d'un certain nombre de choix et d'hypothèses, et leur interprétation ne peut faire l'économie de l'explication rigoureuse des prémisses qui ont guidé leur obtention, au risque de s'enfermer dans un raisonnement circulaire où l'on « découvrirait » dans les résultats d'une expérience les postulats qu'on y a introduits. Pour reprendre la formule de Thomas Kuhn, « les opérations et les mesures que l'homme de science entreprend dans son laboratoire ne sont pas *le donné* de

l'expérience, mais plutôt *l'acquis-avec-difficulté* » (1983 : 176), terme que nous nous empressons de revendiquer.

Examinons les hypothèses sous-jacentes à la construction des indices d'association.

III.2.1.1 modèles de cooccurrence parallèle

Il se trouve que tous les indices que nous allons étudier se basent sur le même type d'observation : la *cooccurrence parallèle*[185] *à l'intérieur des zones alignées*. Intuitivement, si deux unités sont *souvent* cooccurrentes, i.e. si elles rentrent fréquemment à l'intérieur d'un couple de segments alignés, elles sont *probablement* traductions mutuelles. Reste qu'il faudrait préciser le terme « souvent » et évaluer un peu plus précisément ce qu'on entend par « probablement ».

Mais au préalable, nous pouvons évoquer différents type de cooccurrences et différentes façons de les dénombrer : pour définir ces divers modes de calcul nous utiliserons le terme de *modèle de cooccurrence* employé par Melamed (1998).

III.2.1.1.1 Aires de cooccurrence

Les portions de texte alignées où se définissent les cooccurrences peuvent être décrites de manière générale par des aires issues du produit des deux textes. Si l'on situe en abscisse les rangs des unités du premier texte et en ordonnées les rangs des unités du second, on peut représenter chaque couple d'unités cooccurrentes par un point : ces points s'inscriront dans des surfaces délimitées que nous appelons *aires de cooccurrence*.

L'efficacité d'un modèle peut être mesurée à deux propriétés complémentaires de son aire de cooccurrence :

- minimisation du bruit : l'aire doit contenir le moins possible de points ne désignant pas des correspondances lexicales.

[185] Afin qu'il n'y ait pas de confusion avec les cooccurrences *unilingues*, à l'intérieur de chaque moitié du bi-texte, nous désignons par *cooccurrence parallèle* les cooccurrences entre deux parties alignées du bi-texte. Dans la suite de l'exposé, si nous ne précisons pas, nous nous référerons à ce type de cooccurrence.

– minimisation du silence : l'aire doit contenir le plus possible de points désignant des correspondances entre unités équivalentes.

Suivant le modèle de cooccurrence choisi, les aires peuvent prendre des formes diverses, et satisfaire plus ou moins bien aux deux critères énumérés. Nous distinguerons cinq grandes familles de modèles :

1. Aire centrée sur la diagonale (*figure 34*).

Ce type d'aire intègre tous les points dont la distance à la diagonale ne dépasse par un certain seuil. Ce modèle approximatif oblige à choisir entre bruit et silence : une bande étroite minimise le bruit, mais avec un risque important d'augmenter le silence ; et une bande trop large, englobant le chemin en entier, minimise le silence mais engendre un bruit important.

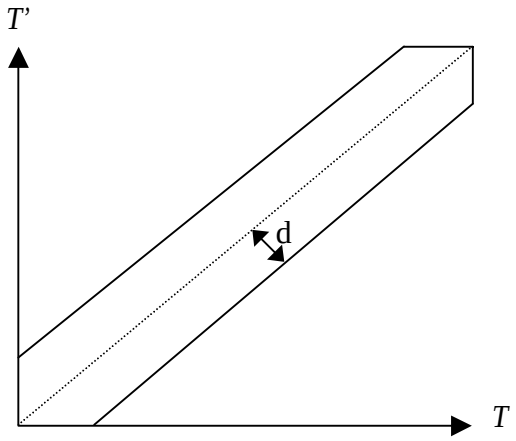


figure 33 : *aire centrée sur la diagonale*

2. Aire basée sur des segments de longueur fixe (*figure 34*).

C'est le modèle utilisé par Fung & Church (1994) dans la méthode K-vec précédemment décrite. Le calcul des points est plus simple que dans le cas précédent, mais ce modèle présente les mêmes défauts, avec probablement un silence et un bruit un peu plus important.

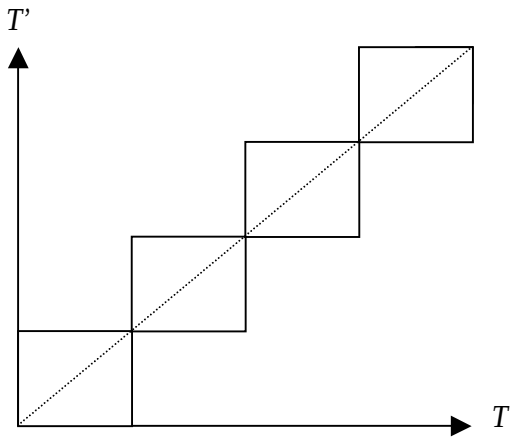


figure 34 : *aire basée sur des segments de longueur fixe*

3. Aire centrée sur un chemin d'alignement interpolé (*figure 35*).

Un tel chemin s'obtient facilement à partir d'une série de points d'ancrage (par exemple les coordonnées des premiers mots des segments alignés), reliés entre eux par des segments de droite. On considère alors tous les points situés à une distance inférieure à un certain seuil *d* du chemin interpolé. Ce type d'aire correspond au modèle basé sur la distance (« *distance-based* ») décrit par Melamed (1998). Si la densité de points d'ancrage est forte, *d* peut être assez faible, sans augmentation du silence. Les résultats dépendent alors beaucoup de la vérification de la propriété de monotonie au niveau des unités lexicales.

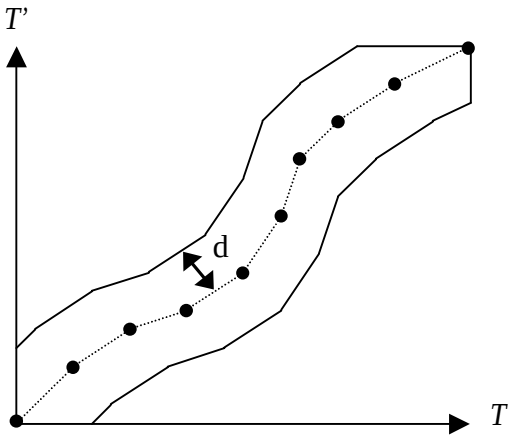
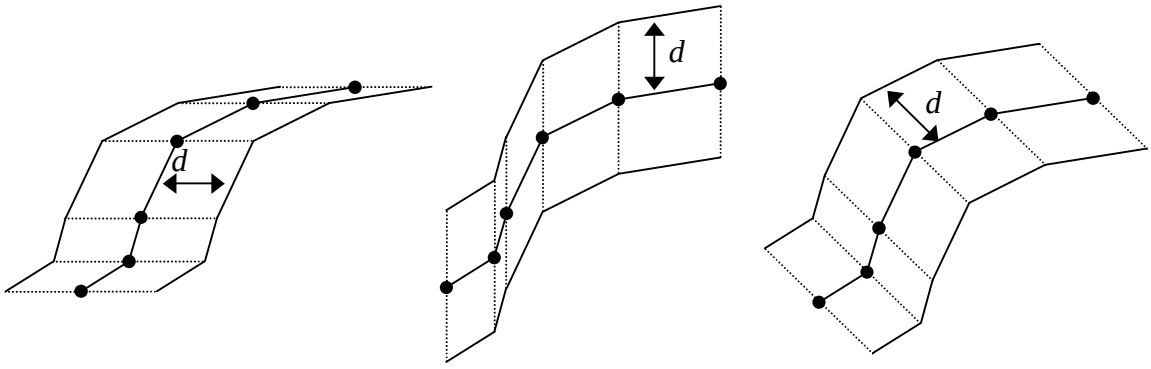


figure 35 : *aire centrée sur un chemin d'alignement*

Notons qu'il existe différentes manières de calculer la distance au chemin (figure 36) :

- la distance *horizontale* δ_1 ou la distance *verticale* δ_2, qui équivalent à la définition de fenêtres de largeur constante respectivement sur T ou sur T', centrées sur la diagonale.
- la distance *mixte* δ_3 combinant les deux directions de manière égale.
- la distance *orthogonale* δ_4, mesurant la distance de la projection orthogonale du point sur le segment de droite le plus proche.



distance horizontale *distance verticale* *distance mixte*

figure 36 : *modes de calcul de la distance à la diagonale*

- distance horizontale : pour $Y_1 \leq y < Y_2$, on a :

$$\delta_1 = \left| x - \left(X_1 + (X_2 - X_1)\frac{(y - Y_1)}{(Y_2 - Y_1)} \right) \right| < d \qquad (40)$$

- distance verticale : pour $X_1 \leq x \leq X_2$, on a :

$$\delta_2 = \left| y - \left(Y_1 + (Y_2 - Y_1)\frac{(x - X_1)}{(X_2 - X_1)} \right) \right| < d \qquad (41)$$

- distance mixte : pour $X_1 - d\frac{(X_1 - X_2)}{2} \leq x < X_2 + d\frac{(X_1 - X_2)}{2}$ on a :

$$\delta_3 = \left| \frac{(y - Y_1)}{(Y_2 - Y_1)} - \frac{(x - X_1)}{(X_2 - X_1)} \right| < d \qquad (42)$$

– la distance orthogonale s'obtient par : $\delta_4 = \frac{\delta_1 \delta_2}{\sqrt{\delta_1^2 + \delta_2^2}}$.

Son domaine de définition est d'un calcul plus complexe, car il existe une zone, entre deux segments consécutifs [(X_1,Y_1),(X_2,Y_2)] et [(X_2,Y_2),(X_3,Y_3)], où la distance peut être calculée en fonction d'un segment ou d'un autre.

4. Aire basée sur des segments alignés (*figure 37*).

Quand on dispose d'un alignement de niveau phrastique, il paraît naturel de calculer les cooccurrences à l'intérieur des segments alignés. Un alignement au niveau des phrases représente un bon compromis pour minimiser à la fois bruit et silence.

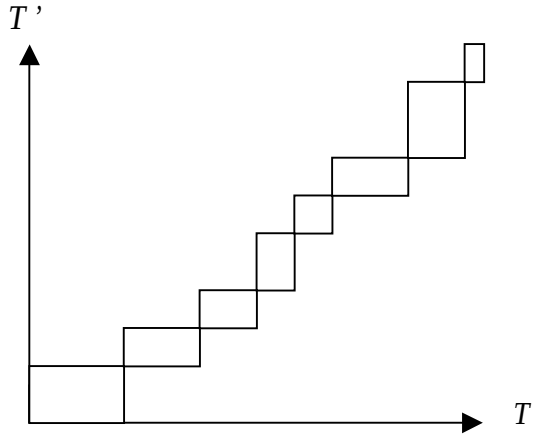


figure 37 : *aire basée sur des segments alignés*

5. Aire basée sur des segments alignés, combinée à un modèle de distance (*figure 38*).

Ce modèle mixte rappelle le « modèle de distorsion » de Gaussier & Langé (1995 : 138). Les auteurs remarquent que la position du mot cible à l'intérieur d'une fenêtre de largeur fixe, centrée autour de la position sur la diagonale, se rapproche empiriquement d'une loi de Laplace-Gauss. On peut donc déterminer la présence du mot cible dans cette fenêtre avec une probabilité supérieure à un certain seuil. Par suite, les cooccurrences sont dénombrées avec des pondérations différentes, respectivement *p* et (*1-p*), suivant que le mot cible se situe à l'intérieur ou à l'extérieur de la fenêtre.

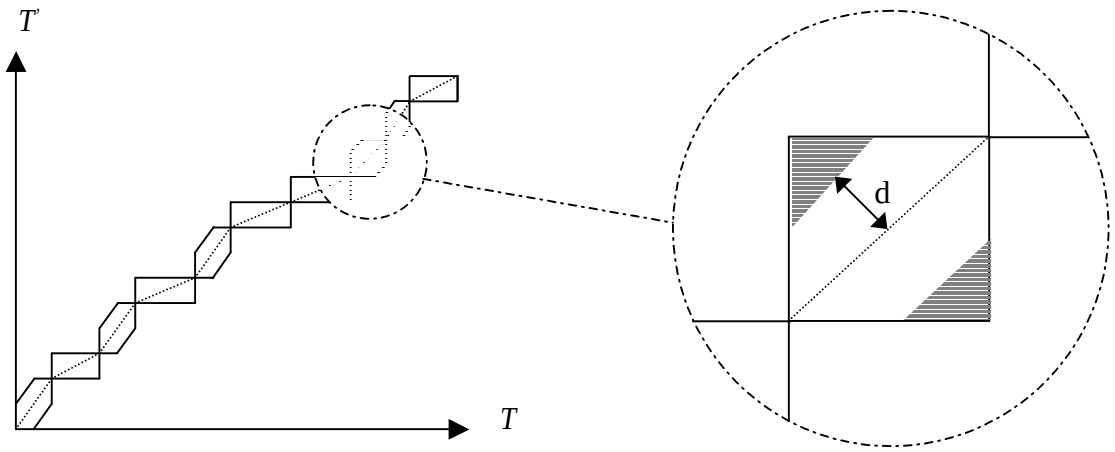


figure 38 : *aire basée sur des segments alignés combinée avec un modèle de distance*

III.2.1.1.2 Comptage des cooccurrences

Une fois déterminé le mode de calcul d'une aire de cooccurrence, il reste à traiter le problème du comptage des cooccurrences à l'intérieur de ces aires.

Plaçons-nous d'abord dans le cas où l'aire est définie par deux segments alignés *S* et *S'*. Si *Occ*(*u*) et *Occ*(*u'*) représentent respectivement les nombres d'occurrences des unités *u* et *u'* dans *S* et *S'*, on peut calculer *Cooc*(*u*,*u'*), le nombre de cooccurrences de *u* et *u'*, de plusieurs manières :

- si on compte tous les points générés par ces occurrences, on obtient :

$$Cooc(u,u') = Occ(u) \cdot Occ(u') \quad (43)$$

  En comptant de cette manière, on risque de surévaluer les cooccurrences : par exemple si le mot (fr.) *le* compte 3 occurrences en *S* et (it.) *il* en compte 4 en S, on aura : *Cooc*(*la*, *il*)=*12*.

- si on rapporte les cooccurrences aux correspondances présumées entre *u* et *u'*, on peut alors effectuer le calcul de deux manières :

$$Cooc(u,u') = min(Occ(u),Occ(u')) \quad (44)$$

$$Cooc(u,u') = max(Occ(u),Occ(u')) \quad (45)$$

Dans le mode de calcul (44), on considère que la traduction d'une occurrence de *u* doit donner lieu à *au moins* une occurrence de *u*' (en faisant l'hypothèse que *u* et *u*' sont traductions mutuelles). Dans le mode (45), on considère que la traduction d'une occurrence de *u* doit donner lieu à *au plus* une occurrence de *u*'.

Dans un modèle de cooccurrence basé sur les distances (modèles (a) et (c)), ces formules ne sont plus valides car *Occ*(…) ne peut plus être limité à un couple de segments.

Le mode de calcul (43) revient à dénombrer tous les points qui mettent en jeu le couple (*u*,*u*') :

$$Cooc(u,u') = card\{(X,Y) \in Aire \,/\, Unité(X) = u \text{ et } Unité'(Y) = u'\} \quad (46)$$

où les fonctions *Unité* et *Unité*' renvoient une unité en fonction de sa coordonnée.

En ce qui concerne (44) et (45), on peut estimer les occurrences en projetant sur l'axe des abscisses et l'axe des ordonnées tous les points impliquant *u* et *u*'.

$$occ_x(u,u') = card\{X \,/\, (X,Y) \in Aire \text{ et } Unité(X) = u \text{ et } Unité'(Y) = u'\} \quad (47)$$

$$occ_y(u,u') = card\{Y \,/\, (X,Y) \in Aire \text{ et } Unité(X) = u \text{ et } Unité'(Y) = u'\} \quad (48)$$

On peut alors appliquer les formules (43) et (44) comme précédemment.

Notons qu'il peut être intéressant de tenir compte de la surface de l'aire de cooccurrence dans le comptage : en effet, plus la surface est importante, plus faible est la proportion des points qui correspondent à des équivalences traductionnelles. On peut ainsi pondérer chaque cooccurrence par une valeur inversement proportionnelle à la surface, afin de donner plus de poids aux cooccurrences qui ont le plus de chance d'être des correspondances. C'est ce que proposent Gaussier & Langé (1995). Le nombre de cooccurrences est calculé comme la somme de l'inverse de la taille *m* du segment cible, pour tous les couples *R* de segments où apparaissent ensemble l'unité anglaise *e* et l'unité française *f* :

$$c(e, f) = \sum_R \frac{1}{m} \quad (49)$$

Pour leur « modèle de distorsion » précédemment évoqué, les auteurs remplacent *1*/*m* par :

$\frac{p}{l_f}$ si *f* appartient à la fenêtre de longueur l_f

$\frac{(1-p)}{(m-l_f)}$ si *f* n'appartient pas à la fenêtre

Comme l'a remarqué Melamed (1998), la plupart des études basées sur le comptage des cooccurrences parallèles n'ont pas donné de description rigoureuse du modèle de cooccurrence employé. Il n'existe pas, à notre connaissance, d'étude empirique centrée sur les spécificités de chacune des méthodes ici décrites. D'après les précédentes considérations qualitatives, nous pensons qu'un modèle de cooccurrence basé sur des phrases alignées, avec un mode de calcul de type (43) ou (44), constitue un bon compromis pour minimiser à la fois le bruit et le silence.

III.2.1.1.3 Dynamic time warping

La méthode DKvec développée par Fung (1995a) intègre une technique originale de calcul des cooccurrences, sans passer par la définition d'une aire de cooccurrence. Le principe en est simple : la distribution d'un mot peut être représentée par le vecteur de ses coordonnées dans le texte. Comparer deux unités *u* et *u*' sur la base de ces vecteurs (resp. $P=(p_i)_{i=1...n}$ et $P'=(p_j)_{j=1..m}$) peut poser problème, dans la mesure où les positions dans *T* et *T*' ne se correspondent pas directement. C'est pourquoi ces vecteurs sont transformés en *vecteurs de différence de position* (« *positional difference vector* »), où figure pour chaque occurrence non plus sa position mais la différence entre celle-ci et la position de l'occurrence suivante : les vecteurs deviennent respectivement $V=(v_i)_{i=1...n}$ et $V'=(v'_j)_{j=1..m}$ avec $v_i=p_i-p_{i-1}$ et $v'_j=p_j-p_{j-1}$. L'auteur se base sur l'idée que « tandis que des mots similaires n'apparaissent pas exactement à la même position dans chaque moitié du corpus, les écarts entre chaque occurrence du même mot sont similaires d'une langue à l'autre »[186] (Fung in J. Véronis, 2000 §11). Sous cette forme, chaque mot peut être représenté par un « signal » représentant les variations de ces écarts, avec des pics plus ou moins marqués, et l'on constate expérimentalement que des mots équivalents présentent effectivement un signal

similaire. Reste ensuite à comparer ces signaux, c'est-à-dire à établir une mesure permettant d'évaluer la distance entre deux vecteurs : celle-ci est calculée comme la longueur du chemin optimal. Ce chemin est constitué de *m* positions successives $i_{1...}i_m$ de *u* correspondant aux *m* positions de *u*' telles que la somme des différences des $|v'_j-v_{i_j}|$ soit minimale. Ce chemin est calculé grâce à un algorithme dynamique, le DTW (pour « *Dynamic Time Warping* »), avec une complexité en $O(n \cdot m)$.

Cette manière de compter la cooccurrence est doublement intéressante, dans la mesure où elle n'implique aucun alignement préalable[187], et où elle débouche directement sur une mesure (la longueur du chemin, i.e. la somme des $|v'_j-v_{i_j}|$) permettant de quantifier la ressemblance entre *u* et *u*'.[188]

III.2.1.2 Les mesures d'association

Examinons maintenant les mesures statistiques qui intègrent les comptes issus des modèles de cooccurrence présentés précédemment.

III.2.1.2.1 Principe

Toutes ces mesures mettent en œuvre le même principe : si le nombre de cooccurrences observé est très supérieur au nombre de cooccurrences estimé dans le cas d'une distribution aléatoire, c'est qu'il y a sans doute un *lien* entre les deux unités.

En effet, dans la mesure où dans un même co-texte (p. ex. une phrase) on ne peut présager des unités qui vont apparaître ensemble, on observe qu'entre des segments alignés de nombreuses cooccurrences ne sont que des coïncidences, imputables au hasard. Pour

---

[186] "It is based on the notion that while similar words do not occur at the exact same position in each half of the corpus, distances between instances of the same word are similar across language"

[187] En fait les points du chemin optimal, après filtrage, permettent de constituer des points d'ancrage. On peut noter une certaine similitude entre cette méthode et nos propres méthodes de préalignement, sauf que dans ces dernières, les vecteurs qui sont comparés ont toujours la même dimension, correspondant au découpage issu de l'étape précédente.

[188] En utilisant cette mesure pour établir des correspondances, Fung obtient une précision de 55,35 % pour les termes techniques d'un petit corpus anglais / japonais, et de 89,9 % sur un corpus anglais / chinois de plus grande taille.

reprendre un exemple précédent, la cooccurrence de *banknote* avec *aveugle* n'est liée à aucune forme de déterminisme.

Ainsi, à partir des distributions de chaque unité à l'intérieur de son propre corpus, on peut estimer le nombre de cooccurrences fortuites attendu. Quand le nombre des cooccurrences observées contredit de façon significative cette attente, l'hypothèse de l'indépendance des distributions est à rejeter.

Comment interpréter ce *lien* de dépendance[189] statistique ? signifie-t-il qu'il y a une relation de traduction ? Sans doute : l'équivalence traductionnelle est la seule forme de corrélation existant entre les deux parties du bi-texte. Mais cette relation peut être plus ou moins directe. On peut distinguer trois cas de figure :

- *correspondance complète* : les deux unités sont employées comme équivalent traductionnel.
- *correspondance partielle* : les unités font partie d'un couple d'équivalents traductionnels mettant en jeu des unités polylexicales plus larges.
- *correspondance indirecte* (ou fausse correspondance) : une des deux unités entretient un lien de dépendance statistique, sur l'axe syntagmatique, à l'intérieur de sa propre partie du bi-texte, avec le correspondant de l'autre unité.

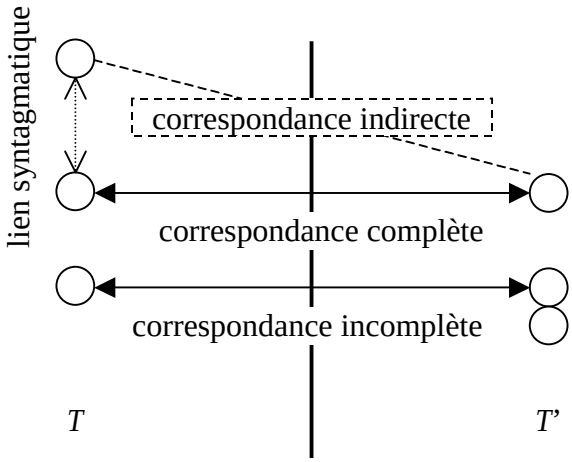


figure 39 : *les divers types de correspondance*

[189] Ici le terme *dépendance* est à entendre sur le plan statistique, en tant qu'il s'oppose au terme indépendance : il n'implique pas de nécessité dans l'enchaînement des occurrences, mais seulement des corrélations.

Comme le montre le schéma de la figure 39, la dépendance statistique concerne deux axes orthogonaux : syntagmatique (intralinguistique) et traductionnel (interlinguistique). La dépendance statistique de deux unités de *T* et *T'* peut donc être le résultat d'une *correspondance indirecte*, qui conjugue un lien de dépendance syntagmatique et un lien de dépendance traductionnelle.

Dans une étude précédente (Kraif, 1995 : 101), sur la base d'un bi-texte anglais - italien, nous avons montré que le phénomène des correspondances indirectes pèse de manière significative sur les mesures d'association. Par exemple, en nous basant sur l'information mutuelle (cf. supra), nous avons observé les associations suivantes :

Unités associées à medals	***Indice d'association***
smalti (« émaux »)	0,9987
vetri (« verrerie »)	0,9987
medaglie (« médailles »)	0,9987
stoffe (« étoffes »)	0,9987
argenterie (« argenterie »)	0,9987

tableau 47 : *effets des correspondances indirectes sur un indice*

On constate que l'indice ne permet pas de départager la correspondance directe de (angl.) *medals* avec (it.) *medaglie*, des correspondances indirectes, erronées. Ceci est dû au fait que *medals* apparaît toujours, dans le corpus, au milieu des mêmes énumérations, dans des phrases du type :

> *The Museum houses paintings from the thirteenth to eighteenth centuries, marble and carved-wood sculptures, bronzes, terracottas, pottery, china, silver, cloths, seals, medals, glassware, tapestries, enamels, etc.*

Par suite, les unités énumérées sont statistiquement liées sur l'axe syntagmatique : ce lien se compose avec le lien traductionnel et fait émerger des correspondances indirectes.

Bien entendu, plus le corpus est vaste, et plus les associations indirectes doivent s'affaiblir, en proportion, car la liberté combinatoire de l'axe syntagmatique finit par en diluer les effets. Mais des liens privilégiés peuvent néanmoins demeurer, pour des raisons sémantiques et thématiques, comme le montrent les cooccurrences du tableau 48 (extraites du corpus JOC).

u	*u*'	*Occ*(*u*)	*Occ*'(*u*')	*Cooc*
Amnesty International	*Amnesty International*	29	26	26
Amnesty International	*rapport*	29	892	12
Amnesty International	*droits de l' homme*	29	544	8
Amnesty International	*avril*	29	502	6
Amnesty International	*torture*	29	49	6
Amnesty International	*cas*	29	1211	3
Amnesty International	*mesures*	29	2147	3
Amnesty International	*exactitude*	29	12	2
Amnesty International	*prisons*	29	11	2

tableau 48 : *cooccurrences de l'unité* Amnesty International
(*sauf les mots outils et les unités cooccurrentes une seule fois*)

Pour utiliser correctement les mesures de dépendance statistique, en vue d'extraire des correspondances, il faudra par conséquent maîtriser les effets des correspondances indirectes.

III.2.1.2.2 Les mesures classiques

Nous avons déjà indiqué (§ II.2.4.2.1) deux mesures permettant de quantifier la force de l'association statistique entre deux unités *u* et *u*' : l'*information mutuelle* (notée IM) et le *t-score* (noté TS):

$$IM = \log\left(\frac{p_{12}}{p_1 p_2}\right) \qquad t \approx \frac{p_{12} - p_1 p_2}{\sqrt{\frac{p_{12}}{n}}} \tag{50}$$

avec $p_1 = \frac{n_1}{n}$ $p_2 = \frac{n_2}{n}$ $p_{12} = \frac{n_{12}}{n}$

où n_{12} est le nombre de cooccurrences de *u* et *u*'
n_1 le nombre d'occurrences de *u*
n_2 le nombre d'occurrences de *u*'
n le nombre total de segments

Dunning constate que les modèles basés sur l'hypothèse d'une distribution normale des occurrences ne sont pas adaptés à l'étude des événements rares. Dans ces modèles, comme avec l'information mutuelle, la cooccurrence de deux hapax devient hautement

improbable. Or, Dunning (1993 : 2) remarque que les événements « rares » sont plutôt fréquents :

> « Quand on compare les taux d'occurrence d'événements rares, [l'hypothèse de normalité sur laquelle se basent certains tests ne tient plus] car les textes sont largement composés de tels événements rares. Par exemple, un simple comptage de mots effectué sur un corpus de taille modérée montre que les mots ayant une fréquence de 1 sur 50 000 constituent environ 20-30 % des bulletins d'informations écrits en anglais courant. Cette tranche « rare » de l'anglais inclut de nombreux mots pleins, et presque tout le jargon technique. »[190]

Pour modéliser de façon plus précise ce genre d'événement, Dunning propose d'assimiler les occurrences de chaque unité au résultat d'un test de Bernoulli : on considère qu'en choisissant une unité au hasard, on a une probabilité constante *p* d'obtenir l'unité *u* (bien entendu *p* dépend de *u*). L'événement « *u* apparaît *k* fois » est interprété comme le résultat de *n* tirages indépendants, avec à chaque fois la probabilité *p* d'obtenir *u* (à la manière de *n* tirages d'une pièce de monnaie).

Le paramètre *p* étant donné, la fonction de vraisemblance *H* d'observer les *k* fois *u* sur *n* tirages suit une loi binomiale :

$$H(p;n,k) = C_n^k p^k (1-p)^{n-k} \tag{51}$$

pour deux distributions observées (k_1,n_1) et (k_2,n_2) de paramètres respectifs p_1 et p_2, on a :

$$H(p_1 p_2; n_1, k_1, k_2, n_2) = C_{n_1}^{k_1} {p_1}^{k_1} (1-p_1)^{n_1-k_1} C_{n_2}^{k_2} {p_2}^{k_2} (1-p_2)^{n_2-k_2} \tag{52}$$

Le rapport de vraisemblance lié à une hypothèse est le rapport entre la valeur maximale atteinte par *H* dans le sous-espace décrit par les paramètres liés à cette hypothèse, et la valeur maximale atteinte par *H* sur tout l'espace décrit par les paramètres (A. Mood *et al.*, 1974).

Le rapport de vraisemblance lié au test de l'hypothèse « $p_1 = p_2$ » est donc :

$$\lambda = \frac{\max_{p} H(p, p; k_1, n_1, k_2, n_2)}{\max_{p_1 p_2} H(p1, p2; k_1, n_1, k_2, n_2)} \quad (53)$$

Ces maxima sont atteints avec :

$p_1 = \frac{k_1}{n_1}$ $p_2 = \frac{k_2}{n_2}$ pour le dénominateur et $p = \frac{k_1 + k_2}{n_1 + n_2}$ pour le numérateur

Si l'on prend le logarithme du rapport de vraisemblance, on a :

$$-2\log\lambda = 2\left[\log p_1^{k_1}(1-p_1)^{n_1-k_1} + \log p_2^{k_2}(1-p_2)^{n_2-k_2} - \log p^{k_1}(1-p)^{n_1-k_1} - \log p^{k_2}(1-p)^{n_2-k_2}\right] \quad (54)$$

On peut maintenant ramener ce modèle au problème des occurrences de deux unités *u* et *u*'. En considérant que p_1 représente la probabilité d'occurrence de *u* avec *u*' et p_2 la probabilité d'occurrence de *u* en l'absence de *u*', on teste alors l'hypothèse $p_1 = p_2$ traduisant l'indépendance des occurrences des deux unités *u* et *u*'. Nous appellerons cette hypothèse *l'hypothèse nulle*.

Soient les valeurs de la table de contingence des deux unités :

	u'	***\ u'***
u	*a*	*b*
\ u	*c*	*d*

On a alors :

$k_1 = a$ $k_2 = b$ $n_1 = c + a$ $n_2 = b + d$ $n = a + b + c + d$

$p_1 = \frac{a}{c + a}$ $p_2 = \frac{b}{b + d}$ $p = \frac{a + b}{n}$

En notant RV l'indice résultant du rapport de vraisemblance, l'équation (54) donne, après simplification :

---

[190] "When comparing the rates of occurrence of rare events, the assumptions on which these tests are based break down because texts are composed largely of such rare events. For example, simple word counts made on a moderate sized corpus show that words which have a frequency of less than one in 50,000 words make up about 20-30 % of typical English language newswire reports. This 'rare' quarter of English includes many of the content bearing words, and nearly all the technical jargon."

$$RV = -2\log\lambda = 2(S^{+} - S^{-})$$
$$S^{+} = a\log a + b\log b + c\log c + d\log d + n\log n \quad (55)$$
$$S^{-} = (a+c)\log(a+c) + (b+d)\log(b+d) + (a+b)\log(a+b) + (c+d)\log(c+d)$$

Cette valeur sera d'autant plus importante que l'hypothèse nulle pour *u* et *u*' sera moins vraisemblable.

Comme l'ont montré Gaussier & Langé (1995), le test du rapport de vraisemblance paraît plus adapté à l'extraction des correspondances lexicales. Là où l'information mutuelle fournit des correspondances avec un rappel de 76 % et une précision de 56 %, le test de vraisemblance, que nous noterons RV, atteint un rappel de 84 % avec une précision de 64 %. C'est actuellement un des indices les plus utilisés pour l'exploitation des cooccurrences (Melamed, 1998a).

III.2.1.2.3 Probabilité de l'hypothèse nulle

Afin de fournir une base de comparaison pour l'évaluation de ces différents indices, nous en proposons un quatrième, basé sur le simple calcul de la probabilité P_0 de cooccurrence en faisant l'hypothèse nulle (i.e. en supposant que toutes les cooccurrences sont aléatoires).

Pour calculer cette probabilité on dénombre toutes les combinaisons avec n_{12} cooccurrences, et l'on divise par toutes les combinaisons possibles. On effectue une série de tirages successifs :

- on tire n_1 occurrences de *u* : $C_n^{n_1}$ possibilités ;
- sur ces n_1 occurrences, on tire de n_{12} cooccurrences de *u* avec *u*' : $C_{n_{12}}^{n_2}$ possibilités ;

– enfin, on tire les n_2-n_{12} occurrences de *u*' restantes, parmi les n-n_1 unités non encore tirées : $C_{n-n_1}^{n_2-n_{12}}$ possibilités.

Au total, on dénombre $C_n^{n_1} C_n^{n_2}$ combinaisons possibles[191].

En reprenant les notations précédentes, on a :

$$P_0(n_{12} / n, n_1, n_2) = \frac{C_n^{n_1} C_{n_1}^{n_{12}} C_{n-n_1}^{n_2-n_{12}}}{C_n^{n_1} C_n^{n_2}} \tag{56}$$

Après simplifications, on trouve, sous une forme plus facilement calculable :

$$P_0(n_{12} / n, n_1, n_2) = \prod_{k=1}^{n_2-n_{12}} \frac{(n-n_1-n_2+n_{12}+k)}{(n-n_2+n_{12}+k)} \prod_{k=1}^{n_{12}} \frac{(n_1-n_{12}+k)(n_2-n_{12}+k)}{k(n-n_2+k)} \tag{57}$$

Plus cette probabilité sera faible, plus l'association entre *u* et *u*' doit être forte. On en déduit l'indice I_0 :

$$I_0 = -\log(P_0) \tag{58}$$

III.2.2 Architectures

Les algorithmes permettant d'extraire les correspondances. Nous exposons ici les plus représentatifs.

III.2.2.1 Association maximale

Lorsqu'on dispose d'un indice d'association, c'est la méthode la plus simple et la plus directe : étant donné une occurrence *u*, on cherche parmi toutes les unités cooccurrentes, celle qui réalise le maximum de l'indice :

[191] Pour simplifier les calculs, on effectue une légère approximation : on fait comme si chaque unité ne pouvait pas apparaître plus d'une fois dans un même segment : dès lors, on est sûr que $n \geq n_1$, $n \geq n_2$ et $n \geq n_1 + n_2 - n_{12}$. Par ailleurs, on néglige l'effet de la taille des segments : il clair que l'occurrence d'une unité est plus probable dans un grand segment que dans un petit – et il serait intéressant, dans un modèle plus élaboré, de tenir compte aussi des longueurs de segments.

$$Corresp(u) = \arg\max_{u' \in C(u)}(Assoc(u,u')) \quad (59)$$

où *Assoc*( ) est une mesure d'association (du type IM, TS, RV, etc.).

C(*u*) est l'ensemble des unités parallèlement cooccurrentes avec *u* dans le modèle de cooccurrence (par exemple toutes les unités du segment aligné correspondant). On notera réciproquement *C*'(*u*') l'ensemble des unités parallèlement cooccurrentes avec *u*'.

Il peut être intéressant d'effectuer un premier filtrage sur les résultats obtenus : par exemple, en ne retenant que les correspondances dont la mesure d'association est supérieure à un certain seuil. L'élimination du bruit permet ainsi d'augmenter la précision sans une forte chute du rappel.

Gaussier et Langé (1995) proposent de rajouter une condition de réciprocité : la valeur maximale atteinte avec *u*' doit aussi être la valeur maximale pour *u*' dans son association avec toutes les unités *w* de *C*'(*u*') :

$$Corresp(u) = \arg\max_{u' \in C(u)}(assoc(u,u')) \text{ et } u = \arg\max_{w \in C'(u')}(Assoc(w,u')) \quad (60)$$

Ce critère de réciprocité du maximum permet de limiter les distorsions dues aux hapax, qui présentent une dépendance statistique importante avec toutes les unités cooccurrentes.

III.2.2.2 Meilleure affectation biunivoque

L'algorithme précédent pose cependant un problème : l'application définie par *Corresp*(*u*) n'est pas injective : deux unités u_1 et u_2 peuvent avoir la même image *u*'. Ceci pourrait certes convenir dans le cadre de correspondances polylexicales, mais nous nous situons dans la perspective où de telles unités ont été préalablement définies et identifiées, et ne doivent pas découler de l'appariement. En outre, l'algorithme implique une dissymétrie entre les deux côtés du bi-texte, puisque la non-injectivité n'est possible que dans le sens $T \rightarrow T'$.

On peut pallier ce problème en imposant la biunivocité des correspondances : pour ce faire, on procède par étapes, en sélectionnant successivement le couple obtenant le meilleur

indice d'association. Une fois appariées, les unités sont mises de côté, car on considère qu'une même unité ne peut être réutilisée dans un autre couple[192]. On peut alors procéder à nouveau à l'élection du meilleur couple. Cet algorithme itératif peut être décrit comme suit :

1. *Initialisation*.
On constitue l'ensemble des candidats *Cand* : on calcule les indices d'association pour tous les couples (*u*,*u*') des phrases (*P*,*P*'). Tous ces couples sont placés dans *Cand*.

2. *Sélection* :
On sélectionne un couple (*u*,*u*') de *Cand* obtenant la meilleure valeur de l'indice. On retient ce couple dans l'ensemble *Corresp* contenant le cumul des correspondances déjà obtenues, et on élimine de *Cand* tous les couples qui mettent en jeu *u* ou *u*'.

3. *Retour* en 2 tant que l'ensemble *Cand* n'est pas vide.

4. *Terminaison*.
Corresp contient le résultat.

Cet algorithme peut être mis en œuvre de façon simple et efficace, en triant par leur score (i.e. l'indice) tous les couples candidats dans un arbre binaire de recherche.

Les correspondances ainsi extraites sont biunivoques, et lorsque les deux phrases ne contiennent pas le même nombre d'unités, la phrase la plus longue contiendra nécessairement des occurrences résiduelles qui n'ont pu être appariées.

L'heuristique de cet algorithme est similaire à *l'heuristique de précision d'abord* : on intègre d'abord les couples les plus sûrs, dont les chances d'être des correspondances correctes sont maximales. L'élimination progressive des unités correctement appariées permet de réduire peu à peu l'espace de recherche, et de minimiser les chances d'erreurs. Nous avions mis en œuvre ce type d'algorithme lors de la seconde campagne d'évaluation du projet ARCADE, où nous utilisions les correspondances lexicales pour l'alignement phrastique.

[192] Il s'agit là d'une simplification par rapport à la forme que nous avons imposée aux correspondances de référence établies manuellement.

III.2.2.3 Algorithme EM

Toute une famille de système dérive de l'algorithme EM (pour *Expectation-Maximisation*) développé par Dempster *et al*. (1977). Dans ces méthodes, on établit tout d'abord un modèle probabiliste permettant d'évaluer la probabilité de générer telle phrase cible à partir de telle phrase source. Ces modèles sont dits paramétriques, car les estimations des probabilités de traduction des phrases sont basées sur tout un ensemble de paramètres de base, comme les probabilités de traduire une unité lexicale par une autre unité lexicale. Le but de l'algorithme est de trouver un ensemble de paramètres maximisant la probabilité globale de traduction d'un corpus d'entraînement. Dans un cadre itératif, moyennant la réalisation de certaines conditions mathématiques, l'algorithme assure la convergence vers les paramètres optimaux. Les correspondances lexicales peuvent ensuite être déduites de ces paramètres.

Ces modèles se basent sur un paradigme que nous avons contesté : l'*alignement* des unités lexicales. En effet, dans la modélisation du processus de génération de la phrase cible, les unités cibles sont censées dériver des unités sources, comme dans une traduction mot à mot. Même si ces modèles, nous le verrons, permettent de considérer l'alignement de groupes de mots avec d'autres groupes de mot, et d'intégrer de façon marginale les ajouts et les omissions, ils parient sur la biunivocité des relations entre les unités sources et cibles. Dans un article très complet, Melamed (1998a) présente ces méthodes sous le titre : *Word-to-Word Models of Translational Equivalence* (« modèles d'équivalence traductionnelle mot à mot »). Vis-à-vis de la pratique de la traduction, nous avons démontré la non-pertinence d'une telle conception. Mais nous pensons néanmoins que ces modèles ont une valeur opératoire, dans la mesure où ils reflètent des phénomènes émergeant sur un plan statistique : s'il est vrai qu'il n'y a pas de déterminisme dans la traduction d'une unité, on peut observer des régularités sur un grand nombre de traductions. Ces régularités découlent naturellement du fait que des unités lexicales appartenant à des systèmes différents, et ayant parfois des significations différentes, ont des référents communs.

Ainsi, dans la mesure où les modèles qui suivent présentent la traduction mot à mot sous un angle probabiliste, ils restent fidèles à la réalité empirique. En d'autres termes, ils

s'adressent à un autre niveau de description, qui n'est pas celui de la traduction en tant que *processus*, mais celui de la masse des *traductions produites*. Même si cette distinction n'est pas toujours clairement établie, par les concepteurs mêmes de ces modèles, elle nous paraît fondamentale.

Dans l'implémentation de l'algorithme EM, tout comme dans celle de l'algorithme de meilleure affectation biunivoque précédemment décrit, le critère de biunivocité présente un avantage pratique certain : il permet de diminuer l'effet des cooccurrences indirectes, dans la mesure où l'affectation de la correspondance la plus probable (*u*,*u*') inhibe les autres correspondances indirectes avec *u* ou *u*'.

Remarquons enfin que ce critère de biunivocité possède une certaine souplesse, dans la mesure où il peut s'appliquer à différents types d'unité. Comme le note Melamed (1998a :5) :

> « L'hypothèse de biunivocité[193] n'est pas aussi restrictive qu'elle le paraît : on peut toujours améliorer le pouvoir explicatif d'un modèle basé sur cette hypothèse, par une redéfinition de ce qu'on entend par mot. »[194]

III.2.2.3.1 Modèles de Brown et al.

P. Brown *et al*. (1993) ont élaboré une des versions les plus complètes de l'application de l'algorithme EM. Etant donné sa puissance et sa généralité, nous en proposons une description schématique, en adaptant les notations des auteurs à nos propres conventions.

L'objectif des auteurs est de construire un modèle probabiliste permettant la traduction automatique. Pour tout couple de phrases *P* et *P'*, le modèle doit permettre d'estimer $p(P'/P)$, la probabilité de traduire la phrase *P* par la phrase *P'*. Dès lors, trouver une bonne traduction de la phrase *P*, revient à trouver une phrase *P'* maximisant la probabilité $p(P'/P)$. En utilisant le théorème de Bayes, on a :

---

[193] Pour désigner cette « hypothèse de biunivocité » nous préférons quant à nous le terme de *critère de biunivocité*, car nous ne faisons aucune hypothèse quant à la biunivocité des unités de traduction.

[194] "The one-to-one assumption is not as restrictive as it may appear: The explanatory power of a model based on this assumption may be raised to an arbitrary level by redefining what words are."

$$p(P'/P) = \frac{p(P').p(P/P')}{p(P)}$$

par suite, une bonne traduction de *P* sera définie par :

$$\hat{P}' = \arg\max_{P'} p(P'/P) = \arg\max_{P'} \frac{p(P').p(P/P')}{p(P)} = \arg\max_{P'} p(P').p(P/P') \quad (61)$$

cette dernière équation étant d'après les auteurs « l'équation fondamentale de la traduction automatique ». On s'écarte du modèle classique analyse - génération qui s'inspirait des deux étapes de compréhension - réécriture identifiées dans la traduction humaine. Mais Brown *et al*. remarquent que d'un point de vue formel, cette équation est adaptée au traitement automatique. Le problème est ainsi décomposé en trois sous-tâches, dont les deux premières sont indépendantes :

1- estimation des paramètres du *modèle linguistique p*(*P'*)

2- estimation des paramètres du *modèle traductionnel p*(*P*/*P'*)

3- génération d'une phrase *P'* maximisant le produit des deux probabilités

Le problème 1 a été abondamment traité dans le champ de la reconnaissance de la parole, notamment avec les modèles markoviens. Le problème 3 est encore peu exploré : les auteurs proposent de générer la traduction de manière itérative, en testant les hypothèses les plus probables, mot par mot. Brown *et al*. se concentrent essentiellement sur le problème 2, i.e. l'élaboration d'un modèle de traduction.

Pourquoi proposent-ils de calculer *p*(*P*/*P'*) au lieu de calculer directement *p*(*P'*/*P*) ? c'est que le modèle de traduction, centré sur la traduction des unités lexicales, ne prend pas en compte le fait qu'une phrase soit bien formée ou non. La maximisation de *p*(*P'*/*P*) mettrait sur un pied d'égalité des phrases grammaticales et des phrases sans structure (simples assemblages de mots). La séparation du modèle de traduction et du modèle linguistique permet de distinguer les deux tâches : trouver un assemblage de mots qui maximise les chances d'un point de vue traductionnel, et trouver un assemblage de mots qui soit syntaxiquement correct dans la langue d'arrivée.

On considérera donc la traduction dans le sens *P'*$\rightarrow$*P*, *P'* étant la source et *P* la cible.

La traduction de *P'* par *P* est conçue comme un processus génératif, constitué de tirages aléatoires successifs. Les unités de *P* sont supposées être générées une par une *en*

connexion avec des unités de *P'*. Un ensemble de connexions entre les unités de *P* et les unités de *P'* est appelé un *alignement A* du couple *P,P'*. La même phrase *P* pouvant être générée suivant plusieurs alignements, on a :

$$p(P/P') = \sum_{A} p(P, A/P') \tag{62}$$

Si les phrases *P* et *P'* contiennent respectivement *l* et *m* mots, alors un alignement *A* peut être représenté comme une suite (a_1,a_2,.., a_m) où chaque a_j est un nombre compris entre 0 et *l*, indiquant que le $j^{ème}$ mot de *P* est connecté avec le $i^{ème}$ mot de *P'* (la connexion avec 0 signifie que le $j^{ème}$ mot est issu d'une « génération spontanée »).

En notant $a_1^m = (a_1 a_2 .. a_m)$ l'alignement et $u_1^m = (u_1 u_2 .. u_m)$ la suite des unités constituant *P*, la formule générale du tirage aléatoire de *P* et *A* est :

$$p(P, A/P') = p(m/P') \prod_{j=1}^{m} p(a_j / a_1^{j-1}, u_1^{j-1}, m, P').p(u_j / a_1^{j-1}, u_1^{j-1}, m, P') \tag{63}$$

On tire d'abord la taille de la phrase *P*, puis les couples (unité, connexion), les uns après les autres. Remarquons que ce modèle n'est pas symétrique dans l'établissement des connexions : un même mot source peut être connecté avec plusieurs mots cibles, mais la réciproque n'est pas vraie.

Sur la base de ce tirage, les auteurs développent progressivement cinq modèles, en effectuant des simplifications importantes qu'ils abandonnent progressivement. Nous ne développerons ici que les deux premiers, plus simples, afin d'illustrer les étapes de l'algorithme EM.

III.2.2.3.2 Modèle 1

Dans le premier modèle, on fait les suppositions suivantes :

- $p(m/P') = \varepsilon$ est une constante
- le tirage de a_j ne dépend que de la longueur de la phrase *P'* :

$$p(a_j / a_1^{j-1}, u_1^{j-1}, m, P') = \frac{1}{l+1} \tag{64}$$

– le tirage de u_j ne dépend que de u_j et u'_{aj} (l'unité qui lui est connectée dans P'). On assimile alors la probabilité de ce tirage à la probabilité de traduction de u'_{aj} par u_j :

$$p(u_j / a_1^{j-1}, u_1^{j-1}, m, P') = t(u_j / u'_{a_j}) \tag{65}$$

Ces approximations sont bien sûr très éloignées de la réalité, car les positions des unités correspondantes (modélisées par a_j) ne sont pas distribuées de façon aléatoire. En outre, ce modèle développe un processus de génération mot à mot : même si un même mot source peut être connecté avec plusieurs mots cibles, les générations de ceux-ci sont supposées être indépendantes les unes des autres.

L'équation (63) devient :

$$p(P, A / P') = \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} t(u_j / u'_{a_j}) \tag{66}$$

Si l'on tient compte de tous les alignements possibles, qui attribuent à chaque a_j une valeur entre 1 et m, on a :

$$p(P / P') = \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^{l} \ldots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(u_j / u'_{a_j}) \tag{67}$$

Les probabilités conditionnelles $t(u_j/u'_{aj})$ sont les paramètres à découvrir.

Pour chaque couple de phrases du corpus d'apprentissage, on peut estimer $c(u/u',P,P')$, le nombre de fois qu'une unité u est susceptible de traduire une unité u' :

$$c(u / u', P, P') = \frac{t(u / u')}{\sum_{j=1}^{l} t(u / u'_j)} Occ(u, P) \cdot Occ(u', P') \tag{68}$$

où $Occ(u,P)$ représente le nombre d'occurrences de u dans P.

Le total du nombre de fois que u traduit u', sur tout le corpus est donc donné par :

$$c(u / u') = \sum_{\substack{(P,P') \\ alignés}} \frac{t(u / u')}{\sum_{k=1}^{l} t(u / u'_k)} Occ(u, P) \cdot Occ(u', P') \tag{69}$$

Dès lors on peut réestimer les paramètres *t*(*u*/*u*') sur la base des comptes *c*(*u*/*u*'), en normalisant par le total des comptes pour toutes les unités *w* du vocabulaire du texte T :

$$t(u/u') = \frac{c(u/u')}{\sum_{w \in T} c(w/u')} \tag{70}$$

L'algorithme est basé sur un raisonnement circulaire : on calcule les comptes en fonction des probabilités, puis l'on réestime les probabilités en fonction des comptes. Chaque paramètre consolide l'autre. Les auteurs décrivent l'application de l'algorithme EM par une suite d'itérations :

1. *Initialisation* :
On initialise tous les paramètres *t*(*u*/*u*') à une valeur identique

2. *Phase E* : *« Expectation »*.
On calcule les comptes *c*(*u*/*u*') pour tous les couples (*u*,*u*')

3. *Phase M* : *« Maximisation »*.
On réestime les probabilités *t*(*u*/*u*') à partir des comptes *c*(*u*/*u*')

4. *Retour à la phase E* jusqu'à stabilité des paramètres *t*(*u*/*u*').

Brown *et al*. montrent que l'algorithme satisfait les conditions de convergence énoncées par Dempster *et al*. : quelles que soient les valeurs initiales (non nulles) choisies pour *t*(*u*/*u*'), l'application alternée des phases E et M entraîne la convergence vers les paramètres optimums, maximisant la probabilité globale du corpus d'apprentissage. La probabilité globale étant une fonction convexe des probabilités *t*(*u*/*u*') (qui sont contraintes, rappelons le, par une condition de normalisation), on est assuré de la convergence vers la probabilité maximum.

A partir des paramètres ainsi obtenus, on peut dériver, pour chaque couple de phrases, « l'alignement de Viterbi », c'est-à-dire l'alignement qui maximise la probabilité :

$$\mathsf{A} = \arg\max_A p(P, A/P') = \arg\max_A \prod_{j=1}^{m} t(u_j / u'_{a_j}) \tag{71}$$

Les correspondances découlent alors des connexions établies par cet alignement de Viterbi.

III.2.2.3.3 Modèle 2

Dans le second modèle, on suppose que la probabilité $p(a_j / a_1^{j-1}, u_1^{j-1}, m, P')$ dépend de a_j, j, m et l. En d'autres termes, à la différence du modèle 1, on suppose que les connexions sont également liées à la position des mots dans la phrase source.

On calcule ainsi la probabilité que la position a_j soit connectée à la position j, dans un couple de phrases de longueurs respectives m et l :

$$p(a_j / a_1^{j-1}, u_1^{j-1}, m, P') = a(a_j / j, m, l) \tag{72}$$

Par suite, la probabilité à maximiser, étant donné P', devient :

$$p(P / P') = \varepsilon \sum_{a_1=0}^{l} \ldots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(u_j / u'_{a_j}) a(a_j / j, m, l) \tag{73}$$

Cette fois, il y a donc deux familles de paramètres à estimer : les probabilités lexicales $t(u/u')$ et les probabilités de position $a(i/j,m,l)$. Il faut donc compter les deux sortes d'événement :

- la traduction de u' par u : on somme, pour tous les couples (P,P') du corpus d'apprentissage, les probabilités liées aux alignements connectant u avec u':

$$c(f / f') = \sum_{\substack{(P,P') \\ alignés}} \sum_{j=1}^{m} \sum_{i=0}^{l} \frac{t(u/u') a(i / j, m, l) \delta(u, u_j) \delta(u', u'_i)}{\sum_{k=1}^{l} t(u / u'_k) a(k / j, m, l)} \tag{74}$$

  où $\delta(x,y)$ est la fonction de Kronecker, prenant une valeur nulle lorsque ses deux arguments sont différents.

- on somme, pour tous les couples (P,P') du corpus d'apprentissage ayant des longueurs respectives l et m, les probabilités d'aligner la position j avec la position i :

$$c(i / j, m, l) = \sum_{\substack{(P,P') \\ L(P)=l \\ L(P')=m}} \frac{t(f_j / f'_i) a(i / j, m, l)}{\sum_{k=1}^{l} t(f_j / f'_k) a(k / j, m, l)} \tag{75}$$

Sur la base de ces décomptes, on peut réestimer les fonctions t et a (cf. équation (70)), par une simple normalisation :

$$t(u/u') = \frac{c(u/u')}{\sum_{w \in T} c(w/u')} \qquad a(i/j,m,l) = \frac{c(i/j,m,l)}{\sum_{k} c(k/j,m,l)} \tag{76}$$

Le processus itératif est identique à celui du modèle 1. On peut notamment, dans la mesure où le premier modèle constitue un cas particulier du modèle 2, réutiliser les paramètres finaux délivrés par le modèle 1 :

1. *Initialisation* :
On initialise tous les paramètres *t*(*u*/*u'*) aux valeurs issues de l'étape précédente. De même, les paramètres *a*(*i*/*j*,*m*,*l*) initiaux peuvent être estimés à partir des *c*(*i*/*j*,*m*,*l*) calculés avec le modèle 1 (en utilisant l'équation (75) sans les probabilités d'alignement).

2. *Phase E* : *« Expectation »*.
On calcule les comptes *c*(*u*/*u'*) pour tous les couples (*u*,*u'*), ainsi que les comptes *c*(*i*/*j*,*m*,*l*) pour tous les quadruplets (*i*,*j*,*m*,*l*) (équation (75)).

3. *Phase M* : *« Maximisation »*.
On réestime les probabilités *t*(*u*/*u'*) à partir des *c*(*u*,*u'*), ainsi que les probabilités *a*(*i*/*j*,*m*,*l*) à partir des comptes *c*(*i*/*j*,*m*,*l*) (équations (76)).

4. *Retour à la phase E* jusqu'à stabilité des paramètres *t*(*u*/*u'*) et *a*(*i*/*j*,*m*,*l*).

III.2.2.3.4 Modèles 3, 4 et 5

Etant donné la complexité formelle des équations qui en dérivent, nous ne décrirons que les hypothèses liées à ces modèles, et le type de tirage aléatoire censé rendre compte du processus de génération.

Dans ces trois modèles, on introduit un nouveau paramètre : la « fertilité » liée à un mot, c'est-à-dire le nombre de mots qui lui sont connectés dans la traduction. De cette manière, on peut connecter plusieurs mots de la source avec plusieurs mots de la cible. Les auteurs donnent l'exemple suivant : lorsque *ne…que* traduit *only*, la fertilité de *only* est de 2. Le processus génératif suit les étapes suivantes :

- pour chaque mot *u'* de *P'*, on tire sa fertilité, i.e. le nombre de mots générés par *u'*.
- pour chaque mot *u'* on tire une liste de mots correspondant à sa fertilité.

– on tire une permutation de tous les mots ainsi tirés, pour obtenir la phrase finale *P*.

III.2.2.3.5 Résultats comparés des modèles 1-5

Les auteurs ont testé leurs méthodes sur un bi-texte tiré du corpus Hansard, comportant 1 778 620 couples de phrases d'une longueur inférieure ou égale à 30 mots. Afin d'éliminer les hapax et les mots erronés dus à des erreurs typographiques, ils ne retiennent que des mots apparaissant au moins deux fois dans le corpus. Au final, le corpus ainsi filtré comporte un vocabulaire de 42 005 unités en anglais et 58 016 unités en français.

A l'initialisation, les probabilités sont toutes instanciées avec une valeur identique, comme si toutes les traductions étaient équiprobables: $t(u/u') \leftarrow 1/58\,016$.

Si l'on ne tient compte que des couples qui cooccurrent parallèlement au moins une fois, on a donc 25 427 016 paires d'équivalents potentiels. En moyenne chaque mot anglais cooccurre avec 605 mots français. Après chaque itération, on ne retient les probabilités estimées $t(u/u')$ que lorsqu'elles sont supérieures à un certain seuil. Les probabilités rejetées sont fixées à une valeur non nulle de 10^{-12} assez petite pour que la normalisation n'en soit pas affectée, mais assez grande pour laisser à la connexion (u/u') la possibilité de resurgir dans une itération ultérieure.

Pour évaluer le degré de convergence des paramètres, les auteurs introduisent une mesure appelée « *perplexity* », que nous traduisons par *incertitude* : puisque le texte français contient 28 850 104 mots, la racine 28 850 104ème de la probabilité globale du bi-texte représente une mesure de la probabilité moyenne liée au tirage d'un mot français sachant le texte en anglais. L'inverse de cette probabilité est la « *perplexity* », l'*incertitude* moyenne liée à la traduction de chaque mot français. Cette mesure d'incertitude est donc inversement proportionnelle à l'information que le texte anglais apporte sur le texte français. Cette information (au sens de Shannon), si elle est significative, ne peut avoir qu'une seule origine : l'équivalence traductionnelle. Ainsi, mieux les paramètres modélisent les équivalences, plus l'information apportée par un texte sur l'autre sera forte, et plus l'incertitude sera faible. En bonne logique, la convergence des paramètres doit donc se manifester par une réduction de l'incertitude, les probabilités $t(u/u')$ se concentrant vers les couples (u,u') qui sont de véritables équivalents potentiels.

Les auteurs résument les résultats, après chaque itération, dans le tableau reproduit ci-dessous (1993 : 283) : les différents modèles ont été testés successivement, le modèle d'arrivée reprenant les paramètres du modèle de départ précédent.

Itération	*Modèle de départ*		*Modèle d'arrivée*	*Nombre de t(u/u') restants*	*Incertitude*
1	1	→	2	12 017 609	71 550,6
2	2	→	2	12 160 475	202,99
3	2	→	2	9 403 220	89,41
4	2	→	2	6 837 172	61,59
5	2	→	2	5 303 312	49,77
6	2	→	2	4 397 172	46,36
7	2	→	3	3 841 470	45,15
8	3	→	5	2 057 033	124,28
9	5	→	5	1 850 665	39,17
10	5	→	5	1 763 665	32,91
11	5	→	5	1 703 393	31,29
12	5	→	5	1 658 364	30,65

tableau 49 : *résultats pour Brown et al*. (*1993*)

A la dernière itération, il ne reste que donc 1 658 364 probabilités significatives (supérieure à un seuil non donné par les auteurs), correspondant à une moyenne de 49 mots français pour chaque mot anglais (au lieu de 605 cooccurrences au départ). En outre l'incertitude diminue considérablement dès la deuxième étape, et se stabilise progressivement aux alentours de 30.

Brown *et al*. notent que les résultats pourraient être améliorés par le recours à un traitement morphologique, qui réduirait la dispersion des traductions possibles en français (1993 : 285) : « Il est clair que nos modèles bénéficieraient d'un traitement morphologique qui limiterait l'exubérance lexicale du français. »[195] Par exemple *should* peut se traduire par : *devrait*, *devraient*, *devrions*, *doit*, *doivent*, *devons*, *devrais*. Les auteurs estiment qu'en lemmatisant ils pourraient réduire d'environ 50 % le vocabulaire du texte français et de 20 % le vocabulaire du texte anglais. Un tel traitement présenterait l'avantage de minimiser l'incertitude des modèles et de diminuer la complexité temporelle et spatiale des calculs.

En commentant ces résultats, Brown *et al*. remarquent que les probabilités obtenues après convergence n'ont pas de pertinence en dehors du corpus : il ne faut pas les interpréter comme des propriétés générales liées aux caractéristiques des langues, car ce sont des paramètres spécifiques dépendant du corpus d'apprentissage, notamment de sa taille (1993 : 286).

L'intérêt de ces modèles est grand : dans la mesure où ils ne s'appuient pas sur des correspondances considérées isolément, mais de façon globale, ils convergent vers la solution la plus vraisemblable au sein d'un espace comptant des millions d'hypothèses concurrentes (les *t*(*u*/*u'*)). Comme le notent les auteurs (1993 : 295), les bi-textes contiennent différents types de corrélations : certaines sont criantes (« *they shout from the data* ») et des méthodes très simples permettent de les appréhender – d'autres parlent d'une voix beaucoup plus faible (« *the quiet call of* (*marquées d'un astérisque*/*starred*) *or the whisper of* (*qui s'est fait bousculer*/*embattled*) », 1993 : 296). Pour capter celles-ci, on ne peut se passer d'une modélisation assez fine des corrélations statistiques.

On pourra certes objecter que ces modèles restent éloignés de la réalité, sur un plan linguistique. Il faut les concevoir comme des simplifications de nature opératoire (et non descriptive), destinées à produire des outils, et non des modélisations de la pratique traductionnelle. En tant que tels, ils ne s'opposent pas aux démarches basées sur une observation minutieuse des phénomènes linguistiques. Nous pensons comme les auteurs (1993 : 296) :

> « nous n'avons pas l'intention d'ignorer, ni de supplanter, la linguistique. Nous voudrions plutôt l'embrasser dans un cadre probabiliste sûr afin que les deux approches puissent chacune tirer des forces de l'autre, et nous guider vers de meilleurs systèmes de TALN en général, et de TA en particulier. »[196]

[195] "It is clear that our models would benefit from some kind of morphological processing to rein in the lexical exuberance of French."

[196] « But it is not our intention to ignore linguistics, neither to replace it. Rather, we hope to enfold it in the embrace of a secure probabilistic framework so that the two together may draw strength from one another and guide us to better natural language processing systems in general and to better machine translation systems in particular. »

III.2.2.3.6 Simplifications possibles

Kupiec (1993) propose une simplification du modèle 1 développé ci-dessus. Son étude est dédiée à l'extraction de correspondances entre composés nominaux. Ceux-ci sont identifiés automatiquement dans une première phase, à partir d'un étiquetage morphologique et de patrons syntaxiques simples. Après quoi, l'auteur met en œuvre une variante de l'algorithme EM, afin d'estimer les probabilités de traduction d'un composé nominal par un autre. A la différence des modèles de Brown *et al.*, les connexions ne sont pas concurrentes, à l'intérieur d'une même phrase : elles sont comptées indépendamment les unes des autres. Au lieu de l'équation (69) on a :

$$c(u/u') = \sum_{\substack{(P,P') \\ alignés}} t(u/u') \cdot Occ(u,P) \cdot Occ(u',P') \qquad (77)$$

La réestimation, elle, n'a pas changé (cf. équation (70)).

Dans ce modèle, très simplifié, on ne considère que des connexions entre composés nominaux prédécoupés : cette fois les découpages ne correspondent pas nécessairement au grain de la compositionnalité traductionnelle (à l'instar de nos expérimentations).

III.2.2.3.7 Modèles A, B et C

Melamed (1998a) propose une variante de l'algorithme EM, où les correspondances sont censées être biunivoques (« *word-to-word* »). Dans ce modèle, on néglige l'ordre des mots : le résultat de la traduction est conçu comme une suite de couples de mots, dont l'ordre n'importe pas, générés à partir d'une suite de concepts. Le processus génératif est donc symétrique vis-à-vis des deux langues :

$$Co_1, Co_2, \ldots, Co_n \rightarrow (u_1,u_1') (u_2,u_2') \ldots (u_n,u_n')$$

Pour appliquer ce modèle à son corpus d'apprentissage, Melamed définit la notion d'« *assignement* » : pour deux phrases *P* et *P'*, un *assignement* correspond à la mise en relation biunivoque des unités de *P* avec les unités de *P'*. S'il n'y a pas le même nombre d'unités dans *P* et *P'*, on introduit une unité vide « *Null* » (correspondant aux cas d'insertion ou d'omission) autant de fois que nécessaire, afin de rééquilibrer la taille des phrases. Un assignement correspond donc à une *permutation* des unités de *P* (ou de *P'*).

Pour des phrases de *n* unités, il existe donc *n! assignements* possibles. Un même couple de phrases peut donc être généré par *n!* tirages, impliquant des séries de couples (*u*,*u*') différents.

La probabilité de génération de deux phrases sera donc le résultat de la somme de toutes ces possibilités :

$$p(T,T') = \sum_{A} p(T,T',A) \tag{78}$$

Les paramètres du modèle sont les suivants (nous adaptons à nos propres notations) :

- *t*(*u*,*u*') la probabilité de générer la paire (*u*,*u*') à partir d'un même concept.
- *c*(*u*,*u*') le nombre de fois que *u* et *u*' sont susceptibles d'être générées ensemble, dans le corpus.
- *i*(*u*,*u*') un indice, exprimant le degré d'association entre *u* et *u*'. Cet indice (appelé « *likelihood* » par Melamed) est lié à *t*(*u*,*u*') et peut recevoir différentes définitions.

Les étapes de l'algorithme sont :

1. *Initialisation* des paramètres du modèle, en première approximation.

2. *Phase E* :
Estimation des *c*(*u*,*u*') à partir des cooccurrences observées dans le corpus, et des paramètres *t*(*u*,*u*').

3. *Phase M* :
Calcul des *i*(*u*,*u*') à partir des *c*(*u*,*u*').

4. *Retour en 2* jusqu'à convergence des paramètres.

5. *calcul des t*(*u*,*u*'), sur la base des comptes, en normalisant :

$$t(u,u') = \frac{c(u,u')}{\sum_{(w,w')\in TxT'} c(w,w')} \tag{79}$$

où *w* et *w*' représentent des unités quelconques de T et T'.

Le but de cet algorithme est de parvenir à l'ensemble de paramètres Ô maximisant la probabilité du corpus d'apprentissage :

$$\hat{O} = \arg\max_{O} p(T, T' / O) \tag{80}$$

Cependant, dans la phase d'estimation, il faudrait calculer les $c(u,u')$ en tenant compte de tous les *assignements* possibles, pour chaque couple de phrases (P,P'). Melamed note qu'un tel calcul n'est pas envisageable, puisque croissant de manière exponentielle avec la longueur des phrases. Il recourt donc à une approximation, déjà mise en œuvre par Brown *et al*. (1993 : 293) : l'estimation peut être faite sur la base de l'assignement le plus probable A_{max}, ou *assignement* de Viterbi.

Si $Z(n)$ représente la probabilité de tirer n couples, on a :

$$\begin{aligned} A_{\max} &= \arg\max_{A}(p(P, A, P' / O)) \\ &= \arg\max_{A}\left(Z(n).n! \prod_{(i,j)\in A} \log t(u_i, u'_j) \right) \\ &= \arg\max_{A}\left(\log(Z(n).n!) + \sum_{(i,j)\in A} \log t(u_i, u'_j) \right) \\ &= \arg\max_{A} \sum_{(i,j)\in A} \log t(u_i, u'_j) \end{aligned} \tag{81}$$

A partir de A_{max} on peut aisément calculer les $c(u,u')$: chaque couple de l'assignement incrémente le compte de 1 :

$$c(u, u') = \sum_{(u,u')\in A \max(P,P')} 1 \tag{82}$$

Mais les algorithmes permettant de déterminer avec certitude l'assignement de Viterbi sont encore trop coûteux, car en temps polynomial (1998a :12). Melamed propose d'effectuer une deuxième approximation basée sur une nouvelle hypothèse : vraisemblablement, le meilleur assignement est celui qui contient les plus grandes valeurs de $t(u_i,u'_j)$, prises individuellement. Cette quantité nous permet de donner une première définition de l'indice i : $i(u,u') = log\ t(u,u')$. A_{max} est donc l'assignement maximisant la somme de l'indice i.

Pour obtenir un assignement proche de A_{max} Melamed propose de mettre en œuvre ce qu'il appelle le « *competitive linking algorithm* », qui n'est autre que l'algorithme déjà cité permettant de trouver la meilleure affectation biunivoque (cf. p. 416).

Melamed montre que cet algorithme permet généralement d'obtenir un des meilleurs *assignements*. Nous noterons A_m l'assignement obtenu.

Par cette approximation, à savoir la prise en compte d'un seul assignement proche de A_{max}, on sort du cadre strict des conditions d'application de l'algorithme EM : la convergence vers les paramètres optimaux n'est plus absolument garantie. Les travaux empiriques de Melamed montrent cependant une convergence, plus irrégulière, mais aussi plus rapide, aboutissant à de meilleurs résultats.

A partir du cadre précédemment exposé, Melamed propose trois modèles, qui mettent en œuvre différentes versions de la fonction *i*(*u*,*u*') (que l'auteur note *like* pour *likelihood*) :

– *Modèle A*

Ce modèle est le plus simple : à l'initialisation, la fonction *i*(*u*,*u*') est identifiée à l'indice tiré du rapport de vraisemblance, RV, calculé à partir des cooccurrences (cf. les paramètres *a*, *b*, *c* et *d*). Puis l'on emploie la formule de *i*(*u*,*u*') déjà donnée. En résumé, les étapes sont :

1. Initialisation :
Pour tous les (*u*,*u*'), on calcule : i(*u*,*u*') ← RV(*u*,*u*')

2. Phase E :
Pour chaque couple (*P*,*P*') on applique le « *competitive linking algorithm* », avec les valeurs courantes de *i*(*u*,*u*'). A partir des $A_m(P,P')$ ainsi obtenus, on calcule, pour tous les (*u*,*u*') :

$$c(u,u') = \sum_{(u,u') \in Am(P,P')} 1 \tag{83}$$

3. Phase M :
Maximisation des probabilités :

$$t(u,u') = \frac{c(u,u')}{\sum_{(w,w') \in TxT'} c(w,w')} \tag{84}$$

L'indice est recalculé sur cette base :

$$i(u,u') = \log t(u,u') \tag{85}$$

4. Retour en 2 jusqu'à convergence (lorsque les variations de *t* sont inférieures à un certain seuil).

Les correspondances sont ensuite simplement déduites des derniers assignements A_m obtenus.

– *Modèle B*

Pour ce second modèle, Melamed s'inspire d'une remarque de Yarowsky (1993) : « un mot ambigu a un seul sens dans une collocation donnée avec une probabilité de 90-99 % »[197]. Par « collocation », il faut ici entendre « contexte proche ». Nous avons déjà évoqué ce phénomène : pour déterminer l'*emploi* d'une lexie polysémique, le co-texte immédiat est une source d'information précieuse, et joue le rôle de sélection d'acception. Melamed propose de tirer parti de ce phénomène dans une configuration bilingue, en notant que la cooccurrence dans le bi-texte est un genre de co-texte. Il fait le raisonnement suivant : quand *u* et *u*' sont des équivalents potentiels, c'est qu'une certaine acception /u_a/ recouvre (au moins partiellement), une certaine acception /u_a'/, et ces emplois apparaissent dans les même contextes. Ainsi, lorsque *u* apparaît dans le contexte de *u*', il est probable que *u* soit employé dans l'acception /u_a/, et réciproquement que *u*' soit employé dans l'acception /u'_a/ : par conséquent il est probable qu'ils soient en relation d'équivalence traductionnelle.

Pour un couple (*u*,*u*') d'équivalents potentiels, le rapport : $r = \dfrac{c(u,u')}{Cooc(u,u')}$ doit donc être voisin de 1, sauf erreur dans le calcul de *c*. Et quand (*u*,*u*') ne sont pas équivalents potentiels, ce même rapport doit être égal à 0, puisque les *c*(*u*,*u*') doivent être égaux à 0, sauf erreur dans le calcul de *c*.

Après une itération de la méthode A, sur 300 000 phrases alignées du Hansard, Melamed constate que la distribution de *r* est effectivement caractérisée par deux pics, la

[197] "for several definitions of sense and collocation, an ambiguous word has only one sense in a given collocation with a probability of 90-99 %".

plupart des couples se concentrant vers 0 ou 1, le phénomène étant plus marqué pour les couples qui cooccurrent souvent.

Melamed propose d'intégrer ce phénomène dans son modèle afin de mieux contrôler les effets des *c*(*u*,*u*') erronés.

On note $\lambda+$ la probabilité que des équivalents potentiels (*u*,*u*') soient connectés quand ils cooccurrent (i.e. quand leur cooccurrence aboutit à l'incrémentation de *c*(*u*,*u*')).

On note $\lambda-$ la probabilité qu'un couple d'unités non-équivalentes (*u*,*u*') soient connectés quand ils cooccurrent.

On supposera que $\lambda+$ (resp. $\lambda-$) est identique quels que soient les couples (*u*,*u*') équivalents (resp. non-équivalents) : ces deux paramètres correspondent aux deux pics de la distribution précédemment évoquée. $\lambda+$ doit donc être voisin de 1, et $\lambda-$ de 0.

A partir de ces paramètres, on peut donner une estimation de la probabilité d'obtenir *c*(*u*,*u*') connaissant *Cooc*(*u*,*u*'), sous la forme d'une loi binomiale, en fonction de deux hypothèses antagonistes :

- hypothèse 1 : si (*u*,*u*') sont des équivalents potentiels :

$$p(c(u,u')/Cooc(u,u'),\lambda^+) = C_{Cooc(u,u')}^{c(u,u')}\lambda^{+c(u,u')}(1-\lambda^+)^{Cooc(u,u')-c(u,u')} \quad (86)$$

- hypothèse 2 : si (*u*,*u*') ne sont pas des équivalents potentiels :

$$p(c(u,u')/Cooc(u,u'),\lambda^-) = C_{Cooc(u,u')}^{c(u,u')}\lambda^{-c(u,u')}(1-\lambda^-)^{Cooc(u,u')-c(u,u')} \quad (87)$$

Dans le premier cas on se situe dans l'hypothèse d'un processus générant des connexions justes, et dans le second on se situe dans l'hypothèse d'un processus générant des connexions erronées.

Pour un couple de paramètres (*c*(*u*,*u*'), *Cooc*(*u*,*u*')), on exprime ainsi le rapport de vraisemblance lié à l'hypothèse 1, et l'on en déduit un nouveau calcul de l'indice *i* :

$$i(u,u') = \log\frac{p(c(u,u')/Cooc(u,u'),\lambda^+)}{p(c(u,u')/Cooc(u,u'),\lambda^-)} \quad (88)$$

L'algorithme lié au modèle B est identique à celui de A. La seule différence tient dans le recalcul de *i*(*u*,*u*') à chaque étape.

Melamed note une propriété intéressante du modèle B : il peut y avoir des unités *u* pour lesquelles $i(u,u') < 0$ quelle que soit *u'* : ce sont des unités pour lesquelles le modèle est dans l'incertitude. Ces unités, dont le comportement échappe au modèle, peuvent ouvrir la voie à des travaux intéressants. Nous verrons plus loin comment tirer parti de ces propriétés.

Pour l'estimation des paramètres $\lambda+$ et $\lambda-$ (qui ne varient pas, comme les autres paramètres, à chaque itération) il faut trouver un couple de valeurs permettant de maximiser la probabilité des distributions de $c(u,u')$ sur le corpus d'apprentissage :

$$(\lambda^+{}_m, \lambda^-{}_m) = \arg\max_{(\lambda^+,\lambda^-)} \prod_{(u,u')} p(c(u,u') / Cooc(u,u'), \lambda^+, \lambda^-) \tag{89}$$

On peut calculer cette probabilité en introduisant un nouveau paramètre : la probabilité τ que deux unités (u,u') tirées au hasard soient équivalentes. La probabilité de $c(u,u')$ devient :

$$\begin{aligned} p(c(u,u') / Cooc(u,u'), \lambda^+, \lambda^-) = \tau . p(c(u,u') / Cooc(u,u'), \lambda^+) \\ + (1-\tau) . p(c(u,u') / Cooc(u,u'), \lambda^-) \end{aligned} \tag{90}$$

Melamed montre qu'on peut calculer τ à partir de λ^+ et λ^-. En effet, si λ représente la probabilité qu'un couple quelconque (u,u') soit connecté, on a :

$$\lambda = \tau . \lambda^+ + (1-\tau) . \lambda^- \tag{91}$$

Or λ peut être calculé empiriquement :

$$\lambda = \frac{\sum_{(u,u')} c(u,u')}{\sum_{(u,u')} Cooc(u,u')} \tag{92}$$

On en déduit une formulation de τ en fonction de λ^+ et λ^- :

$$\tau = \frac{\frac{\sum_{(u,u')} c(u,u')}{\sum_{(u,u')} cooc(u,u')} - \lambda^-}{\lambda^+ - \lambda^-} \tag{93}$$

Ainsi, du fait des équations (90)et (93) la distribution globale des $c(u,u')$ n'a que deux degrés de liberté, les variables λ^+ et λ^-.

Melamed calcule la valeur de la probabilité de ces distributions en fonction de différentes valeurs de λ^+ et λ^- (après une itération de l'algorithme sur 300 000 couples de phrases) :

$$p_c(\lambda^+, \lambda^-) = \prod_{(f,f')} p(c(u,u') / Cooc(u,u'), \lambda^+, \lambda^-)$$

on remarque que la fonction p_c est convexe et qu'elle possède un maximum unique : une méthode classique de recherche d'extremum permet d'estimer les paramètres λ^+_m et λ^-_m qui maximisent la probabilité du corpus d'apprentissage.

– *Modèle C*

Il est possible de raffiner l'algorithme du modèle B en apportant une très légère modification au calcul de λ^+ et λ^-. En effet les valeurs de λ^+ et λ^- dépendent du type des unités impliquées, car toutes les unités n'ont pas la même constance dans la traduction. Ainsi, les mots outils, plus fréquents, et dont le sens est quasiment vide et dépendant du contexte, peuvent avoir des traductions très variables. A l'inverse, des substantifs rares sont susceptibles d'être traduits avec plus de régularité. Si l'on compte les *c*(*u*,*u*') séparément, en fonction d'une typologie des unités, on peut donc calculer un jeu de paramètres (λ^+,λ^-) plus représentatif de la réalité : pour des connexions qui mettent en jeu des unités plus stables, les deux valeurs seront plus éloignées ; à l'inverse, pour des connexions qui impliquent des unités à la traduction variable, λ^+ et λ^- seront moins différenciées.

Dans son évaluation, Melamed établit une typologie rudimentaire, basée essentiellement sur la distinction mot plein / mot vide et les marques de ponctuation. Il considère 7 classes différentes : les ponctuations de fin de syntagme, les ponctuations de fin de phrase, les marqueurs de subordonnée (comme les guillemets et les parenthèses), les symboles (du type ~ ou *), le mot vide (noté *NULL*), les mots pleins (noms, adjectifs, adverbes, verbes non auxiliaires), les mots vides. L'estimation des (λ^+_m,λ^-_m) est effectuée comme précédemment, mais séparément pour les différentes catégories de *u* rentrant dans le calcul de comptes *c*(*u*,*u*').

Le calcul différencié des (λ^+_m,λ^-_m) en fonction du type de connexion permet d'intégrer d'autres formes d'information :

- par la prise en compte simultanée de la partie du discours de *u* et de *u*', afin d'obtenir un jeu de paramètres encore plus précis ;

- par la prise en compte des positions des unités dans la phrase, en classant dans différentes catégories les connexions entre unités éloignées (dans leurs phrases respectives) et unités rapprochées.

III.2.2.3.8 Résultats comparés des modèles A-C

Afin de pouvoir comparer ces modèles avec les méthodes expérimentées plus loin, nous indiquons rapidement les résultats obtenus par Melamed lors de ses tests.

Dans son évaluation, ce dernier met en œuvre les trois modèles A, B et C, ainsi que le modèle 1 de Brown *et al*., servant de base de comparaison. Le corpus d'entraînement est constitué de 29 614 versets de la bible, traduits en anglais et en français.

Le processus itératif se termine lorsque les probabilités de traduction *t*(*u*,*u*') varient globalement de moins de 0,0001 d'une itération à l'autre (nous noterons Δt cette variation moyenne). Les 4 modèles convergent plus ou moins vite (cf. tableau 50) : le modèle 1 est le plus long, et la convergence est aussi plus régulière, puisqu'il y a diminution progressive de Δt. Le modèle A converge plus vite. Les modèles B et C se situent entre les deux.

Modèle	***1***	***A***	***B***	***C***
Nombre d'itérations	40	9	18	26

tableau 50 : *convergence des 4 modèles*

Afin de mesurer la validité des résultats, 250 versets ont été extraits du corpus et ont été annotés manuellement. Chaque mot, des deux côtés du bi-texte, s'est trouvé connecté à zéro, un ou plusieurs mots correspondants dans le segment aligné (cf. la description du projet Blinker, p. 379). L'évaluation est effectuée de manière classique, par comparaison

entre les connexions produites et les connexions établies manuellement dans le corpus de référence. On en déduit les mesures de précision et de rappel.

Deux ensembles de connexions sont pris en compte :

- les connexions entre tous les mots :

  $$1 < A < B < C$$

  La précision et le rappel des résultats du modèle 1 se situent autour de 25 % tandis que le modèle C réalise des scores aux alentours de 40 % (les modèles A et B se situant entre les deux).

- les connexions entre les mots pleins seulement (« *open-class links only* »). On peut classer les résultats dans l'ordre suivant :

  $$A < 1 < B = C$$

  Globalement la précision est légèrement meilleure (aux alentours de 45 %) et le rappel légèrement moins bon (environ 25 %), lorsqu'on évalue les liens entre mots non-fonctionnels seulement.

Le modèle A se comporte assez mal, mais B et C réalisent un gain notable de précision et de rappel par rapport au modèle 1 (entre 10 et 15 points environ). Melamed explique la mauvaise performance de A par le fait que le critère de biunivocité n'est pas vérifié pour les mots non-fonctionnels. La prise en compte des erreurs, dans les modèles B et C permet de mieux contrôler ces distorsions. Le peu de différence entre B et C était prévisible, dans la mesure où le modèle C tire son avantage de l'ajustement de ses paramètres en fonction de la différence entre mots outils et mots pleins.

III.3 Expérimentation

Nous présentons maintenant un travail empirique destiné à fournir une évaluation globale des techniques présentées précédemment, tant sur le plan des mesures statistiques que de l'implémentation algorithmique.

III.3.1 Etude préliminaire des indices

Dans un premier temps, nous nous sommes intéressé aux différents indices d'association. Nous avons déjà présenté quatre indices calculés à partir des seules distributions des unités : l'information mutuelle, le t-score, le rapport de vraisemblance et le logarithme de la probabilité de l'hypothèse nulle. Nous noterons ces indices, respectivement : IM, TS, RV et P0.

A ces indices nous en adjoignons 2 autres, qui intègrent un autre type d'information : la *cognation*. En effet, comme nous l'avons vu dans le chapitre 2, la ressemblance superficielle peut-être un indice d'équivalence traductionnelle.

Dans la comparaison de deux unités, nous avons ainsi isolé les 11 cas de figure (correspondant aux cas du tableau 21, p. 316) :

- cas 1 : transfuge numérique ;
- cas 2 : transfuge de longueur supérieure à 3 ;
- cas 3 : 4-gramme de longueur inférieure à 7 ;
- cas 4-10 : sous-chaînes maximales comportant 4-10 caractères ;
- cas 0 : complémentaire des cas 1-10.

Pour chaque binôme (*P*,*P'*) de notre échantillon, nous avons relevé les cas de figure correspondant à divers degrés de ressemblance entre les couples de lexies manuellement appariées (on notera : *corr*), ainsi qu'entre des lexies non appariées (on notera : *non-corr*).[198] En reprenant les paramètres optimaux utilisés dans la phase d'alignement (notamment avec $r_{seuil} = 2/3$), on obtient les statistiques suivantes :

[198] Plus précisément, nous avons opéré au niveau des formes simples : à l'intérieur de chaque couple (*P*,*P'*), nous avons comparé tous les couples de forme (*f*,*f'*). Nous avons considéré les formes comme étant correspondantes (événement *corr*) lorsqu'elles étaient incluses dans un couple de lexies appariées, et non-correspondantes sinon.

Cas	*corr*	*non-corr*	*p*(*corr*/*cas*)	*p*(*non-corr*/*cas*)
0. nul	18 843	890 112	2,07 %	97,93 %
1. transfuge numérique	3	0	100,00 %	0,10 %
2. transfuge de long>3	1 014	29	97,22 %	2,78 %
3. 4-gram^{+} *pour 3<long<7*	121	47	72,02 %	27,98 %
4. SCM = 4	145	143	50,35 %	49,65 %
5. SCM = 5	159	65	70,98 %	29,02 %
6. SCM = 6	314	87	78,30 %	21,70 %
7. SCM = 7	308	44	87,50 %	12,50 %
8. SCM = 8	268	47	85,08 %	14,92 %
9. SCM = 9	152	17	89,94 %	10,06 %
10. SCM 10…n	97	6	94,17 %	5,83 %

tableau 51 : *probabilités des différents degrés de ressemblance formelle entre les couples d'unités correspondantes et non-correspondantes* ($r_{seuil}=2/3$)

Si l'on ne garde que les n-grammes et SCM respectant la contrainte $r_{seuil} = 0,5$, on obtient un paramétrage plus tolérant, générant plus de bruit, mais limitant le silence dans l'identification des cognats. On obtient alors un autre jeu de probabilités :

Cas	*corr*	*non-corr*	*p*(*corr*/*cas*)	*p*(*non-corr*/*cas*)
0. nul	18 279	888 966	2,01 %	97,99 %
1. transfuge numérique	3	0	100,00 %	0,10 %
2. transfuge de long>3	1 014	29	97,22 %	2,78 %
3. 4-gram^{+} *pour 3<long<7*	151	331	31,33 %	68,67 %
4. SCM = 4	335	307	52,18 %	47,82 %
5. SCM = 5	287	352	44,91 %	55,09 %
6. SCM = 6	432	341	55,89 %	44,11 %
7. SCM = 7	384	174	68,82 %	31,18 %
8. SCM = 8	284	68	80,68 %	19,32 %
9. SCM = 9	158	23	87,29 %	12,71 %
10. SCM 10…n	97	6	94,17 %	5,83 %

tableau 52 : *probabilités des différents degrés de ressemblance formelle entre les couples correspondants et non-correspondants* ($r_{seuil}=0,5$)

Dans un cas comme dans l'autre, on voit que la probabilité de l'événement *corr* est très largement corrélée au cas de figure. On peut s'inspirer de cette classification pour construire un nouvel indice. Nous avons choisi de nous baser sur l'invraisemblance de l'événement « non-corr/cas » :

$$CO = -log(p(non\text{-}corr/cas)) \quad (94)$$

A partir de CO, on peut également forger un nouvel indice, combinant cognation et distribution :

$$PC = P0 + CO \quad (95)$$

Ce dernier indice combine les probabilités correspondant à deux tirages successifs (tirage des distributions et tirage aléatoire de la forme des signifiants) dans l'hypothèse où la cooccurrence serait totalement fortuite. Plus la valeur de cet indice est grande, et moins cette hypothèse est tenable. Nous étudierons CO avec les deux jeux de probabilités donnés ci-dessus, afin de déterminer les impacts relatifs du bruit et du silence dans l'identification des cognats.

Enfin, nous testerons un dernier indice, sans signification, qui servira d'étalon : cet indice, noté AL, est basé sur le tirage aléatoire d'un réel entre 0 et 1. Il n'est construit à partir d'aucune information spécifique. Son seul rôle est de donner une estimation des résultats au cas où l'on se livrerait à une extraction de correspondance au hasard.

En récapitulant, nous serons amenés à étudier les sept indices suivants : IM, TS, RV, P0, CO, PC, AL.

III.3.1.1 Rapport des moyennes des indices : RMI

Une première évaluation consiste à comparer les valeurs des indices pour des couples d'unités quelconques et pour des couples d'unités[199] appariées.

Ainsi, pour tous les couples de phrases (*P*,*P'*) du corpus de référence, nous avons calculé les valeurs des indices en croisant toutes les unités de *P* avec toutes les unités *P'*. Par ailleurs nous avons calculé les indices pour les seuls couples d'unités appariées.

[199] Ces unités sont les « unités de traduction » (simples ou polylexicales) identifiées manuellement lors de l'extraction du corpus de référence. Par commodité, nous référerons désormais à ces unités sous le terme de « lexies », en opposition avec les « lemmes » (i.e. les mêmes unités lemmatisées) et les « formes simples » (i.e. les mots constituant les lexies).

Pour chaque indice, on obtient deux distributions : les valeurs obtenues pour des couples quelconques, et les valeurs pour les couples d'unités équivalentes. Plus ces distributions sont éloignées, et meilleur est le pouvoir discriminant de l'indice. Le rapport des moyennes de chaque distribution est donc un indicateur intéressant.

RMI	***IM***	***TS***	***RV***	***P0***	***COa*** $r_{seuil= 0,66}$	***COb*** $r_{seuil= 0,5}$	***PC***
moyenne 1 (*corr et non-corr*)	1,18	1,65	7,99	8,58	0,01	0,01	8,59
moyenne 2 (*corr*)	2,51	4,82	98,16	98,72	0,12	0,13	98,84
moyenne 2 / *moyenne 1*	2,13	2,92	12,28	11,51	8,96	9,74	11,50

tableau 53 : *rapport des moyennes entre les distributions*
1=« unités quelconques » et 2=« unités appariées »

Notons :

$p(x)=p(indice \geq x)$ dans le cas des couples quelconques
$p_{corr}(x)=p(indice \geq x/corr)$ dans le cas des couples de référence

Le rapport des moyennes, synthétisant l'écart entre les distributions, peut être représenté par le rapport entre les surfaces définies par les courbes de la figure 40, qui représentent les fonctions $p(x)$ et $p_{corr}(x)$. Comme le suggèrent Davis *et al*. (1995), ces statistiques donnent une première estimation de l'efficacité respective des indices : les indices RV, P0 et PC paraissent être nettement plus discriminants. On remarque en outre que RV et P0 ont des distributions extrêmement voisines, ce qui paraît normal puisqu'ils dérivent d'un raisonnement similaire (même s'ils n'ont pas la même signification : RV est homogène à un rapport de probabilités tandis que P0 est homogène à une probabilité, au logarithme près).

On trouvera en annexe (figure 73 à figure 78) l'évolution de ces distributions en fonction des fréquences des lexies mises en jeu. En se basant sur *fréq. min.*, le nombre d'occurrences de la lexie la moins fréquente du couple comparé, nous avons dégagé cinq tranches :

$$\begin{aligned} & fréq.\ min. \leq 3 \\ 3 < & fréq.\ min. \leq 10 \\ 10 < & fréq.\ min. \leq 50 \\ 50 < & fréq.\ min. \leq 500 \end{aligned}$$

$$500 < fréq.\ min. \leq 5000$$

On constate les faits suivants :

– Pour tous les indices, sauf CO, le pouvoir discriminant des indices augmente avec la fréquence des lexies. Ceci est spécialement marqué pour IM et TS. L'efficacité globale des indices est donc corrélée à la taille des corpus : plus un bi-texte est grand, plus ses unités sont récurrentes et plus il est aisé de déterminer leurs correspondants. Plus précisément, ces indices sont mis en échec par l'apparition de plusieurs hapax dans un même binôme. Nous étudierons ultérieurement (cf. p. 485) la probabilité d'un tel événement.

– A l'inverse, l'indice CO (on a représenté la courbe de paramètre $r_{seuil} = 0{,}5$) est insensible à la fréquence des lexies, ce qui était prévisible puisqu'il n'est pas basé sur leurs distributions. Les paliers observés sur les courbes sont liés aux différents cas de figure, en nombre discret. Cet indice se comporte de manière similaire à un indice basé sur un dictionnaire bilingue, qui indiquerait deux cas de figure : 1 si les lexies sont données comme équivalents potentiels et 0 sinon. Mais par rapport à un tel indice, CO est à la fois plus bruyant et plus silencieux.

III.3.1.2 Moyenne des rapports des probabilités : MRP

Cependant, la donnée de ces rapports n'a de sens que si l'on tient compte de la structure des courbes, car le pouvoir discriminant d'un indice n'est qu'indirectement lié à la différence des surfaces ici représentées. En effet ces surfaces sont fonction de la valeur absolue des indices, tandis que l'efficacité d'un indice dépend plus étroitement des valeurs relatives prises par l'indice dans la comparaison des couples. Par exemple, on constate que les rapports du tableau 11 sont favorables à l'indice CO. Or cet indice n'apporte pas d'information pour environ 79 % des couples de référence (avec COb) : ainsi, même si la précision dégagée par CO peut être correcte, le rappel en sera nécessairement faible.

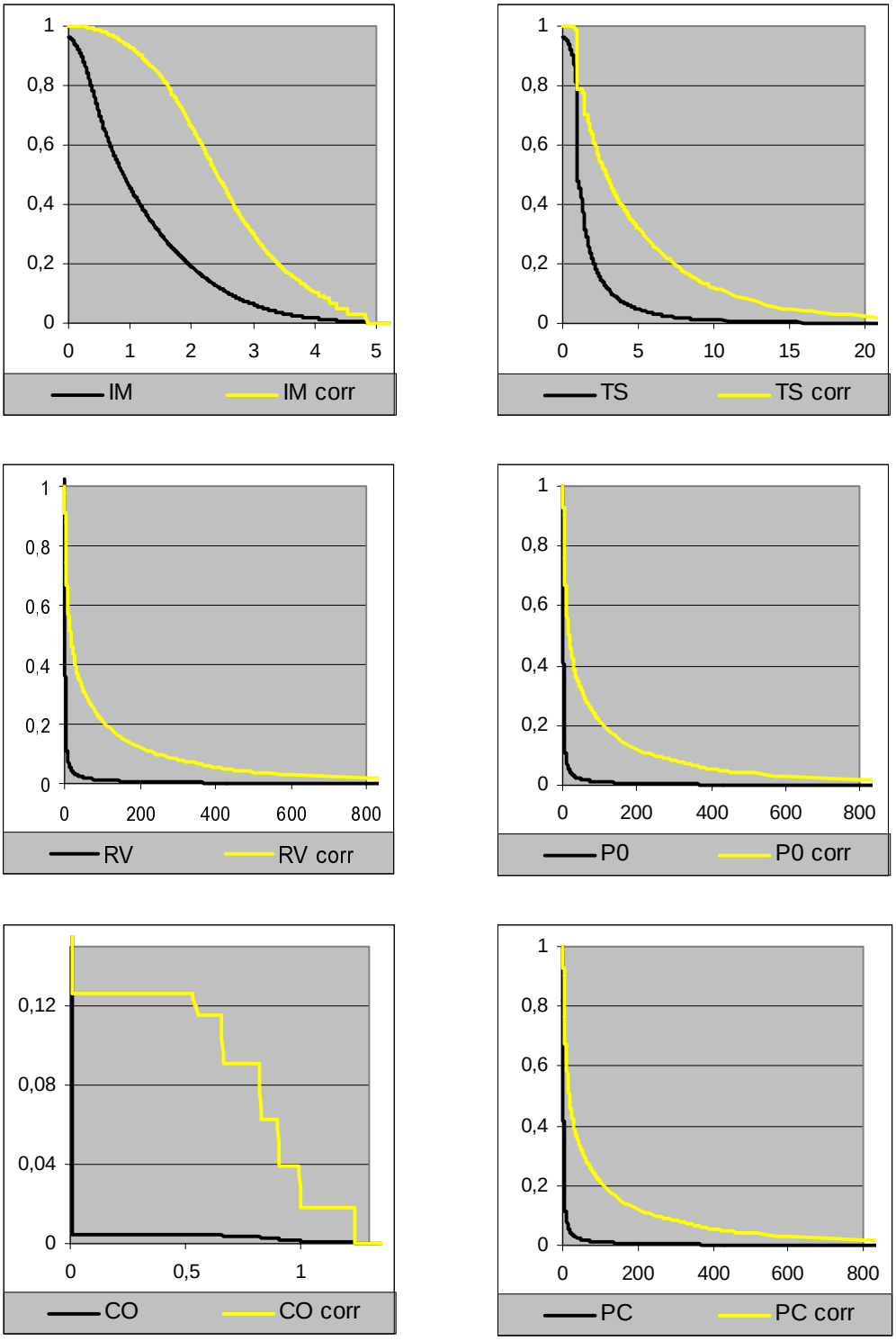


figure 40
Représentations des fonctions $y = p(x)$ et $y = p_{corr}(x)$
pour les indices IM, TS, RV, P0, CO et PC

Pour corriger ce biais, il faut rapporter l'axe des abscisses, non à la valeur absolue de l'indice, mais au nombre de couples quelconques obtenant cette valeur de l'indice. On obtient alors les courbes de la figure 41, avec en abscisse la probabilité $p(x)$, et en ordonnée les probabilités $p(x)$ (ce qui donne la diagonale) et $p_{corr}(x)$. Cette fois, la courbe liée à CO montre les insuffisances de cet indice.

Enfin, pour une valeur x de l'indice, ce n'est pas la différence entre $p(x)$ et $p_{corr}(x)$ qui exprime le caractère discriminant de l'indice, mais leur rapport : il tend vers une valeur maximum[200] lorsque l'indice apporte une certitude, et vers 1 lorsqu'il n'apporte aucune information. Nous avons donc représenté ce rapport pour chaque indice, figure 42 : pour chaque courbe, la moyenne du rapport est calculée par son intégrale de 0 à 1 (cf. équation (96) – notons que la variable n'est plus x, mais $p = p(x)$). Ces moyennes, que nous désignons par le sigle MRP, sont données dans le tableau 54.

$$MRP = \int_0^1 \frac{p_{corr}(indice \leq X(p))}{p} dp \qquad (96)$$

où $X(p)$ représente la fonction réciproque de $p(x)$

	IM	***TS***	***RV***	***P0***	***COa*** $r_{seuil= 0,66}$	***COb*** $r_{seuil= 0,5}$	***PC***
MRP	2,36	2,43	3,14	3,10	1,13	1,21	3,11

tableau 54 : *MRP pour chaque indice*

[200] Ce maximum correspond au cas où une valeur de l'indice permettrait une discrimination totale entre couples de lexies équivalentes et les couples de lexies non correspondantes. Comme l'événement quelconque correspond à l'union des événements complémentaires *corr* et *non-corr* on aurait (en notant *Nc* et *Nn* les effectifs respectifs de *corr* et *non-corr*) : $MRPmax = 1/(Nc/(Nc+Nn)) = 1+Nn/Nc$

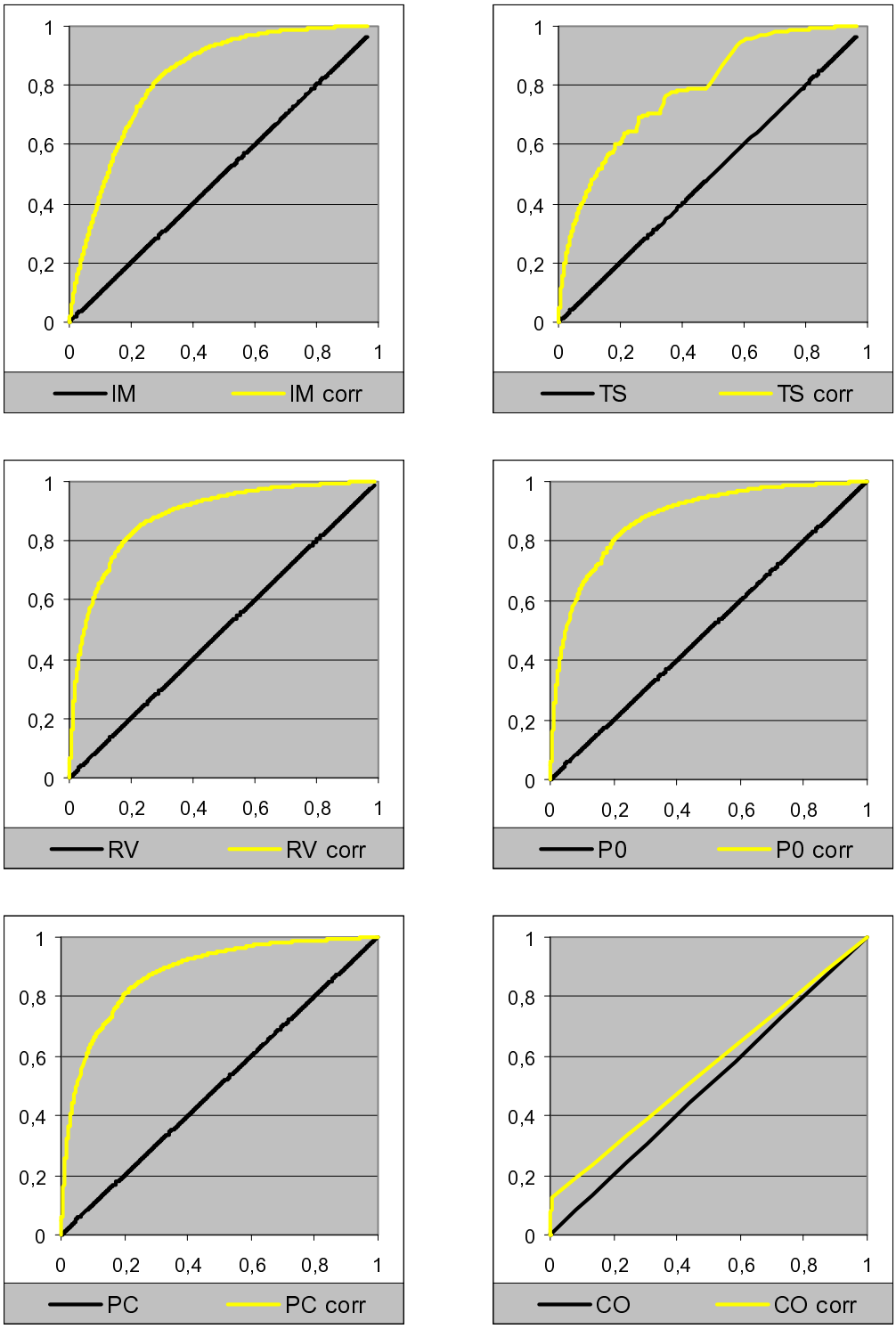


figure 41
Fonctions $y = p_{corr}(X(p))$ et $y = p$ pour les indices IM, TS, RV, P0, CO et PC (avec p en abscisse)

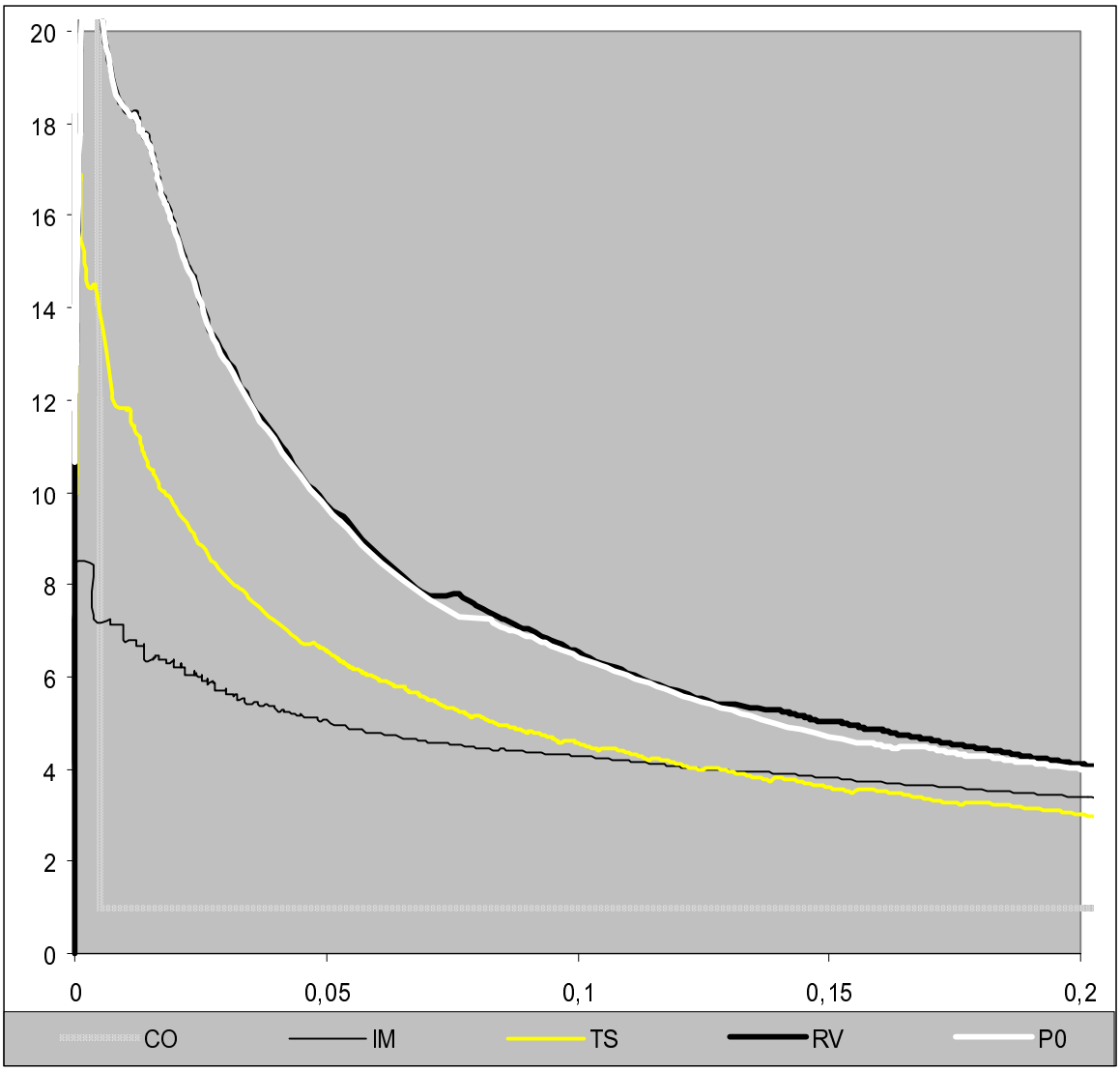


figure 42
Représentation de la fonction $p_{corr}(X(p))/p$ pour les indices IM, TS, RV, P0 et CO.
N.B. : toutes les courbes ayant le même comportement asymptotique, le rapport tendant vers 1 en même temps que l'abscisse, nous n'en présentons ici que la partie la plus intéressante, pour les abscisses comprises entre 0 et 0,2.

Là encore, on assiste à un quasi-chevauchement de RV, P0 et PC. Cette première évaluation nous amène à présumer le classement suivant :

$$COa < COb < IM < TS < P0 < PC < RV$$

Pour l'indice CO, nous avions émis l'hypothèse qu'il était plus intéressant de privilégier la diminution du silence (l'augmentation du rappel) : le meilleur score de COb nous en donne une première confirmation.

III.3.2 Comparaison des algorithmes et des indices

Nous allons maintenant évaluer les indices lorsqu'ils sont utilisés pour l'extraction des correspondances.

III.3.2.1 Définition des tâches d'extraction

Un premier type d'extraction consiste à considérer un ensemble limité d'unités lexicales, choisies en fonction de critères linguistiques. Ce fut la stratégie adoptée lors de la deuxième phase du projet ARCADE (1998), avec l'évaluation du « *translation spotting* ». Pour chaque occurrence des 60 unités sources préalablement choisies, il fallait repérer la (ou les) unités correspondante(s) dans le texte cible. Ce type d'extraction, considéré comme une étape préliminaire à l'extraction complète, a l'avantage de centrer l'évaluation sur des unités clairement identifiées. Dans le projet ARCADE, par exemple, les 60 unités lexicales ont été choisies suivant trois types de critères :

- morphosyntaxiques : les unités se répartissaient en 20 verbes, 20 substantifs et 20 adjectifs.
- de fréquences : toutes les unités avaient une fréquence similaire.
- sémantiques : afin que l'extraction ne soit pas triviale, les annotateurs ont sélectionné les 60 unités considérées comme étant les plus polysémiques, parmi 600 unités de fréquences comparables.

La neutralisation de certaines variables, comme ici la fréquence, permet de mieux cerner l'influence d'autres variables, par exemple la classe morphosyntaxique. En outre, l'exigence de polysémie permet de tester les techniques là où précisément elles rencontrent des difficultés, ce qui permet de limiter l'évaluation aux seuls cas intéressants.

Notons que certaines techniques d'extraction précédemment présentées nécessitent la mise en œuvre d'extractions complètes, où la correspondance d'une unité donnée est déterminée par l'équilibre global des correspondances des autres unités (c'est le cas de toutes les méthodes reposant sur l'algorithme EM). C'est pourquoi, dans la présente

évaluation, nous avons choisi d'évaluer d'abord des extractions complètes[201], quitte à restreindre le champ, par la suite, à certaines unités.

Nous proposons trois tâches pour cette évaluation :

- *L'extraction des correspondances entre lexies* (*LEX*) : on cherche les correspondances entre les unités que nous avons dégagées manuellement dans l'élaboration du corpus de référence.

- *L'extraction des correspondances entre lexies, avec lemmatisation* (*LEM*) : en supprimant les variations morphologiques superficielles des lexies (cf. supra, p. 396) on recalcule les cooccurrences et l'on effectue une extraction similaire à la précédente. Les unités mises en correspondance sont les mêmes que précédemment, mais le calcul des indices s'en trouve modifié, puisque les distributions considérées sont élargies à toutes les unités liées au même lemme. En outre, lors de cette extraction, la cognation est mesurée au niveau des lemmes, et non au niveau des lexies.

- *L'extraction des correspondances entre formes simples* (*FS*) : les unités considérées sont des mots simples (groupes de lettres compris entre des séparateurs, sans tiret ni apostrophe) n'ayant subi aucun pré-traitement. Nous chercherons ainsi à déterminer l'impact de l'identification préalable des unités polylexicales sur les résultats. En effet, on peut supposer que l'identification des unités non compositionnelles facilite la tâche, puisqu'elle permet de se rapprocher d'une certaine biunivocité dans l'extraction des correspondances traductionnelles. Nous émettons donc l'hypothèse que les résultats de cette troisième tâche devraient être moins bons, que ceux de la première.

L'exemple suivant permet d'illustrer ces trois tâches :

[201] A ceci près que nous avons négligé tous les signes de ponctuation, dont la mise en correspondance ne nous a pas paru utile d'un point de vue linguistique : point, virgule, point-virgule, deux points, point d'exclamation, point d'interrogation, parenthèses, tiret, guillemet.

fr. : *Pour la bonne tenue de ces registres, l'évaluation des cas de mortalité constatés par les autorités apporte des informations importantes*.
angl. : *The assessment of the official cause of death is a piece of information vital to these registers*.

Lexies : (*Pour* ; *to*) (*ces* ; *these*) (*registres* ; *registers*) (*l* ; *the*) (*évaluation* ; *assessment*) (*des* ; *of the*) (*cas de mortalité* ; *cause of death*) (*des* ; *a piece of*) (*informations* ; *information*) (*importantes* ; *vital*)

Lemmes : (*Pour* ;*to*) (*ce* ; *this*) (*registre* ; *register*) (*le* ;*the*) (*évaluation* ; *assessment*) (*de le* ; *of the*) (*cas de mortalité* ; *cause of death*) (*de le* ; *a piece of*) (*information* ; *information*) (*importante* ; *vital*)

Formes simples : (*Pour* ; *to*) (*ces* ; *these*) (*registres* ; *register*) (*l* ; *the*) (*évaluation* ; *assessment*) (*des* ; *of*) (*cas* ; *cause*) (*de* ; *of*) (*mortalité* ; *death*) (*des* ; *piece*) (*informations* ; *information*) (*importante* ; *vital*)

Dans cette première évaluation, deux algorithmes simples seront testés : l'algorithme d'association maximale AMAX, et l'algorithme de meilleure affectation biunivoque ABIJ (BIJ comme bijective).

Dans un premier temps, ces algorithmes ne prendront pas en compte la possibilité d'appariement vide. Ils chercheront donc à apparier *toutes* les unités d'un couple de phrases, y compris les occurrences résiduelles (au sens défini p. 362), ce qui diminuera la précision de façon non négligeable (rappelons que ces unités représentent, pour notre corpus, de 21 % à 27 % des formes simples). Nous étudierons dans un prochain chapitre la mise en œuvre de techniques de filtrage, permettant d'éliminer ces appariements résiduels afin d'améliorer la précision sans diminution importante du rappel.

Notons que l'algorithme AMAX est dissymétrique, dans la mesure où il attribue un correspondant à toutes les unités du texte source, la réciproque n'étant pas vraie. Pour cet algorithme, c'est donc le résidu du texte source qui risque d'affecter les résultats.

Quant à l'algorithme ABIJ, de par sa logique de biunivocité, il génère un nombre de couples égal au nombre d'unités de la plus courte des deux phrases appariées. Cet algorithme devrait donc être un peu moins sensible au bruit dû aux occurrences résiduelles.

Pour des raisons d'efficacité, nous avons légèrement simplifié l'implémentation des deux précédents algorithmes :

- D'une part, quand une unité réalise plusieurs occurrences dans une même phrase source, on ne tient compte que d'une seule de ces occurrences. En d'autres termes, dans notre implémentation, une même unité source ne peut apparaître que dans un seul appariement pour un même couple de phrases. Cette approximation diminue certes le rappel des extractions, mais la mise en œuvre en est plus simple (pour une question de gestion des arbres binaires permettant de classer les couples). Environ 8,3 % des couples de référence concernent des unités déjà appariées dans le même binôme, et ne pourront donc être extraits. Avec cette limitation, le rappel de nos extractions sera donc limité à 91,7 %.

- D'autre part, les unités dépassant 5 000 occurrences dans la totalité du corpus d'apprentissage ne sont pas prises en compte. Cela concerne 29 unités en anglais et 38 en français, essentiellement des mots outils et des signes de ponctuation[202]. Cette réduction permet une économie substantielle de calcul et d'espace mémoire[203]. Afin de déterminer avec précision les conséquences de cette réduction sur les résultats, nous effectuerons tout de même une extraction des correspondances avec la totalité des unités (cf. infra, § III.3.3.2.).

De par ces simplifications, 30,2 % des couples de référence ne pouvant être considérés, le rappel sera finalement limité à 69,8 %.

III.3.2.2 Mesures d'évaluation

Les résultats de ces trois extractions seront évalués de manière classique, par comparaison avec les correspondances de référence.

---

[202] En voici la liste, pour les formes simples, entre crochets et séparés par un espace :
- en anglais [*1 ' - $* ( ) , . *? a and are be by Commission Community for in is Member of on Subject that The the to which with this*]
- en français [*1 ' - ( ) , . ? a à au aux ce Commission d dans de des du elle en est et l la La le les n º Objet par pour que qui sur un une*]

[203] On aboutit ainsi à une diminution de 1 138 972 enregistrements pour la table destinée à enregistrer les comptes de cooccurrence.

Pour les deux premières tâches, l'évaluation est simple, car les unités appariées sont les mêmes que dans les couples de référence. Si l'on note C l'ensemble des couples de lexies à évaluer, C_{ref} l'ensemble des couples de référence, on calcule ainsi les mesures de précision, rappel et F-mesure :

$$P = \frac{\left|C \cap C_{ref}\right|}{\left|C\right|} \quad R = \frac{\left|C \cap C_{ref}\right|}{\left|C_{ref}\right|} \text{ et } F = \frac{2 \times (P \times R)}{(P + R)} \tag{97}$$

Dans le cas des correspondances entre formes simples, l'évaluation est un peu plus délicate : par rapport aux correspondances de référence, on obtient parfois des correspondances fragmentaires, lorsque des parties de lexies complètes sont appariées. Pour évaluer ce type de configuration, nous optons pour une certaine tolérance : dans le calcul de la précision, un couple de formes simples est considéré comme valide s'il est inclus dans un couple de lexies appariées manuellement. Certes, cela peut donner lieu à la validation de correspondances peu orthodoxes : comme (*à* ; *because*) qui se trouve inclus dans le couple (*à cause de* ; *because*). Mais dans la mesure où nous mettons en œuvre cette dernière tâche à titre de comparaison, et non dans le but d'obtenir des correspondances exactes, cela ne pose pas problème. Pour le calcul du rappel, le dénominateur correspond au nombre de formes simples engagées du côté source dans les couples de référence.

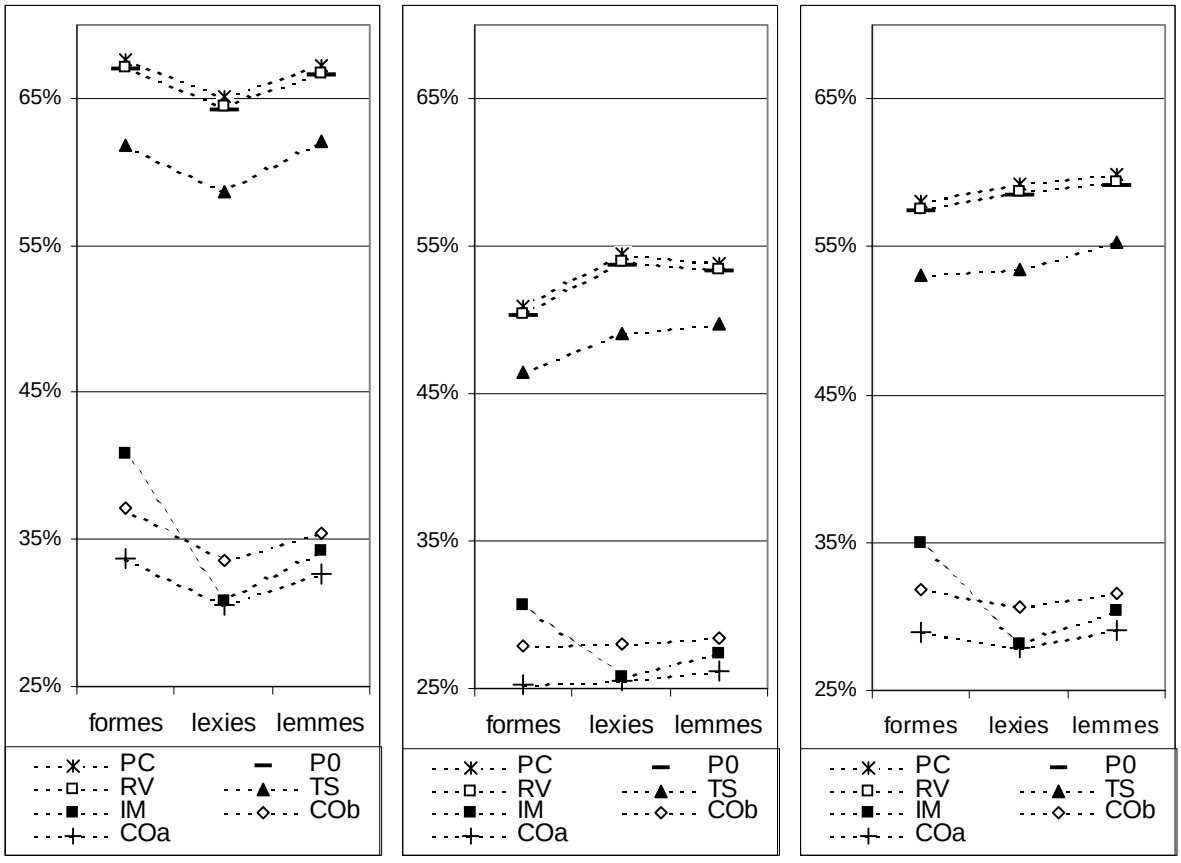
Notons qu'avec ce mode de calcul, nous divergeons des méthodes d'évaluation mises au point pour ARCADE. Dans ce projet, le rappel et la précision sont calculés au niveau de chaque appariement, en comptant pour chaque correspondance les mots corrects et les mots manquant par rapport à la référence. Par exemple, si la référence contient la correspondance (*a* ; *a*',*b*',*c*') et que l'extraction automatique propose d'apparier (*a* ;*a*',*d*') ce couple obtient une précision de 1/2 et un rappel de 1/3. Les résultats globaux sont ensuite obtenus en moyennant les résultats de tous les appariements. En outre, un appariement vide, s'il est correct, donne lieu à un rappel et une précision de 1.

Dans notre évaluation, une correspondance ne peut donner lieu qu'à trois cas de figure, par rapport à la référence : correcte, incorrecte, manquante. Ensuite, précision et rappel sont calculés de façon globale, à partir des comptes globaux de ces trois cas de figure. En outre, les appariements vides de la référence, i.e. le résidu de traduction, ne rentrent pas dans le compte des appariements corrects. C'est un choix délibéré : les

appariements vides n'apportant pas d'information intéressante, dans l'exploitation ultérieure des correspondances, nous avons décidé de ne pas les évaluer positivement.

III.3.2.3 Algorithme d'affectation maximale (AMAX)

Nous avons implémenté les sept indices dans l'algorithme AMAX, et évalué le résultat de chaque extraction.



Précision *Rappel* *F-mesure*

figure 43
Résultats des indices avec l'algorithme AMAX

Bien que ces extractions ne concernent que les binômes constituant le corpus de référence, les comptes d'occurrences et de cooccurrences sont effectués sur la totalité du corpus JOC. Nous ferons donc la distinction entre *corpus d'évaluation* (limité aux binômes

du corpus de référence) et *corpus d'apprentissage*[204], à partir duquel on calcule les indices : on peut interpréter le corpus d'évaluation comme un échantillon permettant de donner une estimation des résultats des extractions si elles étaient lancées sur la totalité du corpus d'apprentissage.

Nous fournissons en annexe, tableau 99, les valeurs numériques de *P*, *R* et *F* pour ces extractions. La figure 43 donne un aperçu de ces résultats.

Notons que *P* et *R* sont étroitement corrélés : les indices obtenant une précision basse obtiennent aussi un rappel bas. Cette corrélation s'explique par le fait que le nombre de couples extraits est à peu près constant pour tous les indices. Ces premiers résultats nous permettent d'effectuer une comparaison directe des indices, dans la mesure où une certaine hiérarchie semble se dessiner :

- On discerne nettement trois groupes d'indices : IM et CO, d'abord, qui occupent le bas du tableau avec des résultats environ deux fois inférieurs aux meilleurs indices ; TS, ensuite qui donne des résultats intermédiaires ; et P0, RV et PC qui se détachent nettement au-dessus des autres indices.

- Dans le groupe de tête, RV et P0 ont des comportements identiques, à tel point que leurs deux courbes se chevauchent. Quant à PC, il se situe en moyenne un demi-point au-dessus. L'information apportée par les cognats semble donc se cumuler avantageusement, quoique dans une faible mesure, avec l'information liée aux distributions.

- Les meilleurs résultats de COb par rapport à COa confirment notre hypothèse de prépondérance du rappel sur la précision, lors de l'identification des cognats. En effet, la nature même de l'algorithme permet de filtrer un surcroît de bruit, car peu de cognats entrent en concurrence dans un même couple de phrases. Par ailleurs,

[204] Ce terme recèle une ambiguïté : il ne s'agit pas d'un apprentissage effectué à partir des correspondances manuellement extraites, comme dans le cas de certains modèles paramétriques – mais de l'apprentissage des indices formels, qui n'utilisent aucune information obtenue manuellement.

le choix de la valeur maximale de l'indice permet de trancher harmonieusement entre deux associations concurrentes.

– IM est légèrement moins bon que COb, sauf pour l'extraction FS. On peut imputer ce phénomène au fait que IM a tendance, avec AMAX, à concentrer toutes les correspondances vers les mêmes unités cibles, de faible fréquence. Or, les unités polylexicales étant nécessairement moins fréquentes que leurs composants, leur présence tend à augmenter la proportion d'hapax, au risque d'accentuer les surévaluations de IM. Ce qui explique les moins bons résultats de IM pour les extractions LEX et LEM.

Si l'on néglige ce comportement fluctuant de IM, cette première évaluation aboutit à un nouveau classement des indices :

$$COa < IM \leq COb < TS < P0 < RV < PC$$

Ce classement confirme peu ou prou la hiérarchie obtenue précédemment, avec MRP (moyenne du rapport des probabilités entre couples de références et couples quelconques), à une différence près : cette fois PC surpasse légèrement RV.

Nous n'avons pas représenté graphiquement les résultats de l'indice aléatoire AL, qui figurent dans le tableau 55 :

Unités	***P***	***R***	***F***
FS	8,4 %	6,3 %	7,2 %
LEX	6,8 %	5,7 %	6,2 %
LEM	7,5 %	6,0 %	6,7 %

tableau 55 : *résultats avec l'indice aléatoire*

Cet indice nous permet d'étalonner notre évaluation, en donnant un « plancher » pour chacune des tâches. Il nous indique par exemple que le mode d'évaluation favorise légèrement la tâche FS au niveau de la précision, car plusieurs appariements sont également acceptables à l'intérieur des unités polylexicales.

On peut donc supposer que les meilleures précisions obtenues dans la tâche FS sont imputables à cette différence de calcul. Malgré cet avantage, la F-mesure globale de

l'extraction FS est légèrement inférieure à celle des autres extractions pour le groupe des meilleurs indices. Ceci pourrait confirmer notre hypothèse préalable, que l'identification des unités non compositionnelles améliore le comportement des indices. Mais les différences sont trop faibles pour pouvoir conclure avec certitude.

De même, entre les extractions LEX et LEM, les résultats sont mitigés : le recours à la lemmatisation semble améliorer sensiblement la précision des résultats, mais au détriment du rappel.

La diminution du rappel est vraisemblablement due à une simplification de la mise en œuvre de notre algorithme, qui néglige les unités répétitives dans un même binôme. En recourant aux lemmes, les effets de cette simplification sont aggravés, car la réduction morphologique augmente le nombre d'unités récurrentes dans une même phrase. Si notre interprétation est bonne, il semble donc que les résultats de l'extraction LEM soient sensiblement supérieurs.

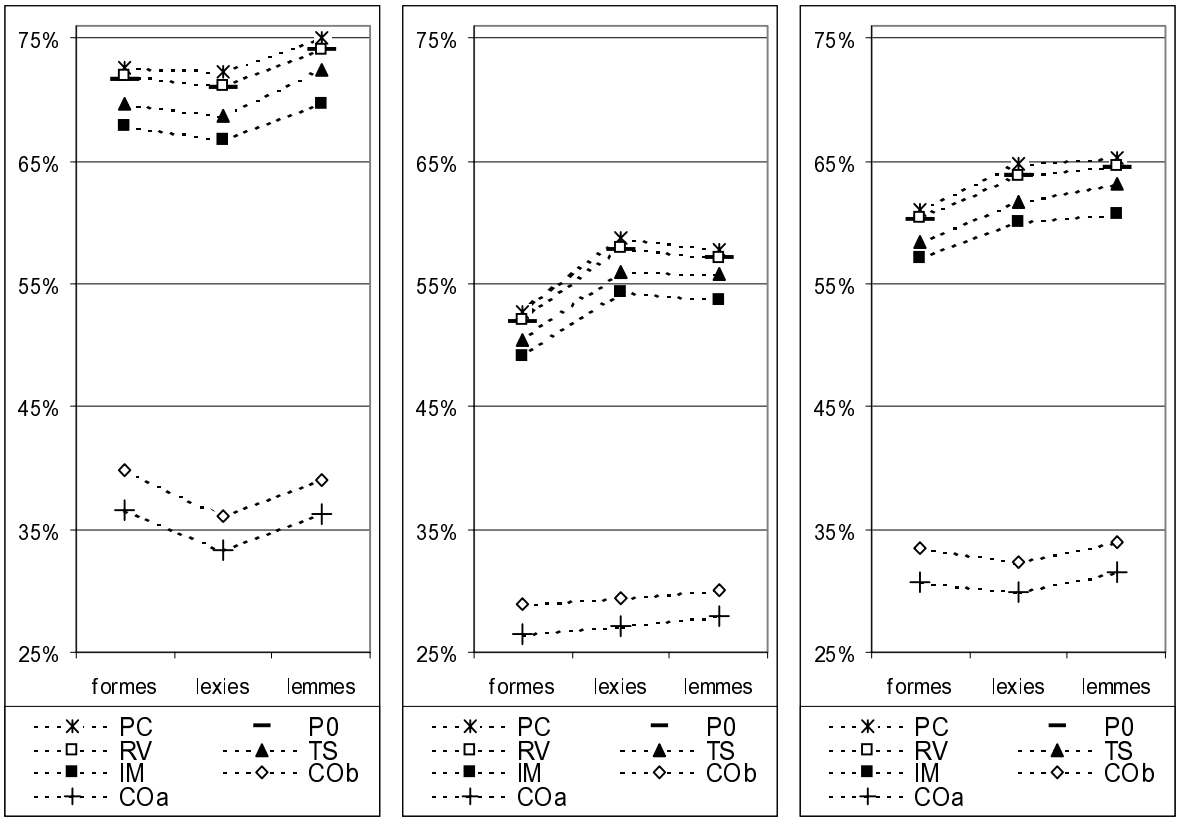


figure 44
Résultats des indices avec l'algorithme ABIJ

III.3.2.4 Algorithme de meilleure affectation biunivoque (ABIJ)

Après avoir implémenté les mêmes indices et les mêmes tâches dans l'algorithme ABIJ, nous obtenons les résultats représentés par la figure 44 (cf. tableau 100 de l'annexe).

Comme le montre le tableau 56, tous indices et toutes tâches confondus, les résultats moyens gagnent avec ABIJ environ 10 points pour la précision, et environ 6 points pour le rappel.

Algorithme	***P***	***R***	***F***
AMAX	55,0 %	43,8 %	48,7 %
ABIJ	65,7 %	50,6 %	57,2 %

tableau 56 : *comparaison des algorithmes*

L'amélioration est nette, ce qui semble confirmer que le critère de biunivocité permet de tirer un meilleur parti des indices. Il serait excessif d'y voir une confirmation, dans l'absolu, de l'hypothèse de biunivocité de I. D. Melamed (1998a : 5) : comme nous l'avons vu précédemment, dans la réalité, celle-ci à ses limites. Simplement, la plus grande efficacité de ABIJ s'explique par sa capacité à filtrer les correspondances indirectes (cf. p. 409) dues à la conjonction de cooccurrences parallèles bilingues et de cooccurrences monolingues sur l'axe syntagmatique.

D'ailleurs, l'indice qui tire le meilleur bénéfice de ABIJ, est justement celui qui était le plus sensible aux correspondances indirectes : c'est l'information mutuelle, dont la F-mesure moyenne (pour les trois tâches) passe de 31 % à 59 %. En comparaison, les résultats de CO sont quasi stationnaires, avec une progression d'environ 2 % : les couples de cognats ne subissant qu'une concurrence très faible avec d'autres couples tirés du même binôme, il est naturel que ABIJ ne modifie les résultats que dans une faible mesure ; en outre, le problème des correspondances indirectes ne se pose pas pour les cognats.

On aboutit donc à un nouveau classement des indices où seule la place de IM a changé :

$$COa < COb < IM < TS < P0 \leq RV < PC$$

Dans la comparaison des trois tâches, les conclusions précédentes restent valides : la précision de FS est légèrement surévaluée du fait de notre méthode d'évaluation, et le

rappel de LEM est à l'inverse sous-estimé de par les simplifications de notre implémentation. Globalement, comme le confirme la F-mesure, les résultats des différentes tâches peuvent être classés dans l'ordre : FS < LEX < LEM.

Les meilleurs résultats sont donc obtenus en lemmatisant, avec l'indice combiné PC dans le cadre de la meilleure affectation biunivoque :

	P	***R***	***F***
PC avec ABIJ pour LEM	75,0 %	57,8 %	65,3 %

tableau 57 : *meilleurs résultats*

Notons que pour une telle extraction, cherchant à apparier toutes les unités des phrases alignées (sauf celles de plus de 5 000 occurrences), on s'approche de la précision maximale, dans la mesure où seulement 85 % des lexies anglaises et 82 % des lexies françaises considérées ont vraiment une correspondance. Du fait de ce résidu, la précision optimale doit avoisiner 80 %.

De même, pour le rappel, la meilleure valeur qui puisse être atteinte n'est pas de 100 %, mais d'environ 70 %, car comme nous l'avons déjà signalé, notre algorithme néglige d'emblée certaines correspondances, avec des unités répétées ou dépassant 5 000 occurrences. Ainsi, pour le rappel, la marge de progression est encore importante, puisque environ 12 % des couples pouvant être extraits sont manquants.

III.3.3 Extractions partielles et extraction intégrales

Il peut être intéressant de mesurer la proportion de mots outils qui entrent dans ces couples manquants, car l'extraction des correspondances a plus souvent vocation à associer des « mots pleins » que des unités de nature grammaticale.

III.3.3.1 Elimination des mots outils et très fréquents

Nous avons donc établi, pour les deux versions de notre échantillon d'évaluation, une liste de mots outils, ainsi que des mots très fréquents. Ces unités incluent des articles (fr. *le*,

un, angl. *the*, *an*, …), des déterminants (fr. *certain*, *chaque*, angl. *all*, *no*, …), des adjectifs démonstratifs possessifs et numéraux (fr. *ces*, *leurs*, *trois*, angl. *its*, *first*, …), des prépositions (fr. *à*, *avec*, *de*, *pour*, angl. *to*, *in*, *amongst*, …), des verbes support et modaux (fr. *être*, *avoir*, *falloir*, *devoir*, angl. *could*, *was*, …), des pronoms (fr. *il*, *qui*, *celui*, angl. *both*, *itself*, …), des conjonctions (fr. *et*, *aussi*, angl. *however*, *then*, *moreover*, …) et des adverbes très fréquents (fr. *même*, *très*, *maintenant*, angl. *well*, *already*, …).

Les listes complètes de mots outils sont données en annexe (§ A-VII) : on aboutit environ à 300 formes différentes en français (en tenant compte des majuscules) et 250 en anglais. Ces listes ne prétendent pas être exhaustives ni totalement cohérentes, dans la mesure où nous avons éludé la définition rigoureuse de ce qu'est un mot outil : les critères de dépendance syntaxique (suivant l'opposition catégorématique *vs* syncatégorématique) et de contenu sémantique (suivant l'opposition mots pleins *vs* mots vides) peuvent donner lieu à des interprétations variables, et là encore on se heurte au problème de zones grises à l'intérieur desquelles il est difficile de trancher. Nous avons appliqué intuitivement ces deux critères plus un troisième, la fréquence, afin de retenir également certaines unités appartenant en quelque sorte au lexique de base (fr. *aujourd'hui*, angl. *early*, …).

Ces listes une fois constituées, nous avons éliminé toutes les unités concernées lors de l'extraction des couples et de l'évaluation (sauf à l'intérieur de lexies polylexicales, bien entendu, celles-ci formant un tout). 32 % des couples de références ont été éliminés. Le tableau 58 donne la répartition des mots outils entre les couples de référence et les lexies résiduelles.

		Couples de référence		***Résidu***	
Anglais	*Lexies*	**9 727**		**2 468**	
	Dont mots outils	2 959	30 %	1 738	70 %
Français	*Lexies*	**9 727**		**4 273**	
	Dont mots outils	2 626	27 %	3 248	76 %
Nombre de couples de références éliminés[205]		3 161	32 %		

tableau 58 : *répartition des mots outils entre les couples de référence et le résidu*

[205] Notons que certains mots outils sont appariés avec des lexies qui ne sont pas comptées comme des mots outils. C'est pourquoi le chiffre global est supérieur à 2 949.

Les résultats obtenus avec ABIJ sont représentés figure 45 (cf. tableau 101 de l'annexe).

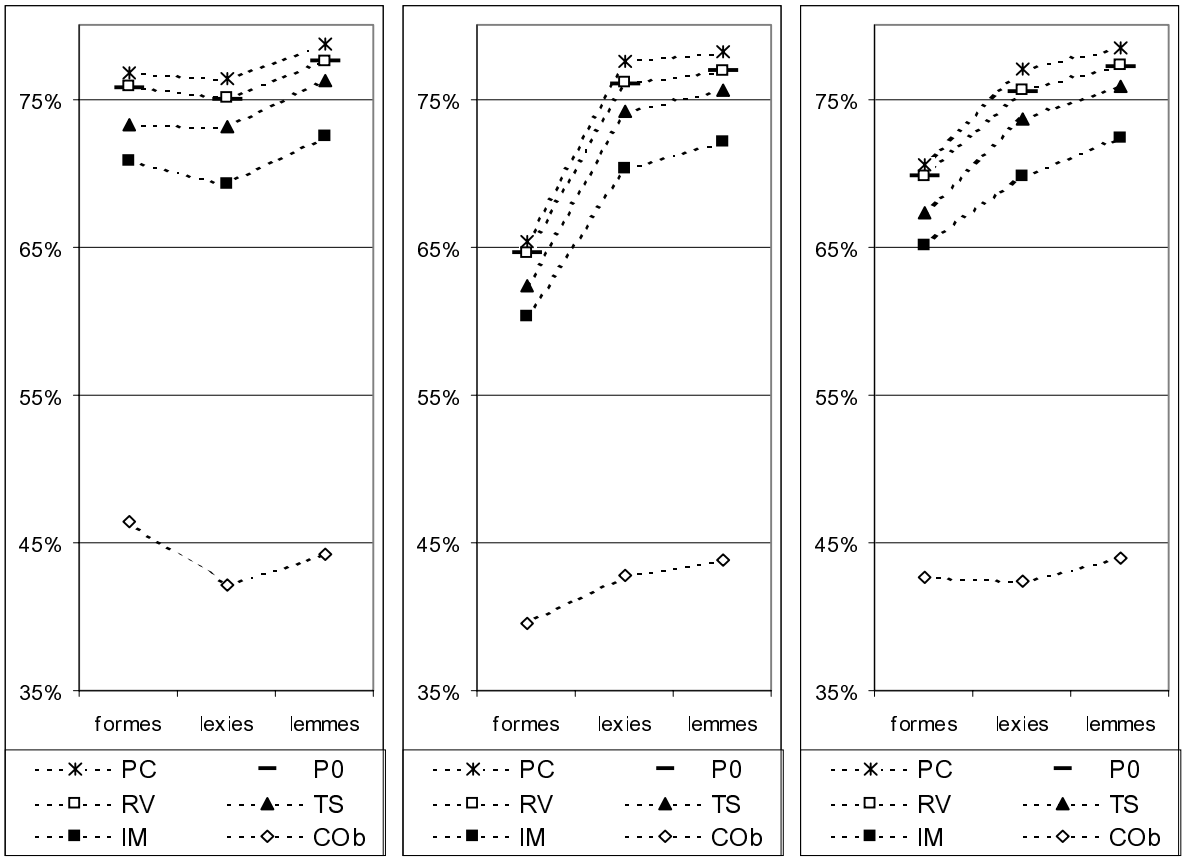


figure 45
Résultats des indices avec l'algorithme ABIJ sans les mots outils

On peut comparer ces résultats avec ceux obtenus précédemment, en prenant la moyenne globale des six indices :

Algorithme ABIJ	***P***			***R***			***F***		
	FS	***LEX***	***LEM***	***FS***	***LEX***	***LEM***	***FS***	***LEX***	***LEM***
Avec les mots outils		65,7 %			50,6 %			57,2 %	
	65,6	64,3	67,4	47,5	52,4	51,9	55,1	57,7	58,6
Sans les mots outils		69,8 %			66,5 %			68,0 %	
	69,8	68,5	71,1	59,4	69,5	70,6	64,2	69,0	70,8

tableau 59 : *comparaison des extractions avec et sans mots outils*

On constate que la progression du rappel moyen est sensible, avec presque 20 % d'augmentation pour l'extraction LEM. Si l'on s'intéresse aux meilleurs résultats obtenus, la progression de *R* avoisine 15 %.

	P	***R***	***F***
PC avec ABIJ pour LEM	78,8 %	78,2 %	78,5 %

tableau 60 : *meilleurs résultats sans les mots outils*

Cette augmentation importante du rappel découle des simplifications de notre implémentation. En effet la plupart des couples négligés ont disparu de l'évaluation, puisqu'ils impliquaient des mots outils. De ce fait, le rappel optimal pouvant être atteint par nos extractions n'est plus de 69,8 %, mais de 91,4 %.

L'amélioration de la précision peut s'expliquer par la diminution de la proportion d'occurrences résiduelles : celles-ci ne constituent plus que 13 % des occurrences considérées (plus exactement 13 % en anglais et 12 % en français). Or, comme on l'a déjà remarqué, une grande part des appariements incorrects est probablement due aux occurrences résiduelles.

A supposer que les couples erronés ne mettent en jeu que des occurrences résiduelles, la précision maximale sans filtrage serait de 87 %. La F-mesure serait alors de 89,7 %. On constate que les résultats du tableau 60 sont situés à un peu plus de 11 % des valeurs optimales.

Par ailleurs, on note que l'amélioration apportée par le recours aux cognats se dessine plus nettement : comme on pouvait s'y attendre, la cognation intéresse plus fréquemment les mots pleins, du moins entre l'anglais et le français.

Il faut préciser que le faible rappel lié à la tâche FS est un artefact de notre évaluation : lorsqu'on considère les formes simples, un certain nombre de mots outils sont éliminés alors qu'ils sont conservés dans les couples de références : ce sont les mots outils inclus dans les unités polylexicales (rappelons que le rappel de FS est calculé sur la base du nombre de formes simples apparaissant dans les unités source des couples de référence).

Enfin, l'écart entre LEX et LEM apparaît plus nettement que précédemment. Cela corrobore nos précédentes hypothèses, selon lesquelles l'écart devrait être encore plus sensible si notre implémentation ne tenait pas compte que d'une seule occurrence pour chaque unité dans un même binôme : cet artefact est ici légèrement estompé du fait de l'absence des mots outils, dont la répétition au sein d'un même couple de phrases est très fréquente.

III.3.3.2 Extractions intégrales

Nous avons voulu déterminer précisément les conséquences de l'abandon des unités, dépassant 5 000 occurrences. Etant donné le surcroît de calculs engendré par la prise en compte de ces unités très fréquentes, nous n'avons lancé qu'une série limitée d'extractions, pour les indices CO, IM, TS et RV et la tâche FS avec l'algorithme ABIJ.

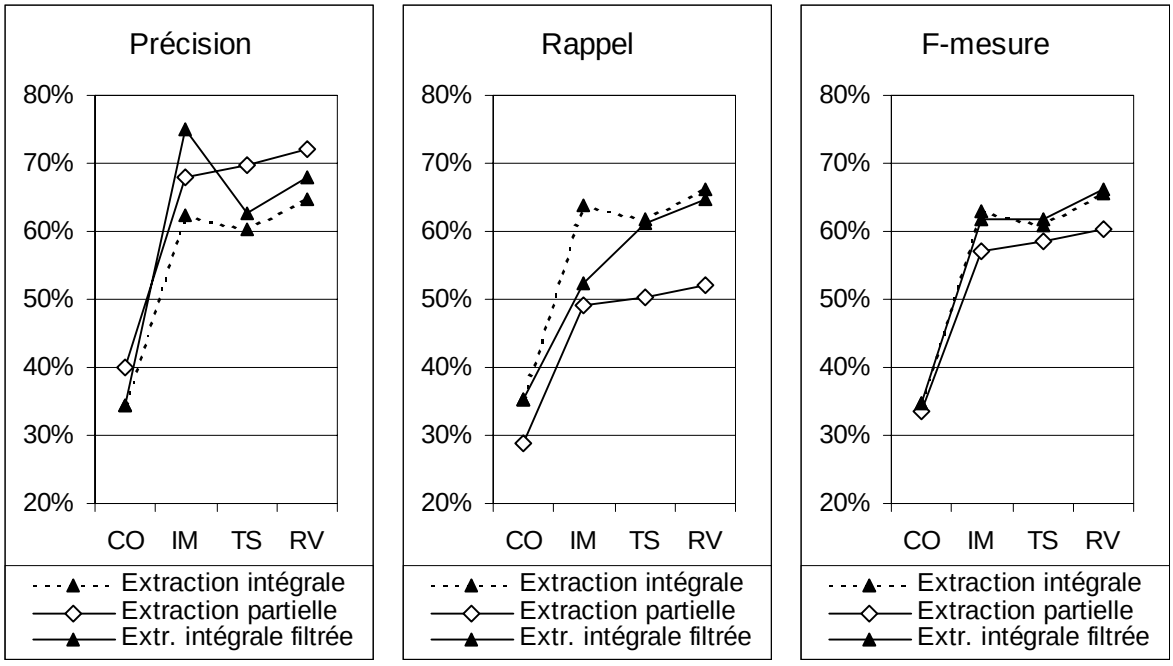


figure 46
Résultats de l'extraction intégrale pour les indices CO, IM, TS et RV avec l'algorithme ABIJ

Comme le montrent les résultats de la figure 46, la prise en compte des unités très fréquentes aboutit à une diminution de la précision et une amélioration du rappel. La F-mesure globale est en légère progression.

Pour que les résultats soient supérieurs, dans l'absolu, il faudrait obtenir une amélioration simultanée de la précision *et* du rappel. Nous avons essayé d'appliquer différents filtrages (dont nous détaillerons plus loin le principe[206]) visant à éliminer les couples erronés, afin de retrouver une précision aussi bonne que celle de l'extraction partielle : mais quels que soit l'indice et la formule de filtrage choisie, nous n'avons pu obtenir une précision égale qu'avec un rappel égal ou inférieur. En d'autres termes, les résultats obtenus avec la totalité des unités ne sont pas meilleurs au titre de la précision et du rappel, mais seulement du rappel et de la F-mesure.

III.3.4 Contrôle des couples erronés

III.3.4.1 Correspondances impliquant les occurrences résiduelles

Afin d'évaluer plus précisément l'incidence négative des occurrences résiduelles, nous avons dénombré les couples erronés qui mettent en jeu au moins une de ces unités. Par exemple, pour l'extraction du tableau 57, on compte 1 227 couples impliquant des occurrences résiduelles, soit 68,7 % des 1 859 couples erronés.

L'effet des occurrences résiduelles peut se situer entre deux extrêmes :

– *Cas de figure 1* : dans le meilleur des cas, les occurrences résiduelles ne se combinent qu'entre elles. Dans ce cas de figure, plus le résidu sera important (i.e. moins la compositionnalité traductionnelle sera forte), et plus la précision sera faible, mais le rappel n'en sera pas affecté.

[206] Cf. le chapitre III.3.4.2. La meilleure valeur de *F* est obtenue avec un filtrage absolu de paramètre 0,5 (troisième courbe de la figure 46).

– *Cas de figure 2* : dans le pire des cas, les occurrences résiduelles se combinent toutes avec des unités possédant une correspondance. Cette fois, plus le résidu est grand et plus rappel et précision sont faibles.

Entre ces deux extrêmes, on peut supposer une situation intermédiaire où la distribution des couples erronés se ferait au hasard. Dans cette dernière hypothèse, il est facile d'estimer la proportion théorique des couples erronés contenant des occurrences résiduelles.

Supposons que les phrases *P* et *P'* contiennent respectivement N_r et N_r' occurrences résiduelles. Notons $N_{corrects}$ et $N_{erronés}$ les nombres respectifs de couples corrects et erronés dans l'extraction des correspondances de (*P*,*P'*). Les unités non résiduelles de *P* entrant dans des couples incorrects représentent une proportion de $(L\text{-}N_{corrects}\text{-}\ N_r)/(L\text{-}N_{corrects})$ des unités de *P* susceptibles d'entrer dans des couples incorrects. Le calcul est le même pour *P'*. Si les couples erronés sont tirés au hasard entre ces unités, on doit avoir une proportion p_{res1} de couples impliquant au moins une occurrence résiduelle :

$$p_{res1} = 1 - \left(\frac{(L - N_{correct} - N_r)}{L - N_{correct}} \right) \left(\frac{(L' - N_{correct} - N_r')}{L' - N_{correct}} \right) \quad (98)$$

On peut aussi évaluer la proportion p_{res2} de couples impliquant deux occurrences résiduelles :

$$p_{res2} = \frac{N_r}{L - N_{correct}} \cdot \frac{N_r'}{L' - N_{correct}} \quad (99)$$

Pour chaque extraction évaluée, nous avons calculé ces proportions théoriques. Il s'avère, aux vues des résultats (cf., tableau 102 de l'annexe), que les associations erronées semblent bien être distribuées au hasard entre les occurrences résiduelles et les autres.

Une traduction très éloignée de l'original, sur le plan de la compositionnalité traductionnelle, risque donc d'aboutir à des résultats faibles sur le plan du rappel (outre la précision qui se dégrade nécessairement), dans la mesure où les occurrences résiduelles rentrent dans la compétition avec les mêmes chances que certaines unités avec correspondance.

Afin d'illustrer les effets négatifs du résidu, nous avons lancé une série d'extractions en supprimant préalablement toutes les occurrences résiduelles. Bien sûr, ce genre d'opération est artificiel, dans la mesure où dans la pratique on ne connaît jamais les occurrences résiduelles avant de lancer une extraction. Cela permet néanmoins de simuler le cas idéal où la compositionnalité traductionnelle serait parfaite (i.e. une traduction mot à mot). Les résultats pour la tâche LEM, obtenus avec l'algorithme ABIJ sont représentés figure 47 (cf. tableau 103 de l'annexe).

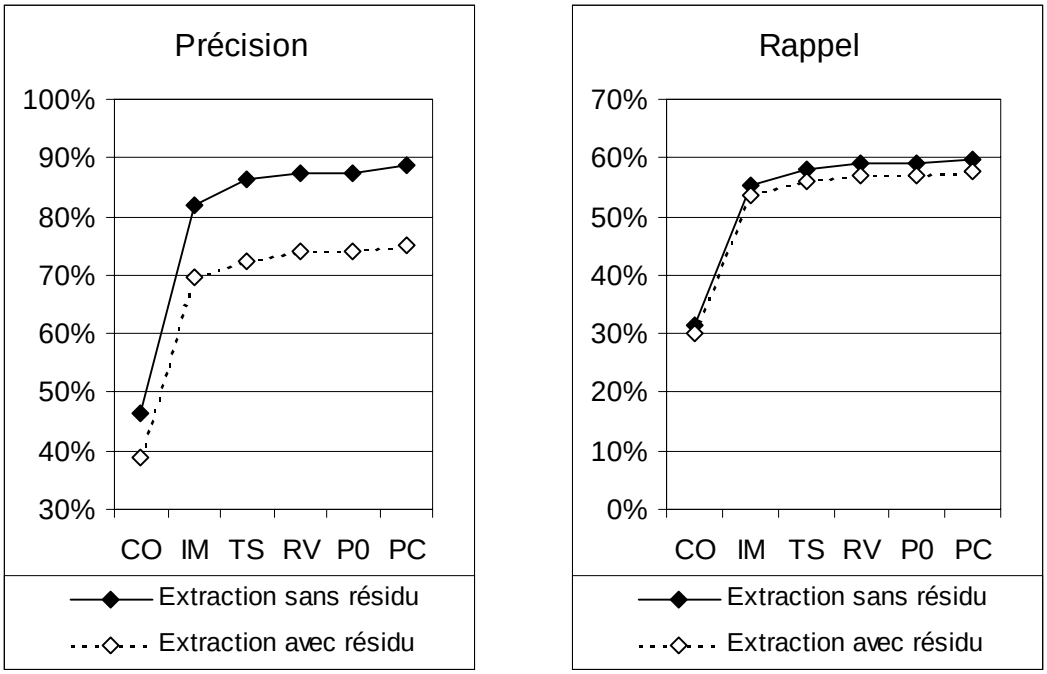


figure 47
Résultats comparés des extractions LEM avec et sans résidu de traduction

On constate que précision et rappel se stabilisent à environ 10 % des valeurs optimales (100 % et 69,8 %). C'est la précision qui a surtout profité de l'élimination des occurrences résiduelles : le rappel, quant à lui est pratiquement stationnaire. Ce résultat semble indiquer que les couples corrects sont les mêmes dans les deux types d'extraction. Ce qui a changé, dans l'extraction sans résidu, c'est que les couples contenant deux occurrences résiduelles (dont la proportion est donnée par p_{res2}) disparaissent et cessent d'affecter la précision. Quant aux occurrences non-résiduelles appariées avec une occurrence résiduelle, elles se recombinent entre elles, mais le plus souvent de façon

incorrecte, de sorte que le rappel n'augmente presque pas. Comme le montre le tableau 61, on observe le même phénomène si l'on fait une extraction intégrale, c'est-à-dire incluant les unités de plus de 5 000 occurrences.

Extraction FS incluant les unités de plus de 5 000 occurrences	***P***	***R***	***F***
RV avec résidu	64,7 %	66,3 %	65,5 %
RV sans résidu	78,6 %	69,2 %	73,6 %

tableau 61 : *extraction intégrale avec et sans résidu* (*algorithme ABIJ*)

Ainsi, on aboutit aux mêmes phénomènes que dans le cas de figure 1, mais pour une autre raison : non pas parce que les occurrences résiduelles ne se combinent qu'entre elles, mais parce qu'une proportion fixe d'unités possédant une correspondance se combinent de manière aléatoire, avec ou sans résidu (de même qu'une proportion fixe d'unités est appariée correctement, avec ou sans résidu). Nous verrons plus loin qu'il s'agit essentiellement d'unités de faible fréquence : pour ces unités, les indices statistiques touchent évidemment à leur limite, et seul le recours à d'autres sources d'information, comme la cognation ou des données dictionnairiques, peut permettre une progression des résultats.

Par ailleurs, notons que pour les extractions LEX et LEM, les valeurs empiriques de p_{res1} sont toutes inférieures aux proportions attendues : en d'autres termes, il semblerait que les indices aient une légère propension à apparier entre elles (de manière incorrecte) les unités n'appartenant pas au résidu. Cette tendance est plus fortement marquée pour l'indice CO : vraisemblablement parce que cet indice est inopérant avec la plupart des mots outils, et que la proportion de mots outils résiduels est forte.

Pour les autres indices, ce phénomène est peut être dû à l'effet des associations indirectes, qui aboutissent à des permutations entre des couples d'unités fréquemment cooccurrentes sur l'axe syntagmatique. Un indice assurant une meilleure inhibition des associations indirectes permettrait sans doute d'améliorer les résultats.

Mais avec les méthodes présentées jusqu'à présent, cette progression n'est possible que dans certaines limites, car les occurrences résiduelles impliquent une proportion incompressible de couples erronés. Nous allons voir comment ces « inévitables » correspondances erronées peuvent être éliminées au moyen de filtres adéquats.

III.3.4.2 Filtrage des résultats

Par filtrage, nous désignons toute opération permettant d'éliminer les couples erronés issus d'une extraction. Une méthode de filtrage efficace se caractérise par sa capacité à éliminer le plus possible de couples erronés en conservant le plus possible de couples corrects : en d'autres termes, il s'agit d'augmenter la précision des résultats en évitant d'en affecter le rappel. Bien entendu, le filtrage doit s'appuyer sur l'information disponible à l'issue de l'extraction : la valeur de l'indice affectée aux couples de d'unités correspondantes.

Pour chaque indice, nous avons testé trois méthodes de filtrage :

- *Filtrage relatif* : étant donné la nature de nos algorithmes, pour chaque couple de phrases, l'extraction fournit une série de correspondances ordonnées de façon décroissante en fonction des valeurs de l'indice. Le filtrage relatif consiste à conserver les meilleurs couples, suivant différentes proportions. Nous avons testé les proportions suivantes : 80 %, 60 %, 40 %, 20 %.

- *Filtrage absolu* : cette fois le seuil de rejet n'est plus relatif, lié au classement des couples à l'intérieur de chaque binôme, mais fixé en valeur absolue. Pour chaque indice, nous avons calculé la moyenne des valeurs obtenues pour des couples quelconques :

CO	***IM***	***TS***	***RV***	***P0***	***PC***
0,0124	1,33	1,68	8,59	10,65	10,66

tableau 62 : *moyennes des indices*

Nous avons ensuite éliminé tous les couples obtenant une valeur inférieure à *x* fois la moyenne, avec différentes valeurs de *x* :

$x = 0{,}25$ $x = 0{,}5$ $x = 1$ $x = 2{,}5$ $x = 4$ $x = 6$ $x = 10$

Nous avons donc testé sept seuils différents pour chaque indice.

– *Filtrage différentiel* : outre la valeur absolue, il existe un autre indicateur de la fiabilité d'un couple de correspondances : il s'agit de l'existence d'associations concurrentes, mettant en jeu une des deux unités du couple. En effet, dans le cas d'associations indirectes, on observe que les valeurs liées aux correspondances indirectes et directes sont très voisines. Cette proximité peut donner lieu au rejet du couple : c'est le principe du filtrage différentiel. Ainsi, pour chaque couple extrait, on calcule le rapport entre la valeur obtenue et la deuxième meilleure valeur atteinte par tout couple concurrent (i.e. qui met en jeu une des deux unités du couple). Plus ce rapport se rapproche de 1, et plus le couple sera considéré comme suspect. On élimine donc tous les couples pour lesquels le rapport est inférieur à un certain seuil *s*. Les seuils suivants ont été testés :
$s = 1{,}05$ $s = 1{,}2$ $s = 1{,}5$ $s = 2$ $s = 2{,}5$ $s = 3$ $s = 4$

Nous avons implémenté les trois méthodes de filtrage, avec ces différents paramètres, sur tous les résultats obtenus jusqu'à présent, concernant les six indices, les trois tâches et les deux types d'algorithme (cf. tableau 104 à tableau 106 et figure 79 à figure 81 de l'annexe, pour les résultats liés aux filtrages de la tâche LEX).

Il apparaît que les comportements de ces méthodes ne dépendent pas du type de tâche. De même, on ne note aucune différence sensible vis-à-vis de l'algorithme employé.

En revanche, les différents indices ne réagissent pas de la même manière au filtrage :

– IM est le seul indice à connaître une baisse de précision, avec les trois méthodes, lorsque le filtrage devient très sélectif (sauf avec AMAX et le filtrage relatif). Cela confirme que les plus fortes valeurs de IM n'indiquent pas les correspondances les plus fiables. Le maximum de *F*, 61,6 %, est atteint avec le filtrage absolu (pour $x = 0{,}5$).

– Avec les filtrages différentiel et absolu, l'indice CO évolue par palier : stabilité puis progression importante de la précision et stabilité à nouveau, entre 80 et

90 %. Ce palier manifeste l'écrémage des couples correspondant au cas 0, qui ne sont pas des cognats. Du côté du rappel, la diminution accompagnant l'augmentation de la précision est faible : le rappel se stabilise aux alentours de 25 %. Le filtrage aboutit donc assez vite à une nette amélioration de la F-mesure. Il permet en outre de montrer clairement les possibilités et les limites du recours aux cognats : avec une précision intéressante, avoisinant 90 % pour le filtrage différentiel, les cognats permettent d'obtenir tout de même environ un quart des correspondances. Le maximum de *F*, 37,9 %, est atteint avec le filtrage absolu ($x = 1$).

– Pour TS, RV, P0 et PC l'évolution de la précision et du rappel est progressive, avec une diminution du rappel plus rapide que l'augmentation de la précision. Globalement, la F-mesure régresse lentement, après avoir atteint un maximum avec le filtrage absolu, pour $x = 0{,}5$. Les valeurs maximales de *F* de TS, RV, P0 et PC sont respectivement de 61,7 %, 65,1 %, 65 % et 66 %.

Globalement, les méthodes de filtrage ont donc des profils différents : le filtrage relatif implique une évolution graduelle et continue de *P* et *R*, tandis que les deux autres méthodes réalisent des variations beaucoup plus brusques. En ce qui concerne *F*, les filtrages relatifs et absolus permettent une très légère amélioration par rapport aux résultats bruts, ce qui n'est pas le cas du filtrage différentiel (sauf pour COb).

Si l'on compare les méthodes de filtrages en représentant, pour chaque paramètre, les valeurs conjointes de précision et de rappel, la spécificité des méthodes apparaît plus nettement : la figure 48 permet de visualiser, indice par indice, les mérites respectifs de chaque filtrage. Toute portion de courbe située dans le quart supérieur gauche par rapport à une autre portion de courbe, indique une supériorité à la fois en précision et en rappel.

Même si les courbes sont incomplètes du côté des valeurs basses du rappel, on constate que pour tous les indices, le filtrage absolu est supérieur si l'on cherche à maintenir un rappel élevé. En revanche, pour TS, RV, P0 et PC, il apparaît clairement que si l'on accepte un rappel inférieur à 50 %, le filtrage différentiel est meilleur (pour CO et IM, c'est un peu moins net, et les points disponibles ne permettent pas de conclure). Le filtrage relatif, globalement, est moins bon.

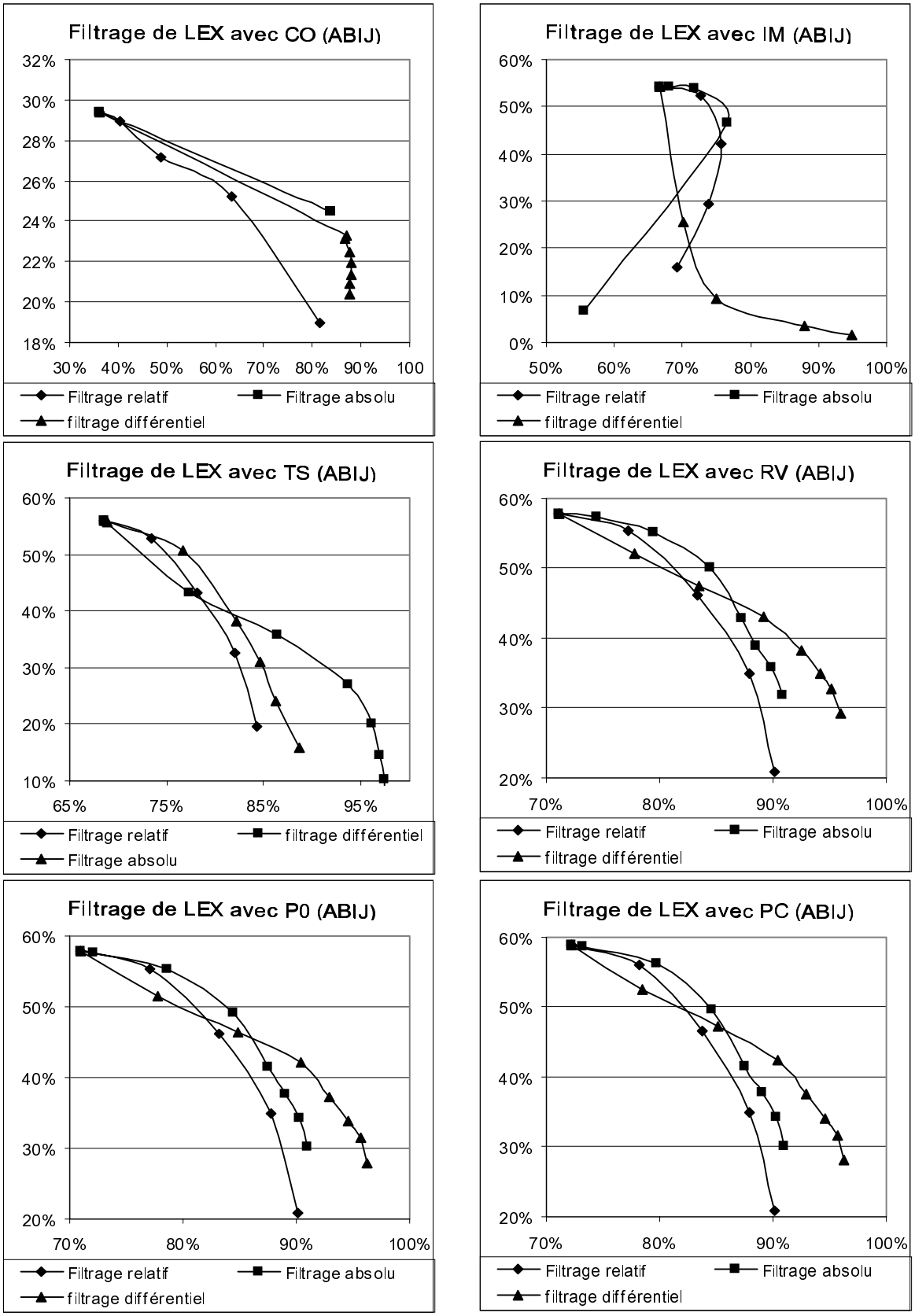


figure 48 : *évolution du rappel en fonction de la précision pour les trois types de filtrage indice par indice*

Le filtrage absolu et le filtrage différentiel peuvent donc avoir une utilisation complémentaire, suivant le type d'extraction qu'on désire mettre en œuvre : pour une extraction la plus complète possible, avec une bonne précision, le recours au filtrage absolu semble plus adapté ; pour une extraction visant à produire des correspondances peu nombreuses mais très fiables, le filtrage différentiel est plus performant. Le filtrage relatif, même s'il est globalement inférieur, présente toutefois un avantage : il permet une évolution graduelle de *P* et *R*, sans discontinuité, et par conséquent un meilleur contrôle de l'équilibre des deux mesures.

III.3.5 Complexité des calculs

Notons que les techniques précédemment présentées, bien que parmi les plus simples, sont relativement coûteuses en calculs et en espace mémoire. Afin de déterminer si elles sont applicables à des corpus plus importants que le corpus JOC, essayons de donner un ordre de grandeur de la quantité de calculs et d'espace mémoire requis en fonction de la dimension du corpus.

III.3.5.1 Production et stockage de la table des cooccurrences

Dans un premier temps, avant même de lancer les différentes extractions, il faut générer une table destinée à stocker les informations de cooccurrence (les *Cooc*(*u*,*u*')). Pour des vocabulaires de taille respective *V* et *V*', il y a donc *V*·*V*' comptes de cooccurrence à stocker, si l'on considère que chaque unité en langue *L* est susceptible de « cooccurrer » avec chaque unité en langue *L*'. Pour le corpus JOC cela fait donc 29 779·36 003 = 1 072 133 337 enregistrements. Sachant qu'il faut en moyenne 28 octets par enregistrement, en comptant les index permettant d'y accéder efficacement, nous arrivons à un total de 30 019 733 436 octets.

En outre, pour calculer le nombre de cooccurrences de deux unités, il faut comparer leurs vecteurs d'occurrence (la suite de leurs coordonnées d'occurrence). Ceux-ci comptant en moyenne 35 occurrences, il y aura approximativement 35 comparaisons pour chaque couple comparé. D'où environ 35 milliards de comparaisons pour obtenir la table des cooccurrences.

A supposer que le vocabulaire soit constant avec l'augmentation de *n*, le nombre de comparaisons à effectuer augmenterait en raison de $n \cdot 10^9$. Mais, comme le remarque Muller (1968 : 156), c'est « une vérité d'expérience aussi bien qu'une évidence linguistique » que « *V* ne cesse pas de croître » avec la dimension du corpus. On peut supposer que la quantité de calcul requis croît de manière bien plus importante que l'estimation précédente.

Il n'existe cependant pas d'estimation théorique de l'étendue du vocabulaire d'un texte en fonction de sa taille : « (…) il est impossible ou difficile d'assigner une valeur théorique à l'étendue d'un vocabulaire, de calculer une espérance mathématique de cette valeur, à laquelle on pourrait ensuite comparer une valeur observée pour déterminer si le texte à un vocabulaire riche, normal ou pauvre. » (Muller, 1968 : 172). La variable *V* dépend trop étroitement des caractéristiques singulières du corpus pour qu'il soit possible d'ériger une norme fiable.

Il nous est difficile, dès lors, d'estimer précisément en fonction de *n*, la complexité en temps et en espace pour la production et le stockage d'une table de cooccurrence, avec la méthode présentée.

Une autre solution permet heureusement de minorer les estimations avancées : dans la réalité, il est évident que toutes les unités de T ne cooccurrent pas avec toutes les unités de T'. Pour tirer parti de ce fait, la construction de la table peut être effectuée de façon plus efficace, en ne comparant que les unités cooccurrentes à l'intérieur de chaque binôme. Ainsi, pour *n* binômes contenant des segments de longueur moyenne *l et l'*, on a environ $n \cdot l \cdot l'$ comparaisons (pour le corpus JOC : approximativement 69 120·15·18 soit 18 662 400). En outre, à chaque comparaison, il n'y a pas à confronter deux vecteurs d'occurrences, mais seulement à incrémenter le compte de cooccurrence des deux unités concernées. L'indexation des enregistrements est bornée, à chaque étape, par une constante multiplicative fois $log(n \cdot l \cdot l')$.

Si l'on considère que la longueur moyenne des segments est bornée (ce qui est vrai lorsqu'on a un alignement au niveau des phrases), l'estimation de la complexité en temps et en espace pour obtenir la table de cooccurrence est d'une complexité inférieure ou égale à $O(n\ log(n))$.

Afin de déterminer plus précisément l'évolution de la taille de la table des cooccurrences avec l'accroissement du nombre de binômes, nous avons enregistré les valeurs empiriques de cette évolution, pour l'ensemble du corpus JOC.

Parallèlement, nous avons tenté d'établir un modèle permettant de donner une estimation du nombre de cooccurrences types[207], en fonction du nombre d'occurrences dans les deux textes.

Cette estimation se base sur deux approximations : d'une part, nous supposons que tous les segments de T (respectivement de T') ont la même longueur l (respectivement l') ; d'autre part, nous supposons que les unités lexicales se répartissent en trois groupes de fréquence :

- les hapax (les unités de fréquence 1) ;
- les mots outils ;
- les mots pleins.

On suppose en outre que toutes les unités ont la même fréquence à l'intérieur de chacun des groupes. Cette hypothèse est certes irréaliste (à part pour les hapax), mais elle simplifie considérablement les calculs et permet d'aboutir à une estimation grossière.

Nous avons enregistré les nombres d'hapax des textes T et T' en fonction de N et N', respectivement les nombres d'occurrences dans T et T'. Comme le remarque Melamed (1998a :19) citant Zipf (1936), on observe, quelle que soit la langue, une corrélation linéaire entre le logarithme de la taille d'un corpus et le logarithme du nombre d'hapax.

Ainsi, les valeurs empiriques, pour nos deux textes, peuvent être modélisées assez précisément sous la forme d'une puissance de N (cf. figure 49) :

$$V_{hapax}(N) \approx N^{p} \qquad (100)$$

avec, respectivement, les deux constantes p = 0,91491 et p'= 0,91827

[207] Une cooccurrence type correspondant à la cooccurrence de deux unités type, sans prendre en compte le nombre de fois que les occurrences de ces deux unités cooccurrent.

Remarquons que ces modèles sont empiriques, dans la mesure où ils ne sont pas déduits d'une construction théorique : nous ne sommes pas en mesure d'en fournir une interprétation.

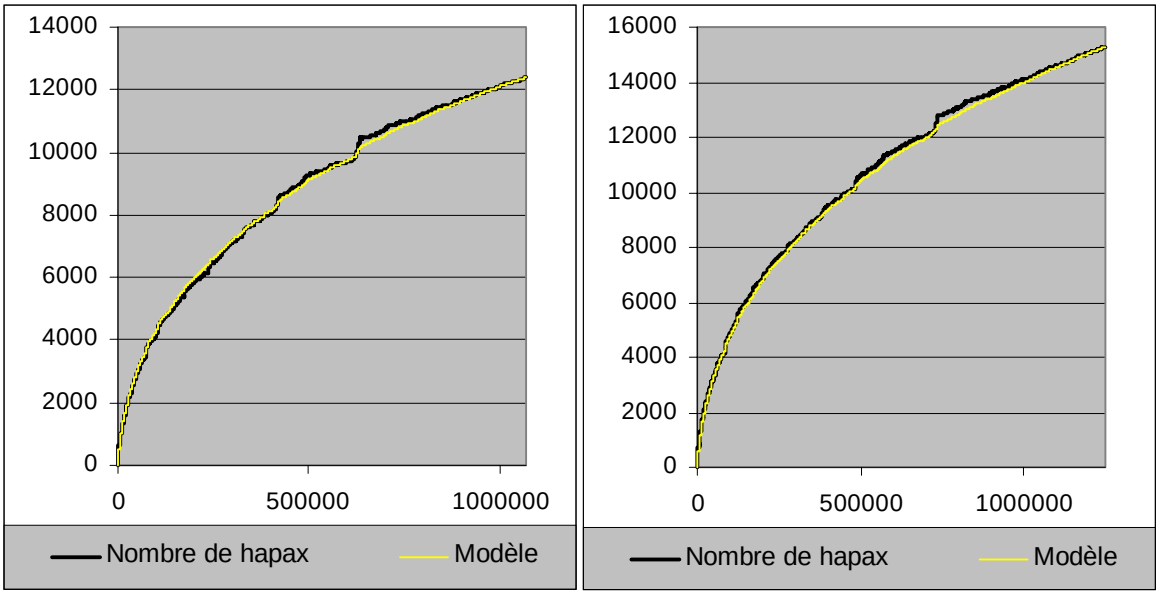


figure 49 : *évolutions empiriques et modèles du nombre d'hapax pour les parties anglaises et françaises du corpus JOC*

D'après nos hypothèses, chaque hapax de T cooccurre avec *l'* occurrences dans T'. Ces occurrences correspondent à un certain nombre d'unités types (ou vocables). Soit $V'(x)$ la fonction donnant une estimation du nombre moyen d'unités types (la taille du vocabulaire) correspondant à x occurrences dans T'. Avec l'aide de $V'(x)$, on peut estimer le nombre de cooccurrences types engendrées par les hapax de T :

$$Cooc_{hapax} = V_{hapax}(N) \cdot V'(l') = N^{p} \cdot V'(l') \qquad (101)$$

En ce qui concerne les mots outils, on fait l'hypothèse que leur nombre est limité à V_{outils} unités et qu'ils constituent une proportion constante t des occurrences. Pour un texte comptant N occurrences, on a donc : $t \cdot N$ occurrences de mots outils, et donc $t \cdot N / V_{outils}$ occurrences par mot outil. Chaque mot outil cooccurre donc avec $(t \cdot N / V_{outil}) \cdot l'$ occurrences

de *T'*, ce qui représente un nombre de cooccurrences types de V'(($t \cdot N/V_{outil}) \cdot l'$). L'ensemble des mots outils de *T* génère $Cooc_{outils}$ cooccurrences types[208] :

$$Cooc_{outils} = V'\left(\frac{t \cdot N}{V_{outils}} \cdot l'\right) \cdot V_{outils} \qquad (102)$$

Pour les mots pleins, enfin, il faut considérer toutes les unités restantes, soient :

$N - t \cdot N - N^p = N \cdot (1-t) - N^p$ occurrences

$V(N \cdot (1-t) - N^p)$ types

et donc :

$\frac{N \cdot (1-t) - N^p}{V(N \cdot (1-t) - N^p)}$ occurrences par type

ce qui représente :

$V'\left(\frac{N \cdot (1-t) - N^p}{V(N \cdot (1-t) - N^p)} \cdot l'\right)$ cooccurrences types

On obtient donc, pour tous les mots pleins, un total de $Cooc_{pleins}$ cooccurrences types :

$$Cooc_{pleins} = V'\left(\frac{N \cdot (1-t) - N^p}{V(N \cdot (1-t) - N^p)} \cdot l'\right) \cdot V(N \cdot (1-t) - N^p) \qquad (103)$$

Le nombre total de cooccurrences types estimé par notre modèle prend donc la forme suivante :

$$cooc = occ^p \cdot V'(l') + V'\left(\frac{t \cdot occ}{V_{outils}} \cdot l'\right) \cdot V_{outils} + V'\left(\frac{occ \cdot (1-t) - occ^p}{V(occ \cdot (1-t) - occ^p)} \cdot l'\right) \cdot V(occ \cdot (1-t) - occ^p) \qquad (104)$$

En ce qui concerne les fonctions $V(x)$ et $V'(x)$, représentant la taille du vocabulaire moyen de deux textes en langue *L* et *L'*, comptant *x* occurrences, nous nous sommes basé

[208] dans ce calcul, on néglige l'éventualité où un mot outil a plusieurs cooccurrences dans une même phrase : ce biais peut néanmoins être compensé par une minoration du taux t_{outils}.

sur les valeurs empiriques concernant le corpus JOC. Il aurait été également possible d'utiliser les modèles de l'accroissement du vocabulaire.

Par exemple, nous avons représenté, figure 50, la courbe empirique de l'évolution du vocabulaire du bi-texte en fonction de son accroissement en nombre d'occurrences (pour chaque nouveau binôme, nous avons pris la moyenne du vocabulaire de T et T', ainsi que la moyenne des occurrences de T et T'). Cette fois, la courbe se modélise assez bien sous la forme d'une racine carrée, à une constante multiplicative près[209] :

$$V(N) = k \cdot \sqrt{N} \qquad (105)$$

Empiriquement, on trouve : k = 31,61

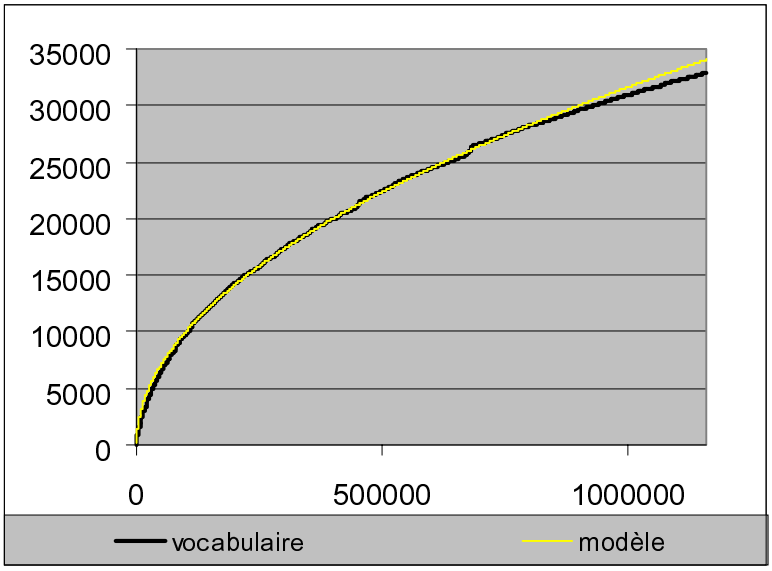


figure 50 : *accroissement du vocabulaire du bi-texte en fonction du nombre d'occurrences* (*données empiriques et modèle*)

Finalement, on peut donc comparer l'accroissement empirique de la table des cooccurrences et le modèle de l'équation (104). Pour les paramètres du modèle, nous avons

[209] On constate cependant que le modèle et la courbe empirique divergent pour les grandes valeurs de *N*. Peut-être une puissance légèrement inférieure à ½ conviendrait-elle mieux, avec un facteur *k* supérieur.

considéré que V_{outils} = 400 mots outils représentaient t = 20 % des occurrences (ce dernier pourcentage est sous-évalué par rapport aux normes de notre corpus, mais ceci permet de compenser le biais signalé à la note 208).

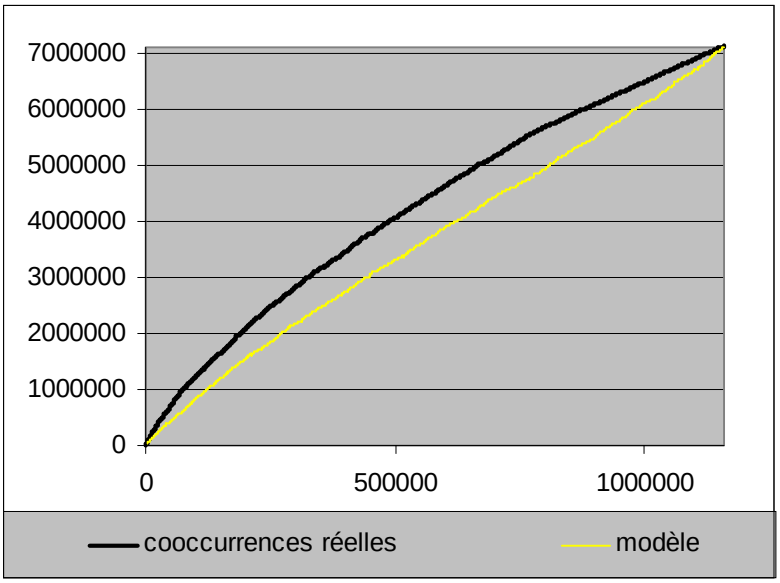


figure 51 : *évolution du nombre de cooccurrences types en fonction du nombre d'occurrences*

Comme on le voit, notre modèle fournit une approximation grossière. En outre, il est à craindre que le modèle diverge des résultats empiriques pour les grandes valeurs de *N*, au-delà de 1 200 000 d'occurrences (la courbe des cooccurrences observées est convexe tandis que celle de notre modèle est concave). Cette évolution asymptotique est cependant intéressante : on peut l'imputer à l'estimation du nombre de cooccurrences pour chaque unité du texte T . Cette estimation se base sur la fonction $V'(x)$, qui donne la taille moyenne du vocabulaire d'une portion de texte : on peut supposer que ce nombre est surestimé lorsqu'il s'agit de réduire les cooccurrences d'une unité à ses cooccurrences types. Ce qui tendrait à démontrer qu'il y a une certaine récurrence du contexte traduit, et qu'une lexie cooccurre toujours plus ou moins avec les mêmes lexies dans la portion alignée.

Enfin, la forme générale de la courbe empirique semble confirmer que l'accroissement des cooccurrences types est légèrement plus lent qu'une fonction linéaire

de la taille du corpus. Le nombre total de cooccurrences types observées, pour la totalité du corpus JOC est de 7 120 212.

III.3.5.2 Extraction des correspondances

Pour deux segments de longueurs respectives *l* et *l'*, l'algorithme d'extraction effectue *l·l'* comparaisons, et indexe chaque couple candidat grâce à un arbre binaire. La complexité globale de la construction de cet index est par conséquent bornée par une constante. Pour *n* binôme, la complexité est en $O(n)$.

Pour un corpus de *n* binômes, la complexité d'une extraction complète de correspondances est donc bornée par $O(n \log(n))$ en temps et $O(n)$ en espace.

III.3.6 Paramètres décisifs

Il peut être intéressant de repérer, au niveau du corpus que nous avons traité et des unités qui sont appariées, les paramètres qui conditionnent les résultats obtenus lors de l'extraction des correspondances lexicales.

III.3.6.1 Taille du corpus d'apprentissage

La taille du corpus d'apprentissage est sans doute le paramètre principal dont dépend le succès des indices statistiques précédemment étudiés. En effet, on peut supposer que plus le corpus est vaste, plus les combinaisons de cooccurrences syntagmatiques sont variées, ce qui permet de faire mieux ressortir les régularités traductionnelles en réduisant l'effet des cooccurrences indirectes.

Nous avons voulu déterminer empiriquement l'effet de la taille du corpus d'apprentissage sur les résultats. Nous avons donc constitué des corpus d'apprentissage de tailles variables, par tirage aléatoire des couples alignés du corpus complet.

– *Corpus d'apprentissage incluant le corpus d'évaluation*

Dans une première série d'observations, les tirages ont été faits de telle sorte que tous les binômes du corpus de référence soient inclus dans les différents corpus d'apprentissage.

1 corpus de référence	***2***	***3***	***4***	***5***	***6***	***7***	***8 corpus entier***
767	1 990	4 502	8 181	15 528	30 238	49 780	69 160

tableau 63 : *tailles des corpus d'apprentissage incluant le corpus d'évaluation*

Pour chaque corpus, on a effectué de nouvelles extractions (LEX avec ABIJ) avec les 5 indices qui dépendent des distributions lexicales. Les résultats de *P* et *R* suivant des évolutions parallèles, nous n'avons représenté que l'évolution de *F* (cf. figure 52, les résultats numériques sont donnés du tableau 107 au tableau 109 de l'annexe) :

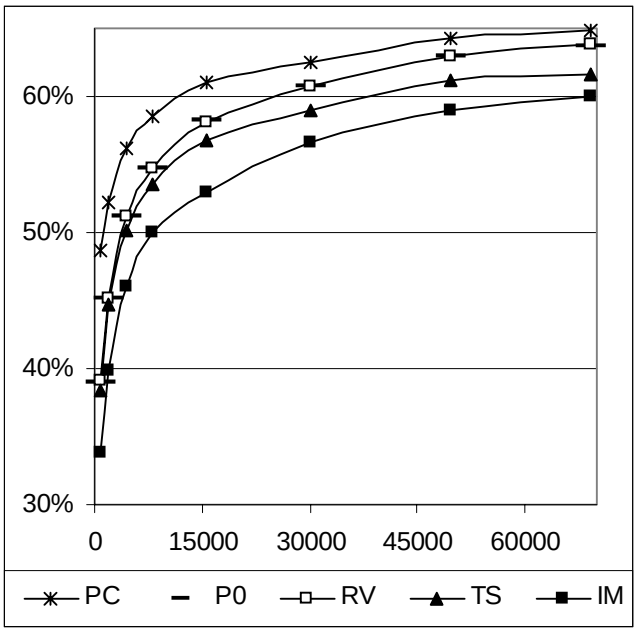


figure 52 : *évolution de la F-mesure en fonction de la taille du corpus d'apprentissage*

On constate pour chaque indice une progression très importante : *F* passe pratiquement du simple au double.

Là encore, les courbes de P0 et RV sont confondues. La courbe de IM effectue une évolution parallèle à ces deux indices, en restant toujours environ 5 points en deçà. Les résultats de TS, en revanche, s'accroissent à un rythme différent : initialement équivalents à ceux de P0 et RV, ils se rapprochent de ceux de IM de manière asymptotique.

Quant à l'indice PC, il bénéficie clairement de la cognation comme source d'information complémentaire : plus l'information distributionnelle fait défaut, plus le recours aux cognats est profitable. Avec le corpus complet, l'écart entre PC et P0 est de seulement 1 point, tandis qu'avec un corpus réduit au corpus d'évaluation, PC domine les autres indices de 10 points.

Ainsi, l'évolution des courbes laisse présager une réduction de l'écart entre PC et P0 avec l'accroissement du corpus : au-delà de 80 000 binômes, l'avantage tiré du recours à la cognation deviendrait vraisemblablement insignifiant.

On remarque enfin que la progression des indices, initialement très rapide, s'essouffle au-delà de 30 000 binômes : entre 767 et 30 238, avec P0, *F* progresse de 21,7 % ; entre 30 238 et 69 160, la progression n'est plus que de 3 %. Vu la quantité de calcul requise par ce genre d'extraction, on peut se demander si la prise en compte des 39 000 derniers binômes en vaut la peine. D'autant plus qu'avec la cognation, on se situe alors à seulement 1,3 % des meilleurs résultats de RV.

On voit que le recours au cognat permet une certaine économie de calcul : à partir de 15 528 binômes, PC obtient de meilleurs résultats que RV avec un corpus deux fois plus grand (30 238 binômes).

– *Corpus d'apprentissage sans le corpus d'évaluation*

Dans la pratique, on peut avoir besoin de lancer l'extraction des correspondances sur de nouveaux textes en faisant l'économie d'un nouveau processus d'apprentissage (i.e. sans mettre à jour la table des cooccurrences). Ceci peut être envisageable, par exemple, quand les nouveaux textes sont homogènes à ceux sur lesquels on a préalablement compté les cooccurrences.

Ainsi, il peut être intéressant de renouveler l'étude précédente, à partir d'un corpus d'apprentissage de taille variable, sans inclure les 767 binômes constituant le corpus

d'évaluation. Dès lors, l'extraction des couples du corpus d'évaluation est effectuée à partir de données extérieures à ce corpus.

Le rappel et la précision montrant cette fois des évolutions différentes, nous les avons représentés séparément, figure 53. A titre de comparaison, nous avons également représenté les résultats avec le corpus d'apprentissage complet, comptant les 69 160 binômes (ces résultats ne sont pas reliés aux autres par un trait plein). Les tailles des corpus d'apprentissage s'échelonnent entre 629 et 68 393 (cf. du tableau 110 au tableau 112 de l'annexe).

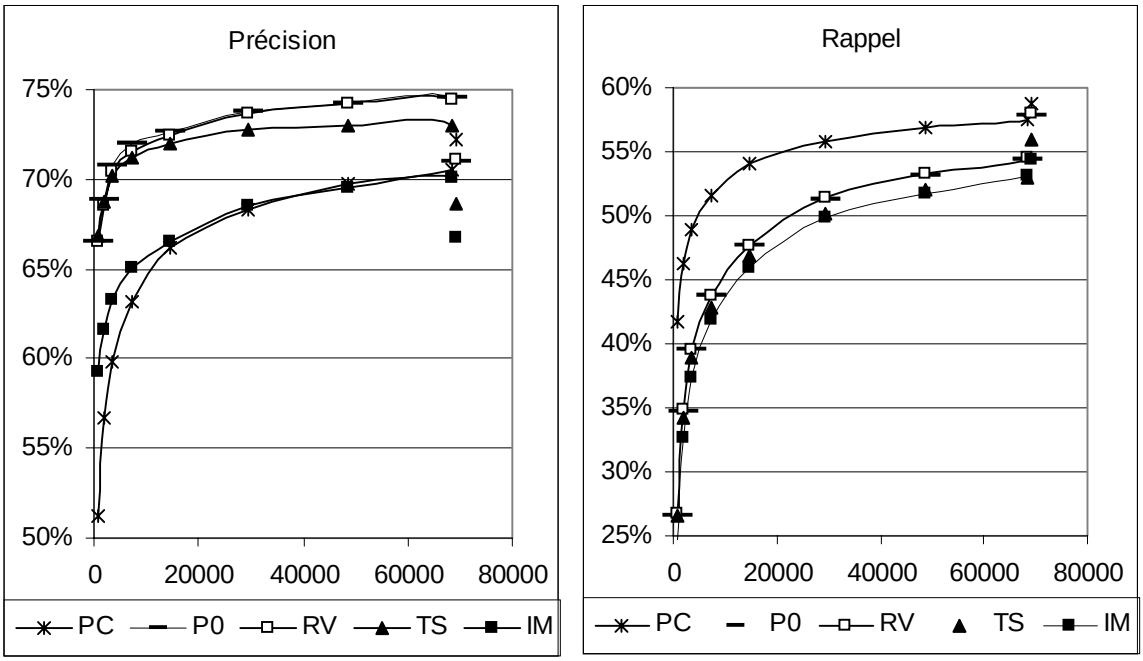


figure 53 : *évolution des résultats en fonction de la taille du corpus d'apprentissage* (*corpus d'évaluation exclu*)

Comparés à ceux des extractions précédentes, les résultats ainsi obtenus laissent apparaître des phénomènes nouveaux. D'une part, on constate que les valeurs de précision sont en général supérieures à celles obtenues précédemment : même avec seulement 629 binômes d'apprentissage, on obtient des précisions de plus de 66 % avec P0 et RV. Seul PC connaît une dégradation de la précision de ses résultats, car pour un certain nombre de couples n'apparaissant pas dans le corpus d'apprentissage, la cognation est la seule source d'information, et CO apporte beaucoup de bruit.

On constate que la précision dépasse même les valeurs atteintes avec le corpus d'apprentissage complet, avec un maximum de 74,5 % pour RV et P0.

Du côté du rappel, on assiste à une légère diminution par rapport aux résultats de la figure 52. Le rappel de PC se conserve mieux que celui des autres indices, qui obtiennent des valeurs de rappel très rapprochées.

Nous pensons que ces phénomènes, augmentation de la précision et diminution du rappel, sont tous deux imputables à l'élimination des hapax du corpus de référence, c'est-à-dire aux unités qui n'y apparaissent qu'une fois. Cela tendrait à confirmer que les hapax sont sources de bruit et engendreraient des associations erronées susceptibles d'amoindrir la précision.

Afin de confirmer cette hypothèse, il faut maintenant mener une étude des résultats en fonction de la fréquence des unités comparées. De manière générale, il peut être intéressant de regrouper les unités lexicales afin d'évaluer les résultats pour des classes plus homogènes. Nous avons opéré de tels regroupements pour les lemmes correspondant aux lexies du corpus de référence. Deux types de classement ont retenu notre attention : les classes de fréquence, et les classes morphosyntaxiques.

III.3.6.2 Classes de fréquence

Comme indiqué au tableau 64, nous avons réparti les lemmes de chaque langue en 10 tranches de fréquences d'égale importance, la tranche 1 représentant les hapax :

Tranches	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
fréquences	> 0 ≤ 1	> 1 ≤ 3	> 3 ≤ 8	> 8 ≤ 18	> 18 ≤ 33	> 33 ≤ 59	> 59 ≤ 101	> 101 ≤ 175	> 175 ≤ 400	> 400 < 5000
nombre de lemmes anglais	310	291	355	331	338	336	335	342	348	343
nombre de lemmes français	347	345	392	427	371	399	330	338	361	326

tableau 64 : *répartition des lemmes en dix tranches de fréquence*

A l'intérieur de chaque classe de fréquence, le calcul de la précision et du rappel est effectué séparément entre les deux langues. Au cours de l'évaluation, on enregistre trois comptes pour chaque lemme :

- n_{ref} : le nombre de fois que le lemme apparaît dans les couples de référence.
- $n_{corrects}$: le nombre de fois que le lemme est correctement apparié dans les couples évalués.
- $n_{incorrects}$: le nombre de fois que le lemme est apparié de façon incorrecte dans les couples évalués.

Pour chaque classe *C*, dans chaque langue, on a donc :

$$P(C) = \frac{\sum_{lemme \in C} n_{corrects}}{\sum_{lemme \in C} (n_{corrects} + n_{incorrects})} \qquad R(C) = \frac{\sum_{lemme \in C} n_{corrects}}{\sum_{lemme \in C} n_{ref}} \tag{106}$$

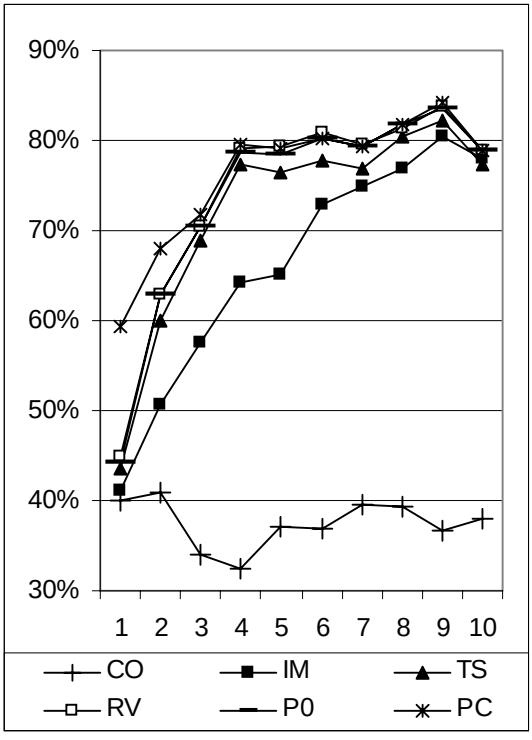


figure 54
évolution de F en fonction des tranches de fréquence en anglais (*LEM avec ABIJ*)

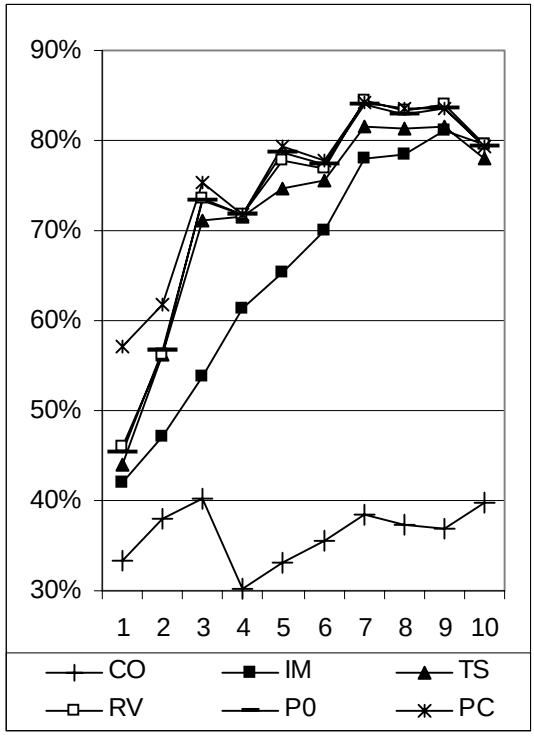


figure 55
évolution de F en fonction des tranches de fréquence en français (*LEM avec ABIJ*)

Nous avons représenté, figure 54 et figure 55, les valeurs de F-mesure pour chaque indice en fonction des tranches de fréquence, pour l'anglais et le français (cette fois, *P* et *R* sont très proches en valeur absolue, le rappel étant légèrement supérieur – pour le détail des résultats, cf. du tableau 113 au tableau 118 de l'annexe).

Notons que les simplifications algorithmiques qui affectaient le rappel n'ont ici pratiquement plus d'effet :

- les lemmes comptant plus de 5 000 occurrences sont rejetés hors de nos classes ;
- seule la classe 10 est susceptible de compter des unités fréquemment répétées plusieurs fois dans un même binôme.

Les courbes obtenues nous révèlent des phénomènes intéressants :

- D'une façon générale les indices basés sur les distributions s'améliorent lorsque la fréquence augmente. Sur les premières tranches la croissance est très rapide. Au-delà d'un certain palier, la progression s'essouffle : en anglais, on obtient de bons résultats avec RV et P0 dès la tranche 4 (fréquences supérieures à 8 occurrences), tandis qu'en français il faut attendre la tranche 5 (fréquences supérieures à 18 occurrences). Notons que la courbe des lemmes français est nettement plus irrégulière : on observe un creux au niveau de la tranche 4. Il s'agit sans doute d'un phénomène linguistique, mais nous n'en avons pas trouvé d'interprétation satisfaisante.

- On observe une légère diminution pour la tranche 10 : on peut l'imputer à la représentation importante des mots outils, dont les résultats moyens sont faibles.

- La cognation n'est pas clairement corrélée à la fréquence, et elle donne des résultats relativement stables. On observe toutefois un creux au niveau de la tranche 4, dans les deux langues. Là aussi, il existe peut être une explication linguistique : par exemple, une plus grande représentation de lemmes comportant des suffixes longs communs aux deux langues, comme *–ment* ou *–tion*. Nous n'avons pas vérifié une telle hypothèse.

– A part pour PC, les hapax obtiennent des scores assez faibles, loin derrière les lemmes apparaissant au moins deux fois (il y a environ 15 % d'écart, avec RV). C'est à ce niveau que la cognation présente un réel intérêt, et fournit un bon complément à l'information statistique. Pour les lemmes dépassant 3 occurrences on constate que la cognation n'est plus d'un grand intérêt.

L'amélioration des résultats avec l'augmentation de la fréquence confirme les observations précédentes : les cooccurrences indirectes se dispersent à mesure que les cooccurrences directes (entre unités équivalentes) s'affermissent.

– *Cas particulier des hapax*

A l'inverse, l'appariement des hapax devient aléatoire chaque fois que plusieurs hapax cooccurrent dans une même phrase. Melamed (1998a :20) donne une estimation de la probabilité p_{cooc} d'un tel événement sous la forme d'une loi binomiale :

si l est le nombre d'unités d'un segment source (on suppose tous les segments de même longueur) et p_h la probabilité de tirer un hapax en langue source, on peut estimer la probabilité d'avoir exactement k hapax dans un segment :

$$p(k) = C_l^k {p_h}^k (1 - p_h)^{l-k} \tag{107}$$

d'où :

$$p_{cooc} = 1 - p(0) - p(1) = 1 - (1 - p_h)^l - p_h (1 - p_h)^{l-1} \tag{108}$$

Mais nos observations nous indiquent que ce nombre est largement sous-estimé. Sur le corpus anglais, par exemple, on trouve 14 segments avec 8 hapax, événement dont la probabilité estimée est pratiquement nulle.

k	***Nombre de segments anglais avec k hapax***	
	Nombre observé	*Nombre théorique*
0	60 360	57 752
1	6 786	10 149
2	1 272	832
3	408	42
4	167	1
5	68	0
6	44	0
7	19	0
8	14	0
…	…	…
54	1	0

tableau 65 : *cooccurrences des hapax dans un même segment*

Ces distorsions sont dues en partie à l'hypothèse de constance de la longueur des segments, qui est évidemment une simplification. Mais surtout, il s'agit là d'un phénomène textuel, car les hapax ont tendance à apparaître en groupe, et leur distribution véritable n'a rien d'aléatoire. Ceci s'explique simplement par des phénomènes de changement thématique, ainsi que par la présence d'énumérations, de tableaux de chiffres, etc. La formule de l'équation (108) est en quelque sorte une estimation « *a minima* » de la proportion de segments contenant plus de un hapax. Même si elle est incorrecte en valeur absolue, elle permet de suivre les variations de p_{cooc} en fonction de la taille du texte.

De manière plus précise, on peut estimer la quantité minimale d'appariements incorrects engendrés par les hapax en langue source. Considérons un segment contenant *k* hapax, et supposons que ces hapax doivent être appariés avec *k* unités dans l'autre partie du bi-texte. Un indice basé sur les seules distributions aboutira à un appariement aléatoire avec les hapax. Sur *k!* appariements possibles, on peut calculer la probabilité des cas de figure suivants[210] :

– 2 erreurs : $\frac{1}{k!} \cdot C_j^k \cdot 2!(1-\frac{1}{1!}+\frac{1}{2!})$

[210] D'abord on tire les *j* unités qui vont être appariées de façon erronée : C_k^j possibilités ; puis, pour ces *j* unités, on calcule le nombre de permutations sans point fixe, ou *dérangements* : j! (1-1/1!+1/2!+...+$(-1)^j$/j!)

– 3 erreurs : $\frac{1}{k!}\cdot C_k^3 \cdot 3! \cdot (1-\frac{1}{1!}+\frac{1}{2!}-\frac{1}{3!})$

– …

– j erreurs : $\frac{1}{k!}\cdot C_k^j \cdot j! \cdot \left(1-\frac{1}{1!}+\frac{1}{2!}-...+\frac{(-1)^j}{j!}\right)$

L'espérance du nombre d'erreurs engendrées dans un segment contenant k hapax est donc donnée par l'équation (116) :

$$E(N_{erreurs},k)=\sum_{j=2}^{k}\frac{j}{k!}\cdot C_k^j \cdot j! \cdot \left(1-\frac{1}{1!}+\frac{1}{2!}-...+\frac{(-1)^j}{j!}\right)=\sum_{j=2}^{k}\frac{j}{(k-j)!}\cdot\left(1-\frac{1}{1!}+\frac{1}{2!}-...+\frac{(-1)^j}{j!}\right) \quad (109)$$

En tenant compte des probabilités d'obtenir de tels segments, on a donc l'espérance suivante pour le nombre total d'appariements erronés engendrés par les hapax :

$$\begin{aligned} E(N_{erreurs}) &= \sum_{k=2}^{l} p(k)\cdot\sum_{j=2}^{k}\frac{j}{(k-j)!}\cdot\left(1-\frac{1}{1!}+\frac{1}{2!}-...+\frac{(-1)^j}{j!}\right) \\ &= \sum_{k=2}^{l} C_l^k {p_h}^k (1-p_h)^{l-k}\cdot\sum_{j=2}^{k}\frac{j}{(k-j)!}\cdot\left(1-\frac{1}{1!}+\frac{1}{2!}-...+\frac{(-1)^j}{j!}\right) \end{aligned} \quad (110)$$

Dans ce calcul, nous n'avons tenu compte que des hapax en langue source. Or les hapax en langue cible ne sont pas nécessairement en correspondance avec ces hapax. Sur les couples de référence, on compte qu'environ 48 % des hapax, dans chaque langue, n'est pas apparié avec un hapax. Les erreurs engendrées de part et d'autre se cumulent donc (sauf pour les hapax appariés ensemble).

$$E_{total} = E(N_{erreurs}) + 0{,}48 \cdot E(N'_{erreurs}) \quad (111)$$

Nous avons représenté, l'évolution de E_{total} avec le nombre de binômes du bi-texte : d'une part, figure 56, en nous basant sur les valeurs observées des nombres d'occurrences, de la taille du vocabulaire, du nombre d'hapax ; d'autre part, figure 57, en utilisant les valeurs théoriques fournies par nos modèles d'accroissement du vocabulaire et de la proportion des hapax (cf. figure 49 et figure 50, p. 476).

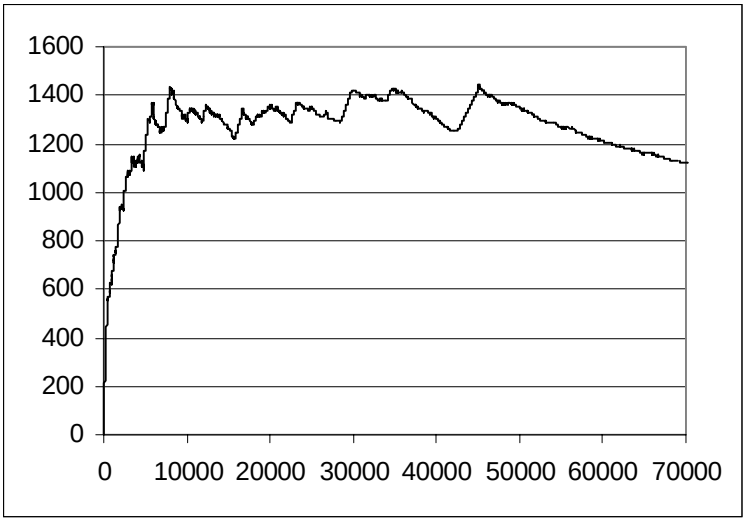


figure 56 : *évolution de E_{total} en fonction du nombre de binômes du bi-texte sur la base des observations de Occ, V*(*Occ*) *et V_{hapax}*(*Occ*)

Après une augmentation rapide, E_{total} semble se stabiliser, puis décroître de manière continue.

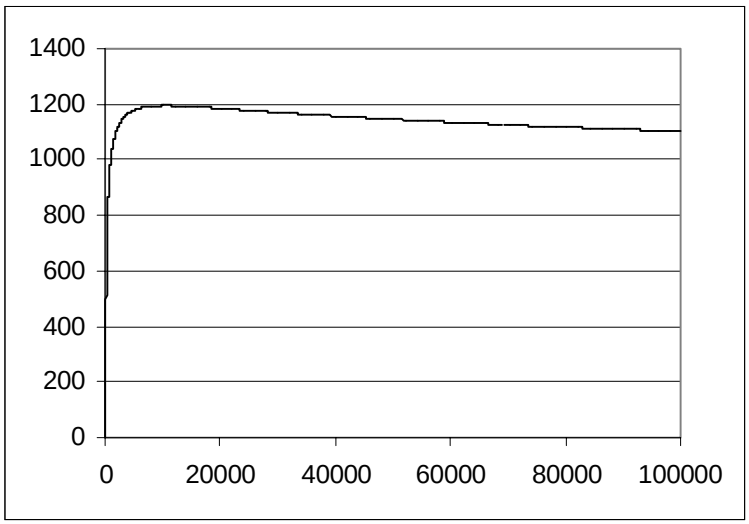


figure 57 : *évolution de E_{total} en fonction du nombre de binômes du bi-texte sur la base des estimations théoriques de Occ, V*(*Occ*) *et V_{hapax}*(*Occ*)

Cette évolution est confirmée par la figure 57, qui donne une image lissée la courbe de E_{total}, où les irrégularités dues à l'accroissement du vocabulaire et du nombre d'hapax sont gommées. Cette courbe théorique fait apparaître plus nettement la forme globale de

l'évolution : on constate que le nombre d'erreurs estimé atteint un maximum (vers 12 500 binômes) puis décroît lentement. Dans une première phase, la densité d'hapax est assez importante pour que chaque nouveau segment soit susceptible de contenir plus d'un hapax. Passé un certain seuil, l'accroissement du nombre d'occurrences étant plus rapide que celui du nombre d'hapax, les hapax s'éparpillent et apparaissent de plus en plus souvent de manière sporadique et isolée : dès lors, le nombre d'erreurs engendrées décroît progressivement.

Ce phénomène confirme que la cognation perd une partie de son intérêt avec l'augmentation de la taille du corpus. Remarquons cependant qu'elle se révélera toujours utile pour une certaine catégorie d'unités : les transfuges numériques et les noms propres (au sens large, incluant toponymes, noms d'organisation, sigles, etc.) qui forment la majeure partie des nouveaux hapax lorsque le corpus est de grande dimension.

III.3.6.3 Classes morphosyntaxiques

Par ailleurs, nous avons effectué des regroupements en fonction des parties du discours. Ce classement ayant été effectué hors du contexte syntaxique, la nature de certaines unités était équivoque. Par exemple, de très nombreux substantifs anglais peuvent se confondre avec des verbes (p. ex. *measure*, *aid*, *use*, *report*, *provision*, *support*, *plan*, etc.) ainsi que quelques verbes français (*p. ex. devoir*, *pouvoir*, *avoir*, *remarque*[211], *contrôle*, etc.). Le même problème se pose pour de nombreux adjectifs, qui peuvent être substantivés (p. ex. angl. *public*, *human*, *standard*, *future*, *Italian*, *local*, *responsible*, *material*, *directive*, fr. *politique*, *parlementaire*, *particulier*, *financier*, *intérieur*, *total*, *moyen*, *producteur*, *scientifique,* etc.). Pour ces cas d'ambiguïté nous avons créé des classes mixtes, tenant compte de l'ambivalence morphosyntaxique. Notons que ces ambiguïtés peuvent être de natures différentes : l'emploi substantivé de *parlementaire* est régulier dans la langue, tandis que la double interprétation de *moyen* relève de l'ambiguïté. Mais notre classement ne prétend pas rendre compte des ses subtilités : nous cherchons

[211] Rappelons que la lemmatisation effectuée est partielle, les formes ambiguës (hors contexte) étant restées telles quelles.

seulement à enregistrer les virtualités morphosyntaxiques les plus communes de chaque unité.

Finalement, nous avons retenu huit classes : substantif, substantif / verbe, nom propre, verbe, adjectif, adjectif / substantif, adverbe, mot outil. Les classes ne concernent que les formes simples : nous n'avons pas tenu compte des unités polylexicales. Par ailleurs, d'autres unités à statut variable n'ont pas été classées : les participes passés, participes présents et gérondifs ; les unités non traitées dépassant 5 000 occurrences, les numériques, les codes alphanumériques (p. ex. 79/409/CEE), et certaines unités ambiguës (p. ex. substantif / conjonction fr. *or*, substantif / verbe / adjectif angl. *draft*, substantif / verbe / adverbe / adjectif angl. *close*, etc.).

Catégorie	***sub.***	***nom propre***	***verbe***	***verbe/ sub.***	***adjectif***	***adjectif/ sub.***	***adverbe***	***mot outil***	***non classé***
Nombre de lemmes anglais	683	193	221	269	223	72	62	157	1 459
Nombre de lemmes français	874	170	335	85	235	112	39	155	1 639

tableau 66 : *répartition des lemmes en huit catégories morphologiques*

classe	***Anglais***	***Français***
substantif	*Council, country, authorities, article, information, area, year, regulation, development, proposal*	*membre, réponse, Conseil, nom, projet, pays, question, cadre, article, action*
nom propre	*OJ, EEC, EC, Greece, Europe, Spain, Italy, PPE, NI, France*	*CE, JO, CEE, Grèce, Europe, Espagne, Italie, PPE, France, Allemagne*
verbe	*give, take, make, provide, consider, adopt, include, ensure, implement, receive*	*donner, prendre, adopter, produire, indiquer, prévoir, concerner, assurer, envisager, estimer*
mixte verbe / sub.	*answer, programme, measure, concern, aid, use, question, report, market, provision*	*pouvoir, mesure, programme, aide, compte, devoir, mise, contrôle, vue, base*

classe	*Anglais*	*Français*
adverbe	*currently, particularly, recently, directly, closely, specifically, especially, notably, approximately, properly*	*notamment, récemment, directement, particulièrement, clairement, pleinement, essentiellement, rapidement, principalement, largement*
adjectif	*national, new, particular, social, specific, environmental, certain, necessary, financial*	*communautaire, honorable, national, relatif, économique, social, nouveau, commun, agricole, spécifique*
mixte adjectif /sub	*European, directive, Greek, public, possible, human, general, standard, future, Italian*	*directive, politique, parlementaire, général, grec, particulier, financier, espagnol, intérieur, total*
mixte mot outil	*it, not, from, at, its, or, what, under, their, can*	*pas, son, si, que, ne, faire, ou, t, leur, plus*

tableau 67 : *unités les plus fréquentes de chaque catégorie*

Rappelons que les unités classées comme mots outils ne représentent pas une catégorie morphologiquement homogène : elles incluent des verbes et des adverbes aussi bien que des conjonctions, des prépositions ou des articles. Nous avons cependant conservé cette classe telle quelle afin de rapprocher les nouveaux résultats de ceux précédemment obtenus. Les classes mixtes sont donc au nombre de trois.

Etant donné leur comportement spécifique vis-à-vis de la traduction, nous avons classé séparément, dans la deuxième colonne du tableau 66, les noms propres, sigles et toponymes.

Les résultats obtenus pour les indices distributionnels sont similaires. Nous donnons, tableau 68, pour les deux langues, les valeurs de précision et rappel du meilleur représentant de ces indices, l'indice RV (pour les autres indices, cf. du tableau 119 au tableau 121 de l'annexe) :

Précision				***Rappel***			
Anglais		*Français*		*Anglais*		*Français*	
Catégorie	*P*	*Catégorie*	*P*	*Catégorie*	*R*	*Catégorie*	*R*
mot outil	46,3 %	mot outil	47,0 %	mot outil	70,3 %	verbe	73,2 %
adverbe	62,2 %	verbe	68,9 %	adverbe	70,9 %	mot outil	79,9 %
verbe	70,7 %	verbe/sub.	76,0 %	verbe	77,1 %	adjectif	80,9 %
verbe/sub.	80,0 %	adjectif	77,5 %	verbe/sub.	84,4 %	verbe/sub.	82,4 %
adjectif	80,1 %	adverbe	79,0 %	adjectif	85,1 %	substantif	84,5 %
substantif	85,0 %	substantif	83,9 %	substantif	86,6 %	adjectif/sub.	86,0 %
nom propre	89,7 %	adjectif/sub.	88,4 %	adjectif/sub.	89,5 %	adverbe	87,5 %
adjectif/sub.	90,4 %	nom propre	91,7 %	nom propre	90,5 %	nom propre	91,3 %

tableau 68 : *résultats triés par ordre croissant de l'indice RV pour les catégories morphosyntaxiques en l'anglais et en français*

Globalement, les résultats liés à une catégorie dépendent principalement de trois facteurs :

- la fréquence moyenne des unités de la catégorie.
- la proportion des occurrences résiduelles des unités de la catégorie.
- la plus ou moins grande stabilité des traductions de ces unités.

Les deux derniers facteurs mettent en jeu un véritable complexe de phénomènes traductionnels : divergences grammaticales, lexicales, choix pragmatiques, etc. On ne peut en donner une interprétation qu'au cas par cas, par une analyse fine des traductions. On peut néanmoins énoncer quelques principes généraux :

- plus une unité est polysémique, et plus ses traductions risquent d'être variables, et donc les résultats mauvais.
- de même, une unité qui n'a pas, dans l'autre langue, d'équivalent lexical satisfaisant sur le plan sémantique, recevra des traductions parfois aléatoires. En outre, dans ce cas, la compositionnalité traductionnelle est moins souvent respectée au niveau lexical.

Mais avant d'interpréter les résultats liés à chaque catégorie, il faut déterminer si ces catégories sont statistiquement indépendantes des tranches de fréquence. Après avoir établi des tables de contingences croisant tranches *i* et catégories *j* (cf. du tableau 122 au tableau

123 de l'annexe) nous avons calculé, en pourcentage, la répartition des catégories sur les dix tranches.

Ces pourcentages sont représentés, catégorie par catégorie, figure 58 pour le corpus anglais et figure 59 pour le corpus français.

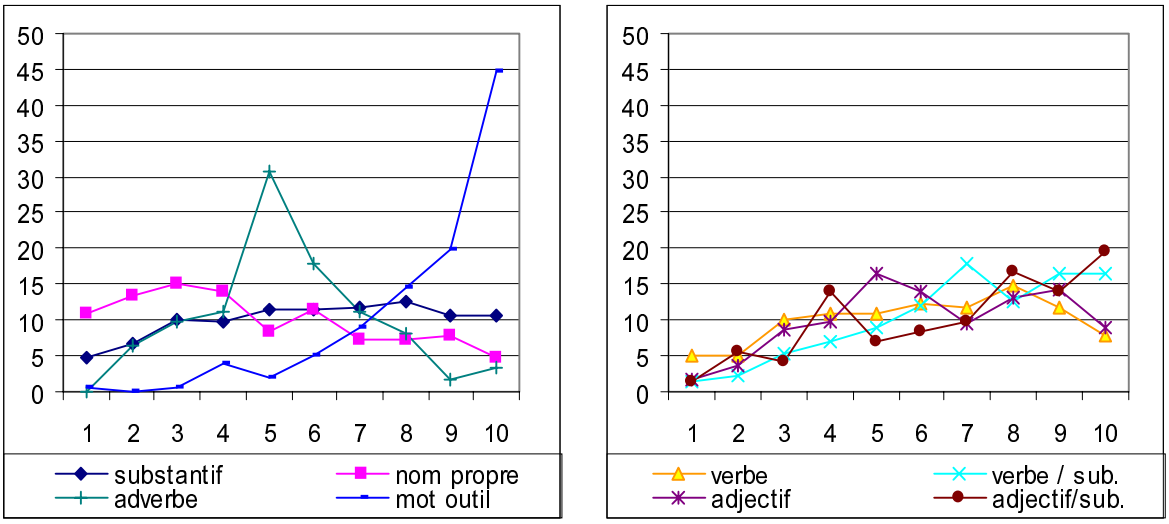


figure 58 : *répartition des catégories anglaises dans les dix tranches fréquence*

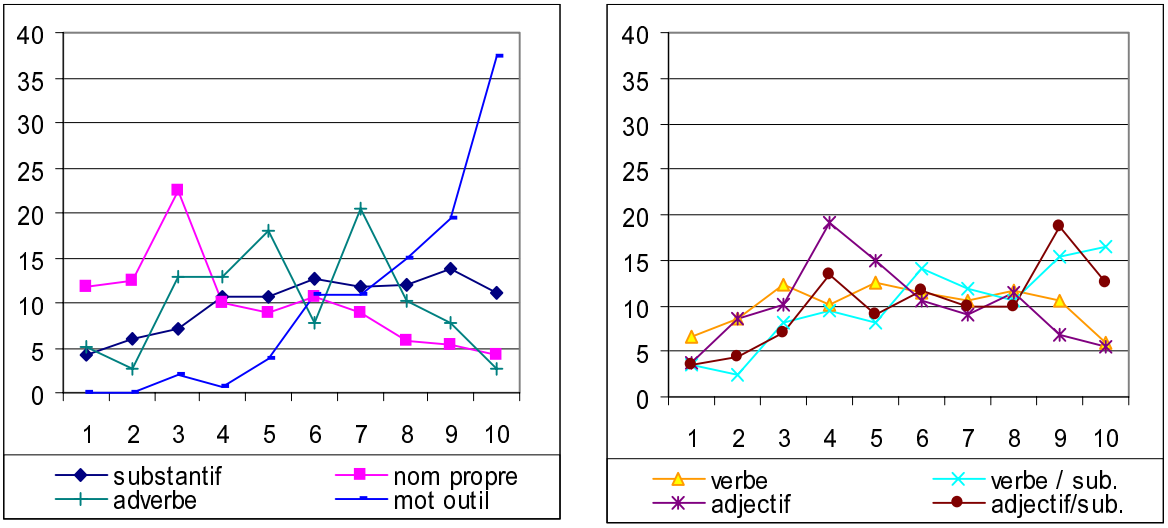


figure 59 : *répartition des catégories françaises dans les dix tranches fréquence*

On constate que la répartition de certaines catégories est assez nettement liée aux tranches de fréquence :

- les noms propres sont mieux représentés dans les basses fréquences ;[212]
- à l'opposé, les mots outils se concentrent dans les tranches 9 et 10 ;
- les catégories mixtes adjectif / substantif et verbe / substantif ont une nette majorité de représentants dans les tranches 6 – 10 ;
- les adverbes se concentrent de façon marquée dans les tranches médianes 5 - 7.

Pour la plupart des catégories, en dehors des mots outils et des noms propres, la répartition est suffisamment homogène autour des valeurs médianes pour qu'on ne puisse soupçonner une incidence forte sur les résultats.

En revanche, si la fréquence moyenne des unités d'une catégorie était le seul facteur susceptible d'affecter les résultats des indices statistiques, il faudrait s'attendre (cf. la figure 54 et la figure 55, p. 483) à ce que les mots outils obtiennent de bons résultats, et les noms propres de mauvais.

Or, en anglais comme en français, on assiste au phénomène inverse : les mots outils atteignent la F-mesure la plus basse (resp. 55,9 % et 59,2 % en anglais et en français) tandis que les noms propres obtiennent la meilleure valeur de *F* (resp. 90,1 % et 91,5 % en anglais et en français).

Il faut donc y voir une tendance marquée liée à la stabilité des traductions : les unités de notre catégorie « nom propre » ont des traductions stables et conformes à l'hypothèse de compositionnalité traductionnelle au niveau lexical, à la différence des mots outils qui présentent les caractéristiques inverses.

[212] Mais cela ne signifie pas que la majorité des noms propres soient des hapax : seulement 10 % des noms propres sont de fréquence 1, ce qui explique que les résultats liés aux noms propres restent bons.

III.3.6.3.1 Contrôle des artefacts

L'élimination des unités de plus de 5 000 occurrences n'affecte pas les résultats du tableau 68, puisque ces unités ne sont pas classées. En revanche, il est possible que l'autre simplification algorithmique, à savoir la prise en compte d'une seule occurrence de chaque unité par binôme, soit à l'origine de certains des écarts présentés au niveau du rappel.

En effet, certaines classes ont peut-être une plus grande tendance à la répétition que d'autres, du moins en ce qui concerne les occurrences appariées. C'est ce que nous avons voulu déterminer, en calculant la proportion des occurrences ayant un équivalent traductionnel traitées par notre algorithme, pour chaque classe morphosyntaxique (cf. tableau 69).

Anglais		***Français***	
Catégorie	*Occurrences traitées*	*Catégorie*	*Occurrences traitées*
adjectif/sub.	95,7 %	adjectif/sub.	94,3 %
substantif	96,1 %	verbe/sub.	95,6 %
mot outil	96,9 %	substantif	96,6 %
nom propre	97,5 %	nom propre	97,0 %
verbe/sub.	97,5 %	mot outil	97,9 %
adjectif	98,1 %	adjectif	98,0 %
verbe	98,1 %	verbe	99,1 %
adverbe	98,8 %	adverbe	100,0 %

tableau 69 : *proportion d'unités traitées pour chaque catégorie*

Même si les variations sont assez faibles, il semble que toutes les catégories n'aient pas le même comportement vis-à-vis de la répétition. Si l'on néglige les classes mixtes, on retrouve en effet le même classement en anglais et en français :

substantif < nom propre < adjectif < verbe < adverbe

La plus grande récurrence des substantifs est peut-être liée à des raisons thématiques. En ce qui concerne les adverbes (ceux qui n'ont pas été classés parmi les mots outils), ils semblent peu enclins à la répétition, sans doute pour éviter les pesanteurs sur le plan du style (la plupart sont des mots longs suffixés par *–ly* en anglais et *–ment* en français).

De toutes façons, ces variations sont trop faibles pour expliquer les différences de rappel entre les catégories (cf. tableau 68). Nous négligerons donc l'effet de cet artefact dans l'interprétation des résultats.

III.3.6.3.2 Occurrences avec correspondance

Nous avons ensuite cherché à déterminer la proportion des occurrences avec correspondance (i.e. les occurrences non-résiduelles) à l'intérieur de chaque catégorie. Dans la mesure où ces proportions donnent une estimation de la précision optimale théorique d'une extraction complète, nous avons également indiqué, dans le tableau 70, la différence avec la précision obtenue précédemment (indice RV, algorithme ABIJ).

Cette différence permet de jauger la marge de progression des méthodes d'appariement, catégorie par catégorie.

Anglais			***Français***		
Catégorie	*Occurrences appariées*	*Différence avec P*	*Catégorie*	*Occurrences appariées*	*Différence avec P*
mot outil	59,2 %	12,8 %	mot outil	48,8 %	1,8 %
adverbe	84,0 %	21,8 %	verbe	82,3 %	13,3 %
verbe	85,7 %	15,1 %	verbe/sub.	82,6 %	6,7 %
verbe/sub.	88,2 %	8,2 %	adjectif	85,4 %	7,8 %
adjectif	89,1 %	9,0 %	substantif	85,5 %	1,6 %
substantif	91,0 %	6,0 %	adverbe	87,5 %	8,5 %
adjectif/sub.	93,2 %	2,9 %	adjectif/sub.	90,1 %	1,7 %
nom propre	95,6 %	5,9 %	nom propre	96,3 %	4,7 %

tableau 70 : *proportion d'unités avec correspondance pour chaque catégorie et différence avec la précision obtenue avec RV*

On constate que la précision des résultats est étroitement corrélée à la proportion d'unités avec correspondance, puisque le classement des catégories est pratiquement identique à celui du tableau 68.

Cette fois il apparaît très clairement que les mauvais résultats des mots outils sont dus en grande partie à une forte proportion d'occurrences résiduelles. Cette hypothèse est

confirmée par le fait que les mots outils obtiennent un rappel bien supérieur à leur précision : les correspondances 1-0 ou 0-1 de ces occurrences résiduelles sont sources d'appariements erronés qui affectent la précision sans compromettre le rappel.

Par ailleurs, dans les deux langues, ce sont les verbes et les adverbes qui présentent l'écart le plus important entre la précision obtenue et la précision optimale. On ne peut expliquer ces difficultés par la fréquence moyenne de ces catégories : la cause est sans doute à chercher du côté d'une certaine variabilité traductionnelle, découlant du contenu sémantique assez vague ou de la polysémie des unités en question.

III.3.6.3.3 Hiérarchie globale issue du rappel

Ainsi, en se basant sur le rappel, on peut donner un classement reflétant d'une façon générale la variabilité traductionnelle des catégories. Par ordre croissant de rappel on observe, dans les deux langues, une hiérarchie similaire :

(*mot outil*, *verbe*) < (*verbe* / *sub*., *adjectif*) < *substantif* < *adjectif* / *sub*. < *nom propre*[213]

Etrangement, la catégorie mixte adjectif / substantif dépasse la classe des substantifs, alors que les adjectifs présentent des résultats légèrement inférieurs. Un examen minutieux nous révèle que cette catégorie mixte contient un grand nombre d'ethnonymes : *European*, *Greek* , *espagnol*, *balte* ou de termes renvoyant à des notions précises dans des textes à vocation juridique, comme les noms de métiers, les statuts juridiques, les noms d'appartenance à un groupe, etc. : angl. *scientific*, *resident*, *catholic*, *public*, *fr*. *douanier*, *informatique*, *producteur*, *victime*, *militaire*, *combustible*, etc. La plupart de ces termes, bien que grammaticalement ambigus, ont un sens précis dans le domaine considéré et donc des traductions relativement stables.

Doit-on supposer que l'emploi substantivé de l'adjectif (ou adjectival du substantif) implique plus fréquemment des unités dont le contenu sémantique s'approche de la monosémie ? les données dont nous disposons sont bien sûr insuffisantes pour généraliser

[213] Les classes entre parenthèses ne sont pas dans le même ordre relatif dans les deux langues.

cette observation, mais il est possible que ce genre de phénomènes statistiques se dessine avec plus de netteté sur un corpus plus important.

A notre niveau, nous nous contenterons de conclure que les verbes aboutissent en général à des traductions plus variables que les substantifs et les adjectifs, ce qui dénote une polysémie plus importante de cette catégorie. A l'opposé, les noms propres, par définition monosémiques, se caractérisent par des correspondances récurrentes, plus faciles à extraire automatiquement.

On peut supposer que les usages terminologiques, dans la mesure où ils tendent eux aussi à la monosémie, obtiendraient aussi de bons résultats dans ce type d'extraction (mais dans la mesure où nous nous ne disposions pas de liste normative des termes de notre corpus, nous n'avons pas étudié une telle catégorie).

Nous n'avons pas donné de classement pour la classe des adverbes, dont les résultats sont disparates dans les deux langues : une F-mesure de 66,3 % en anglais contre 83,1 % en français. Il est difficile d'expliquer cet écart, quand on sait que les adverbes sont pratiquement insensibles aux artefacts évoqués précédemment. On constate que la proportion d'adverbes résiduels est plus élevée en anglais qu'en français, ce qui pourrait indiquer un manque de consistance sémantique chez certains adverbes anglais. Il existe peut-être une autre raison à ce phénomène : le changement catégoriel, susceptible d'augmenter la variabilité des traductions.

III.3.6.3.4 Changement catégoriel

Comme nous l'avons montré dans le chapitre I, il n'est pas rare que la traduction ne conserve par les catégories morphosyntaxiques des unités impliquées. Or, vis-à-vis des résultats précédents, il peut être intéressant de déterminer le degré de stabilité, ou d'instabilité, de telle ou telle classe morphosyntaxique : il est possible qu'il y ait une corrélation entre l'instabilité catégorielle et les résultats.

En outre, la mise en évidence de la conservation, ou du passage d'une catégorie à une autre, peut parfois être instructive vis-à-vis des préférences idiomatiques de chaque langue,

et caractériser les phénomènes de « stylistique comparée » tels que ceux évoqués par Vinay & Darbelnet.

Nous avons ainsi recensé, pour tous les couples de références, les catégories impliquées (parmi les huit que nous avons retenues). Les nombres d'appariements types pour chaque couple de catégories sont donnés dans le tableau 71.

anglais / ***français***	***sub.***	***nom propre***	***verbe / sub.***	***verbe***	***adverbe***	***adjectif***	***adjectif / sub.***	***mot outil***	***non classé***
sub.	1688	7	454	35	1	18	13	11	20
nom propre	1	276	2	1			6		
verbe / sub.	36		173	3		1	4	46	1
verbe	18		141	408		5		14	10
adverbe	0				32	10		5	11
adjectif	47	6	14	1	5	269	18	19	280
adjectif / sub.	51	8	3	1	6	68	172	1	71
mot outil	3	2	3	4	9	10	3	2595	11
non-classé	85	40	58	82	26	45	4	289	1987

tableau 71 : *correspondances entre catégories – les zones en gris dénotent l'identité* (*gris foncé* : *identité certaine*, *gris clair* : *identité probable*)

On constate que la tendance générale est à la conservation : ce n'est pas étonnant pour un couple de langues aussi proches que le français et l'anglais. Cependant, d'autres combinaisons sont relativement fréquentes. Nous avons relevé, dans les deux sens, les changements catégoriels concernant plus de 2 % des unités de la catégorie.

– *De l'anglais vers le français* :

  – 8 % des adverbes anglais sont appariés avec un adjectif / substantif ;

    Exemple : segment n° 67 643
    angl. : (…) *to create a new framework to facilitate, both legally and* ***financially****, the distribution* (…)
    fr. : (…) *créer un nouveau cadre visant à faciliter, sur les plans législatif et* ***financier****, la circulation* (…)

  – 7 % des verbes anglais sont appariés avec un substantif ;

    Exemples : segments n° 5 985 et 2 120
    angl. : *Thus the United States* ***applies*** *the reduced rate* (…)
    fr. : *Les États-Unis d'Amérique accordent ainsi directement l'****application*** *du taux réduit* (…)

    angl. : (…) *to prevent the Athens-Delphi road being* ***widened***
    fr. : (…) *afin qu'il ne soit pas procédé à l'****élargissement*** *de la route Athènes-Delphes*

  – 6 % des adverbes anglais sont appariés avec un adjectif ;

    Exemple : segments n°31 517
    angl. : (…) *thus excluding the only properly* ***democratically*** *elected institutions*.
    fr. : (…) *d'où est donc exclue la seule institution issue d'élections* ***démocratiques*** *appropriées*.

  – 4 % des adjectifs anglais sont appariés avec un substantif ;

    Exemple : segments n° 327
    angl. : (…) *assurances that the Bishops would be in no danger and* ***free*** *to move about* (…)
    fr. : (…) *l'assurance que les évêques ne seraient pas en danger, qu'ils bénéficieraient de la* ***liberté*** *de mouvement* (…)

  – 3 % des adjectifs / substantifs anglais sont appariés avec un nom propre ;

    Exemple : segments n° 36 791
    angl. : ***Danish*** *enterprise zones*
    fr. : *Zones d'activités économiques au* ***Danemark***

Dans ce cas de figure, il s'agit essentiellement de passage entre nom de pays et nationalité.

– 2 % des substantifs français sont appariés avec un adjectif ;

Exemple : segments n° 8 072
angl. : (…) *the loans in question should be guaranteed under the* ***Community*** *budget*.
fr. : (…) *la garantie des prêts en question par le budget* ***communautaire***.

– 2 % des verbes / substantifs anglais sont appariés avec un adjectif ;

Exemple : segments n° 27 248
angl. : (…) *the US* ***trade*** *law* (…)
fr. : (…) *la loi* ***commerciale*** *américaine* (…)

Dans les deux derniers cas, angl. *Community* et *trade* jouent des rôles de modifieurs de substantif, fonction fréquente chez les substantifs en anglais.

– *Du français vers l'anglais* :

– 17 % des adverbes français sont appariés avec un adjectif ;

Exemple : segments n°12 733
angl. : *Does the Commission intend to provide* ***economic*** *assistance for those farmers* (…)
fr. : *La Commission a-t-elle l'intention d'adopter des mesures destinées à venir* ***financièrement*** *en aide aux agriculteurs* (…)

– 10 % des adjectifs français sont appariés avec un substantif (dont 3 % avec une forme ambiguë verbe / substantif) ;

Exemple : segments n° 62 734
angl. : (…)*Croatian* ***health*** *certificates*.
fr. : (…)*certificats* ***sanitaires*** *croates*.

Fait intéressant : la quasi-totalité des adjectifs concernés sont des adjectifs relationnels : *alimentaire*, *artisanal*, *auditif*, *budgétaire*, *céréalier*, *climatique*, *communautaire*, *législatif*, *maritime*, *sanitaire*, *tarifaire*, *touristique*, *écologique*, *minoritaire*, etc. Comme précédemment, les substantifs anglais correspondants ont une fonction de modifieur.

– 2 % des adjectifs / substantifs anglais sont appariés avec un adverbe ;

Exemple : segment n° 67 643 (ci dessus).

– 2 % des adjectifs / substantifs anglais sont appariés avec un nom propre ;

Exemple : segments n° 36 486
angl. : (…) *imports of such products from* ***Europe***.
fr. : (…) *importations* ***européennes*** *de ce type de produit*.

N.B. : rappelons que ces statistiques ne concernent que les unités classées, qui sont toutes des formes simples.

Ces exemples montrent à quel point il est difficile de catégoriser des unités hors de leur contexte syntaxique, une grande part des lexies pouvant assumer des fonctions différentes. Ce genre d'étude serait sans doute beaucoup plus concluant, du point de vue de la stylistique comparée, si l'on se basait sur une analyse syntaxique des unités afin d'étudier les véritables transformations structurelles mises en jeu. Mais notre point de vue étant essentiellement lexical, ce type d'analyse dépasse le cadre de notre étude.

Par ailleurs, ce que nous enseignent les différents cas de changement catégoriel énumérés ci-dessus, c'est que certaines catégories sont plus instables que d'autres. En totalisant, pour chaque catégorie, le pourcentage minimum d'appariements impliquant un changement (en laissant de côté les unités non-classées, les mots outils et les cas d'ambiguïté), on aboutit au classement du tableau 72 :

Anglais		***Français***	
adverbe	17,2 %	adverbe	15 %
adjectif	11,1 %	adjectif	8 %
verbe	3,9 %	verbe	7 %
nom propre	3,5 %	nom propre	4 %
substantif	2,7 %	substantif	4 %

tableau 72 : *proportion de couples impliquant un changement catégoriel*

Ces chiffres sont sans doute sous-estimés, puisque nous n'avons tenu compte que des cas de changements avérés. Mais ils permettent de dessiner des tendances globales au niveau de chaque catégorie. Or, l'instabilité catégorielle peut être à l'origine d'un surcroît de variabilité traductionnelle, une même lexie pouvant être traduite par différentes unités

de la même famille dérivationnelle, portant toutes le même sens : par exemple, angl. *probably* peut donner *probablement, il est probable que, avec une certaine probabilité, il n'est pas improbable que,* etc. En tout cas, on observe une certaine corrélation entre le classement du tableau 72 et les résultats globaux obtenus avec RV.

En particulier, il semblerait que ce phénomène touche plus fortement la classe des adverbes, fréquemment traduits par des adjectifs ou des substantifs. En outre, en anglais, une forte proportion d'adverbes (30 % contre 12 % en français) correspondent à des unités polylexicales françaises qui connaissent une forte variabilité dans ce type de corpus :

> *à l'évidence, de toute évidence*
> *dans une grande mesure, dans une très grande mesure, dans une forte mesure*
> *pour une grande part, pour une large part, pour une part importante*, etc.

Ces deux raisons conjuguées expliquent vraisemblablement les résultats plus faibles de la classe des adverbes en anglais.

Ainsi, une lemmatisation plus approfondie, regroupant les variantes des expressions polylexicales ainsi que les unités dérivées les unes des autres, quelle que soit leur classe morphosyntaxique, permettrait d'éliminer ces variations superficielles, et peut-être d'améliorer sensiblement les résultats.

III.3.6.3.5 Cognation et catégories : résultats de l'indice CO

Il est vraisemblable que les différentes catégories se comportent différemment vis-à-vis d'un autre type d'indice : la cognation. En effet, si celle-ci n'est que secondairement dépendante des fréquences et des distributions, elle a directement partie liée avec la morphologie, puisqu'elle se base sur les ressemblances de surface.

Afin d'étudier ce type de relation, nous avons consigné les résultats de l'indice CO dans le tableau 73 :

Précision				***Rappel***			
Anglais		*Français*		*Anglais*		*Français*	
Catégorie	*P*	*Catégorie*	*P*	*Catégorie*	*R*	*Catégorie*	*R*
mot outil	5,1 %	adverbe	3,7 %	adverbe	6,3 %	adverbe	3,6 %
adverbe	5,8 %	mot outil	5,5 %	mot outil	7,7 %	mot outil	9,5 %
verbe	24,2 %	verbe	23,3 %	verbe	26,5 %	verbe	25,6 %
verbe/sub.	38,6 %	adjectif	42,2 %	verbe/sub.	40,8 %	adjectif	44,3 %
adjectif	47,0 %	verbe/sub.	43,8 %	adjectif	50,1 %	verbe/sub.	46,6 %
substantif	51,2 %	substantif	46,7 %	substantif	51,9 %	substantif	48,8 %
adjectif/sub.	62,2 %	adjectif/sub.	59,2 %	nom propre	62,7 %	adjectif/sub.	57,8 %
nom propre	62,4 %	nom propre	73,7 %	adjectif/sub.	62,7 %	nom propre	68,9 %

tableau 73 : *résultats de l'indice CO par catégorie morphosyntaxique pour l'anglais et le français avec l'indice RV*

Il est normal que les résultats soient très similaires dans les deux langues, puisque la plupart des couples de cognats appartiennent aux mêmes catégories. Le classement global des résultats donne :

adverbe < mot outil < verbe < adjectif < substantif < nom propre

Bizarrement, on retrouve pratiquement le même ordre qu'avec un indice distributionnel. Mais il est difficile d'en donner une interprétation. Certains phénomènes sont sans doute de nature morphologique : notamment, il y a une confusion entre le suffixe *–ment* de nombreux adverbes français et celui de substantifs anglais assez fréquents (comme *employment*, *improvement*, *amendment*, etc.). Ceci peut expliquer les mauvais résultats des adverbes français, et, par voie de conséquence, anglais.

Bien évidemment, en dehors de ce genre d'interférence superficielle, les résultats doivent être interprétés sur un plan historique, relatif à la génétique des deux langues. Par exemple, un grand nombre de substantifs anglais des domaines concernés (politique, juridique, économique) sont d'origine commune (latin, grec, ancien français).

Dans une certaine mesure, il est possible que le classement précédent soit lié à des aspects sémantiques : on peut supposer que les cognats se partagent des concepts communs standardisés, plus près de la mono-référentialité et moins équivoques que certaines unités « endémiques » dont le sens s'enracine dans un terreau culturel spécifique. En outre,

l'existence de termes apparentés a peut être tendance à stabiliser les traductions. La conjugaison de ces deux facteurs permettrait alors d'expliquer la convergence des résultats entre indice distributionnel et indice superficiel.

III.3.6.3.6 Résultats finaux : l'indice PC

On peut maintenant établir une dernière comparaison entre les différentes classes, à partir des résultats finaux obtenus avec l'indice mixte PC. Le tableau 74 donne la synthèse des difficultés respectives de chaque catégorie vis-à-vis de l'extraction de correspondances.

Anglais		***Français***	
Catégorie	***F***	***Catégorie***	***F***
mot outil	55,9 %	mot outil	58,9 %
adverbe	66,3 %	verbe	71,6 %
verbe	74,2 %	verbe/sub.	79,2 %
substantif	86,9 %	adjectif	81,0 %
verbe/sub.	82,5 %	adverbe	83,1 %
adjectif	83,5 %	substantif	86,6 %
adjectif/sub.	90,7 %	adjectif/sub.	87,4 %
nom propre	93,8 %	nom propre	96,0 %

tableau 74 : résultats de l'indice mixte PC par catégorie morphosyntaxique

III.3.7 Constitution de « dictionnaires »

Dans l'évaluation précédente, nous avons remarqué que l'exploitation des ressemblances superficielles autorisait un certain bruit dans l'identification des cognats. L'information apportée par ces ressemblances étant rare, il s'est avéré plus intéressant de réduire le silence que le bruit.

Nous proposons maintenant d'examiner une source d'information comparable, mais moins rare : la présence, dans un dictionnaire bilingue, du couple de lexies comparées. Malheureusement, dans le cadre de cette évaluation, nous ne disposons pas d'un tel dictionnaire validé par l'humain.

Nous nous sommes donc contenté de faire une simulation, en nous basant sur un dictionnaire constitué automatiquement à partir des techniques précédentes. Concrètement, nous avons lancé une extraction de correspondances, au niveau des formes simples, sur l'intégralité du corpus JOC (avec ABIJ et P0)[214]. L'ensemble des correspondances ainsi obtenues représente en quelque sorte un « dictionnaire » bilingue sommaire, relatif aux formes simples constituant le corpus[215].

Nous avons ensuite filtré les résultats de cette extraction avec différents paramètres, afin d'obtenir d'autres dictionnaires, contenant moins de bruit mais moins complet. On peut utiliser l'évaluation de ces extractions filtrées, sur l'échantillon de référence, comme indicateur du bruit et du silence lié à chaque dictionnaire.

III.3.7.1 Dictionnaires issus du filtrage des extractions

Dans un premier temps, nous avons extrait huit dictionnaires à partir du filtrage absolu appliqué à l'extraction sur le corpus entier. On obtient donc une série de neuf dictionnaires du plus large au plus sélectif, avec pour chacun d'eux les valeurs de rappel, précision et F-mesure liées à l'extraction dont ils sont issus (cf. tableau 75).

Extraction	***complète***	***abs 1***	***abs 2***	***abs 3***	***abs 4***	***abs 5***	***abs 6***	***abs 7***	***abs 8***
Paramètre	**0**	**0,25**	**0,5**	**1**	**2,5**	**4**	**6**	**10**	**20**
P %	71,7	72,6	78,3	82,9	86,8	88,8	90,5	91,7	93,2
R %	52,0	51,8	50,2	45,9	39,4	35,9	32,5	28,8	22,7
F %	60,3	60,5	61,2	59,1	54,2	51,2	47,8	43,8	36,5

tableau 75 : *paramètres de filtrage absolu et évaluation des extractions correspondantes*

La colonne en gris correspond au dictionnaire issu de l'extraction non filtrée.

[214] Ne disposant pas d'une caractérisation des unités polylexicales pour l'ensemble du corpus, nous ne pouvions lancer cette nouvelle extraction qu'au niveau des formes simples.

Dans un second temps, nous avons recouru au filtrage différentiel, avec 11 paramètres différents.

Extraction	***complète***	***diff 1***	***diff 2***	***diff 3***	***diff 4***	***diff 5***	***diff 6***	***diff 7***	***diff 8***	***diff 9***	***diff 10***	***diff 11***
Paramètre	**1**	**1,05**	**1,2**	**1,5**	**2**	**2,5**	**3**	**4**	**5**	**6**	**7**	**10**
$P_{dico\ \%}$	71,7	77,0	84,0	89,5	92,8	94,8	95,6	96,2	96,7	97,2	97,4	98,1
$R_{dico\ \%}$	52,0	45,4	40,2	35,6	30,7	27,8	25,6	22,4	20,1	18,3	16,8	13,5
$F_{dico\ \%}$	60,3	57,2	54,4	50,9	46,1	42,9	40,4	36,4	33,2	30,8	28,7	23,7

tableau 76 : *paramètres de filtrage différentiel et évaluation des extractions correspondantes*

Ces dictionnaires peuvent ensuite être utilisés pour constituer un nouvel indice. Pour deux unités *u* et *u'*, nous avons calculé la probabilité conditionnelle d'obtenir une correspondance avec *u'* connaissant *u* :

$$d = p(u'/u) = \frac{n_{12}}{n_1} \tag{112}$$

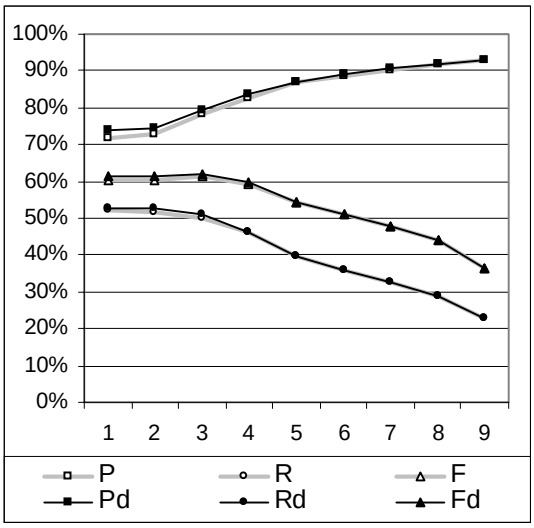
où n_1 est le nombre de couples de correspondances où *u* apparaît et n_{12} le nombre de couples (*u*,*u'*) extraits.

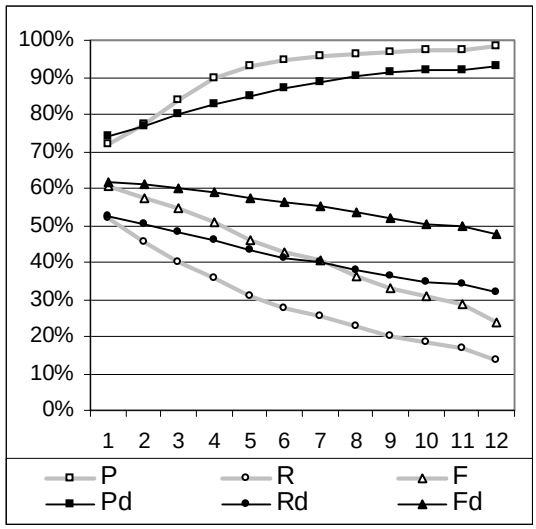
III.3.7.2 Extractions issues des dictionnaires

Pour chaque dictionnaire, nous avons effectué une nouvelle extraction (sur le corpus de référence seulement) en nous basant sur cet indice, avec l'algorithme ABIJ. Le résultat de chaque extraction a ensuite été évalué avec les 3 mesures habituelles, notées P_d, R_d et F_d.

Nous avons représenté, figure 60 les valeurs de *P*, *R* et *F* liées à chaque extraction de dictionnaire (en blanc) et les résultats P_d, R_d et F_d des extractions basées sur les indices *d* déduits de ces dictionnaires (en noir). Les résultats chiffrés figurent dans tableau 124 et tableau 125 de l'annexe.

[215] Ce « dictionnaire » présente cependant la particularité suivante : il enregistre le nombre de fois qu'une correspondance apparaît. Il permet donc de stocker, pour chaque couple d'équivalent traductionnel, la probabilité liée au couple. Par abus de langage nous utiliserons le terme *dictionnaire* pour désigner cet ensemble de probabilités issu d'une extraction.





filtrage absolu *filtrage différentiel*

figure 60 : *précisions, rappels et F-mesure des dictionnaires et des extractions déduites de l'indice d*

Ces résultats dénotent des comportements très différents suivant la méthode de filtrage impliquée :

- En ce qui concerne le filtrage absolu, on constate un net chevauchement des trois courbes : pour chaque extraction, l'indice *d* propose exactement les mêmes associations que celles du dictionnaire dont il est issu. Cela montre les limites de notre simulation : un véritable dictionnaire apporterait des informations exogènes, non inscrites dans le comportement statistique des lexies du corpus. Tandis que ces dictionnaires n'apportent rien de nouveau, dans la mesure où ils se contentent d'enregistrer les résultats déjà obtenus par les méthodes purement formelles.

- Bien que les précédentes remarques restent vraies avec le filtrage différentiel, on constate que les résultats obtenus avec *d* dépassent toutes les valeurs liées aux extractions de dictionnaires.

Pourquoi cette différence de comportement ? L'explication en est simple :

- Avec le filtrage absolu, les couples retenus sont toujours les mêmes, quels que soient les couples de phrases où ils apparaissent ; ainsi les couples rejetés par le

filtrage à l'intérieur du corpus d'évaluation sont de même rejetés à l'extérieur, dans le reste du corpus. L'extraction sur le corpus entier n'apporte donc pas d'information supplémentaire par rapport au corpus d'évaluation, et le dictionnaire filtré aurait donné peu ou prou les mêmes résultats si l'on s'était contenté d'y enregistrer les seules correspondances du corpus d'évaluation.

– A l'opposé, le filtrage différentiel est sensible au contexte, et les appariements retenus dépendent étroitement du couple de phrases où ils apparaissent. Le principe de ce filtrage étant la compétition entre les associations, il suffit qu'un couple apparaisse une fois dans un contexte où il n'a pas de concurrence pour qu'il soit retenu. Or, par la vertu de la combinatoire linguistique, il arrive toujours qu'un couple correct ne subisse plus de concurrence, et « survive » au filtrage le plus draconien. Ainsi, de nombreux couples corrects perdus par le filtrage différentiel au niveau du corpus d'évaluation, réapparaissent plus loin, à un autre endroit du corpus complet.

Par exemple, le dictionnaire *diff11* a été obtenu à partir d'une extraction filtrée totalisant un rappel de 13,5 % sur le corpus d'évaluation. L'extraction issue de *d* obtient un rappel de 31,8 % : ainsi environ 18 % de couples corrects, éliminés dans le corpus d'évaluation, ont été récupérés à l'extérieur de ce corpus.

Cette caractéristique du filtrage différentiel est intéressante et montre qu'il est plus approprié pour la constitution d'un dictionnaire : à précision et rappel égal, il fournit des informations plus diversifiées que le filtrage relatif.

Enfin, il peut être intéressant de confronter les résultats des différentes extractions liées à l'indice *d* avec les résultats filtrés des extractions simples liées à P0. On ne constate pas une amélioration décisive des résultats, mais les résultats issus de *d* correspondent globalement aux meilleurs couples (*P*,*R*) :

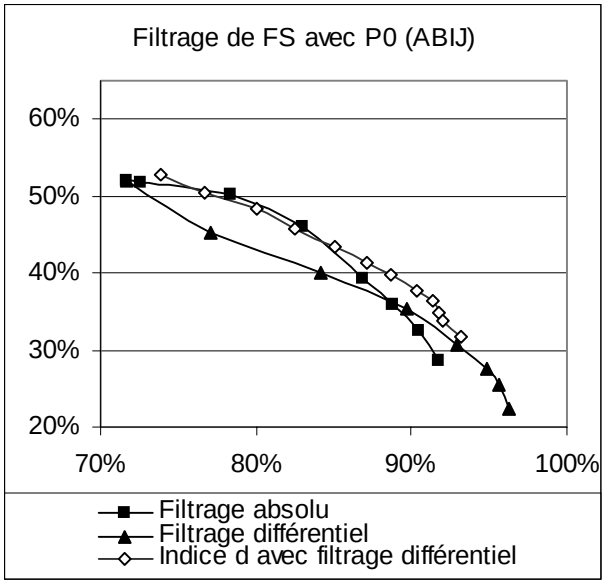


figure 61 : *comparaison des extractions simples de P0 avec les extractions déduites de l'indice d* (*avec filtrage différentiel*)

III.3.8 Algorithme EM

Pour que cette évaluation soit complète, il nous faut mettre en œuvre les méthodes basées sur l'algorithme EM qui ont donné d'excellents résultats. Nous avons choisi d'implémenter le modèle B exposé précédemment (cf. p. 433).

En reprenant les équations (86),(87) et (88), on a :

$$\begin{aligned} i(u,u') &= \log\left(\frac{\lambda^+}{\lambda^-}\right)^{c(u,u')} \cdot \left(\frac{1-\lambda^+}{1-\lambda^-}\right)^{Cooc(u,u')-c(u,u')} \\ &= c(u,u') \cdot \log\left(\frac{\lambda^+}{\lambda^-}\right) + (Cooc(u,u') - c(u,u')) \cdot \log\left(\frac{1-\lambda^+}{1-\lambda^-}\right) \\ &= c(u,u') \cdot k_1 + (Cooc(u,u') - c(u,u')) \cdot k_2 \end{aligned} \quad (113)$$

En nous basant sur les valeurs empiriques obtenues par Melamed (1998a), nous avons fixé les constantes à : $\lambda+ = 0{,}8$ et $\lambda- = 0{,}004$, ce qui donne les deux facteurs multiplicatifs : $k_1 = 2{,}30103$ et $k_2 = -0{,}69723$

Cette fois, l'algorithme EM requiert le lancement de l'extraction sur le corpus JOC *entier* à chaque itération, afin d'en tirer les comptes de connexion *c*(*u*,*u*') de manière exhaustive. Dans la mesure où nous n'avons effectué la lemmatisation et l'identification des unités polylexicales que sur le corpus d'évaluation (correspondant à 769 binômes sur 69 160), il ne nous était pas possible de lancer les tâches LEX ou LEM. Nous nous sommes donc borné à l'extraction des formes simples.

En ce qui concerne l'évaluation, elle a été effectuée comme précédemment sur la base des couples de référence : les valeurs obtenues ne correspondent donc pas à l'évaluation de *tous* les couples extraits, pour les 69 160 binômes, mais aux seuls couples du corpus de référence. On peut donc les comparer, avec exactitude, aux valeurs des résultats antérieurs.

A l'initialisation, le processus de convergence est amorcé avec l'indice P0 (au lieu de RV, comme Melamed, 1998a) [216]. Le processus se termine lorsque 99,9 % des couples ont atteint la stabilité d'une itération à l'autre.

A notre grande surprise, trois itérations ont suffit à la convergence vers des appariements stables (pour les extractions de Melamed, la convergence n'est atteinte qu'au bout de 18 itérations). Les résultats sont consignés dans le tableau 77 :

	P	***R***	***F***
Amorçage : P0	71,6 %	51,9 %	60,2 %
Itération 1	73,0 %	52,4 %	61,0 %
Itération 2	73,3 %	52,5 %	61,2 %
Itération 3	73,5 %	52,5 %	61,3 %

tableau 77 : *résultats des itérations successives avec l'algorithme EM pour la tâche FS* (*modèle B*)

In fine, la F-mesure réalise une progression très modeste de 1,1 %. P0 seule atteint pratiquement les résultats optimaux. Par rapport aux résultats de l'indice mixte PC, obtenant une F-mesure de 61 %, le gain devient négligeable.

On peut supposer que la progression aurait été plus importante avec un corpus d'apprentissage plus petit, qui aurait demandé un plus grand nombre d'itérations (comme celui de Melamed, 1998a, comprenant 29 614 versets de la Bible). L'algorithme EM tire le meilleur parti des indices distributionnels et du critère de biunivocité, car la concurrence entre appariements joue sur tout le corpus, et non seulement à l'intérieur de chaque binôme. Dès lors, les associations indirectes sont correctement inhibées, et disparaissent itération après itération : c'est ce qui fait la supériorité de la méthode itérative. Mais lorsque le corpus atteint une taille critique, à partir de laquelle la plupart des couples deviennent redondants et entretiennent une combinatoire variée (sur le plan syntagmatique), les indices distributionnels comme P0 ou RV suffisent à résoudre la plupart des cas d'association indirecte.

Visiblement, le corpus JOC n'est pas loin d'une telle taille critique. On peut même faire l'hypothèse que la progression aurait été moindre, voire nulle, si l'on avait pu effectuer la tâche LEM, qui exploite de façon optimale le critère de biunivocité (grâce à l'identification des unités polylexicales) tout en réduisant la variabilité traductionnelle.

III.3.9 Propriétés formelles des extractions : l'entropie conditionnelle comme indicateur *a priori*

Les résultats des méthodes précédemment étudiées ont donné lieu à des évaluations basées sur la comparaison avec des couples de référence. Or ces évaluations peuvent également être corrélées avec les propriétés formelles des extractions, calculables sans recourir au « *gold standard* ».

En effet, on peut supposer que plus une extraction est proche des correspondances de référence, plus elle présente de régularités dans ses appariements. A rebours, il est vraisemblable que les couples erronés soient distribués de façon plus aléatoire et irrégulière que les couples corrects, car la relation de traduction est le seul « ordre » sous-jacent à la mise en relation d'unités cibles avec des unités sources. C'est ce que suggère l'exemple du

[216] En fait nous avons aussi effectué ces calculs avec RV, mais un bogue ne nous a pas permis d'en exploiter les résultats, et les calculs étant très longs nous n'avons pas relancé de nouvelle extraction. De toutes façons, P0 donne des résultats quasiment identiques.

mot anglais *against*, dont les correspondances de référence donnent trois équivalents type sur dix occurrences, alors qu'un tirage aléatoire en donne dix sur dix :

Correspondances extraites manuellement	***Correspondances extraites au hasard***
(*against, à l'encontre de*)	(*against, par*)
(*against, à l'encontre de*)	(*against, procédure*)
(*against, à l'encontre de*)	(*against, moratoire*)
(*against, au détriment de*)	(*against, à l'encontre de*)
(*against, contre*)	(*against, dont*)
(*against, contre*)	(*against, contre*)
(*against, contre*)	(*against, effectivement*)
(*against, contre*)	(*against, charges*)
(*against, contre*)	(*against, Etat membre*)
(*against, contre*)	(*against, qui*)

tableau 78 : *les correspondances extraites manuellement sont plus régulières que les correspondances extraites au hasard*

Ainsi si l'on prend un couple (*u*,*u*') tiré au hasard, la connaissance de *u* ne permet pas de prévoir *u*'. Autrement dit, l'information apportée par *u* sur l'occurrence de *u*' est minimale. Tandis que si les couples sont correctement appariés, l'information apportée par *u* sur *u*' doit être, au contraire, non négligeable.

Supposons que l'unité *u* ait trois occurrences, et que pour chacune de ses occurrences, elle soit appariée avec *u*'. On aura :

$$p(u'/u)=1$$

L'information apportée par *u* sur l'occurrence de *u*' est maximale. Autrement dit, l'information apportée par l'occurrence de *u*', sachant *u*, est nulle :

$$I(u)=-\log(p(u'/u))=0$$

Maintenant, supposons que pour chacune de ses occurrences, *u* soit appariée avec une unité différente : u'_1, u'_2 et u'_3.

L'information apportée par l'occurrence de u'_1, sachant *u* ,est de :

$$I(u)=-\log(p(u'_1/ u))=-\log(1/3)$$

L'information moyenne apportée par les occurrences de u'_1, u'_2 ou u'_3, sachant *u*, est donc de :

$$I(u) = -1/3 \log(p(u'_1/ u)) \; -1/3 \log(p(u'_2/ u)) \; -1/3 \log(p(u'_3/ u))$$

Sous une forme plus générale, cette information s'écrit, en considérant toutes les unités *u*' du vocabulaire *V*' :

$$I(u) = -\sum_{u' \in V'} p(u'/u) \cdot \log p(u'/u)$$

Cette information exprime ce que les unités appariées avec *u* ont d'inattendu. Plus souvent *u* est impliquée dans des appariements erronés, plus ses correspondances seront imprévisibles, et plus cette information sera importante.

Si l'on prend la moyenne de cette information sur toutes les unités *u* du vocabulaire *V*, on obtient une évaluation globale de la régularité des correspondances. Notons $H(V'/V)$ cette quantité :

$$H(V'/V) = -\sum_{u \in T} \sum_{u' \in T'} p(u)\, p(u'/u) \cdot \log p(u'/u) = -\sum_{u \in T} \sum_{u' \in T'} p(u,u') \cdot \log p(u'/u) \qquad (114)$$

Il s'agit de l'entropie conditionnelle de *V*' par rapport à *V* (à travers les correspondances). On peut calculer la réciproque :

$$H(V/V') = -\sum_{u \in T} \sum_{u' \in T'} p(u')\, p(u/u') \cdot \log p(u/u') = -\sum_{u \in T} \sum_{u' \in T'} p(u',u) \cdot \log p(u/u') \qquad (115)$$

Nous avons calculé l'entropie conditionnelle, dans les deux sens, pour six groupes d'extractions concernant les lexies (tâche LEX) :

- l'extraction maximale tirée des couples de références (i.e. les couples de référence n'impliquant pas de répétition dans un même binôme, ni de lexies dépassant 5 000 occurrences) ;
- sept extractions complètes avec différentes pondérations de l'indice aléatoire AL et de l'indice P0, afin d'obtenir une gradation des résultats, allant de $F = 6$ % à $F = 65$ % ;
- les extractions complètes avec AMAX, pour les six indices ;
- les extractions complètes avec ABIJ, pour les six indices ;
- des extractions avec AMAX, pour les six indices, après un filtrage différentiel de rapport 2 ;

– des extractions avec ABIJ, pour les six indices, après un filtrage différentiel de rapport 2.

Ces groupes ont été choisis pour présenter un large éventail de résultats et de structures dans la distribution des correspondances. Les valeurs de précision s'échelonnent entre 6 et 100 %, celles de rappel entre 3 % et 69 %, et celles de *F* entre 3 % et 82 %.

Du point de vue de la structure formelle des extractions, certaines sont biunivoques à l'intérieur de chaque binôme (avec ABIJ) et d'autres non (avec AMAX). Certaines sont complètes, avec un nombre à peu près constant d'appariements, et d'autres sont filtrées et présentent un nombre très faible de couples (jusqu'à 148, pour le filtrage de IM). Bien entendu, il serait inutile de chercher une quelconque corrélation entre *H* et les valeurs de rappel, car *H* est censée refléter la proportion de couples erronés pour chaque lexie et non la complétude d'une extraction par rapport à la référence.

En revanche, la corrélation entre *H* et *P* est patente. Plus précisément, il nous est apparu que la corrélation était plus forte avec H_{max}, la plus grande des valeurs $H(V'/V)$ et $H(V/V')$ (c'est-à-dire qu'on tient compte du sens le moins favorable). Rappelons que le sens n'est pas indifférent dans la mesure où l'algorithme AMAX est antisymétrique, ce qui explique que le sens français / anglais obtienne souvent la plus faible entropie.

Pour chaque extraction, nous avons représenté la précision en fonction des valeurs de H_{max}. On obtient les nuages de points de la figure 62.

Sur la base de ces résultats empiriques, on constate que la précision est presque une fonction linéaire de H_{max}. A l'intérieur de chaque groupe, la corrélation est encore plus nette : elle apparaît très clairement avec les différentes valeurs liées aux tirages aléatoires, qui sont très espacées. Dans tous les cas, la croissance de H_{max} indique clairement une dégradation de la précision (sauf pour certaines valeurs de *P* très rapprochées).

Par ailleurs, le nombre de correspondances de l'extraction s'avère être un paramètre important : moins une extraction contient d'appariements, et moins elle dégage d'entropie, ce qui paraît normal puisque la variabilité des traductions diminue forcément avec le nombre de couples proposés : cela explique que les couples de référence n'obtiennent pas la plus petite valeur d'entropie. Néanmoins, si l'on ne tient pas compte des extractions filtrées, l'entropie atteint son minimum avec les couples manuellement extraits.

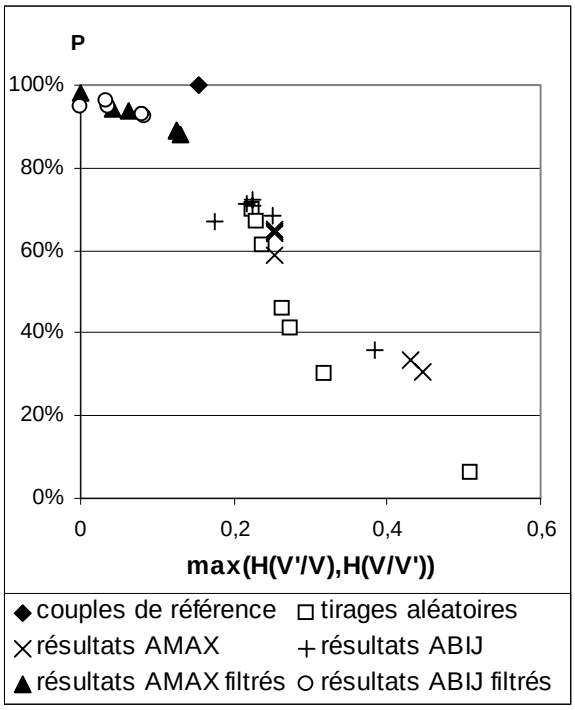


figure 62
précision des extractions en fonction de l'entropie conditionnelle

Une telle mesure présente donc un double intérêt : d'une part, elle permet d'effectuer l'évaluation sommaire d'un ensemble d'extractions de correspondances, en l'absence de couples de référence. Par ailleurs, elle fournit une caution aux couples de référence obtenus manuellement.

Certains auteurs proposent d'autres mesures, comme l'information mutuelle, qui permet d'évaluer « à quel degré une variable aléatoire permet d'en prédire une autre »[217] (Melamed, 1997d : 2). Avec nos notations, l'information mutuelle des distributions des lexies sources et cibles dans les correspondances s'écrit :

$$I = \sum_{u \in T} \sum_{u' \in T'} p(u', u) \cdot \log \frac{p(u, u')}{p(u)\, p(u')} \qquad (116)$$

Nous avons testé une autre mesure, plus élémentaire dans son principe : le nombre total de correspondances entre lexies types. L'idée est simple : plus les correspondances sont régulières, et moins les correspondances sont variées. Si l'on note C_{types} l'ensemble de

[217] "*how well one random variable predicts another.*"

correspondances entre lexies types (chaque correspondance n'étant considérée qu'une fois), on a :

$$N_{Ctypes} = \left| C_{types} \right| \tag{117}$$

Les valeurs observées de précision en fonction de l'information mutuelle et de N_{Ctypes} sont représentées respectivement figure 63 et figure 64.

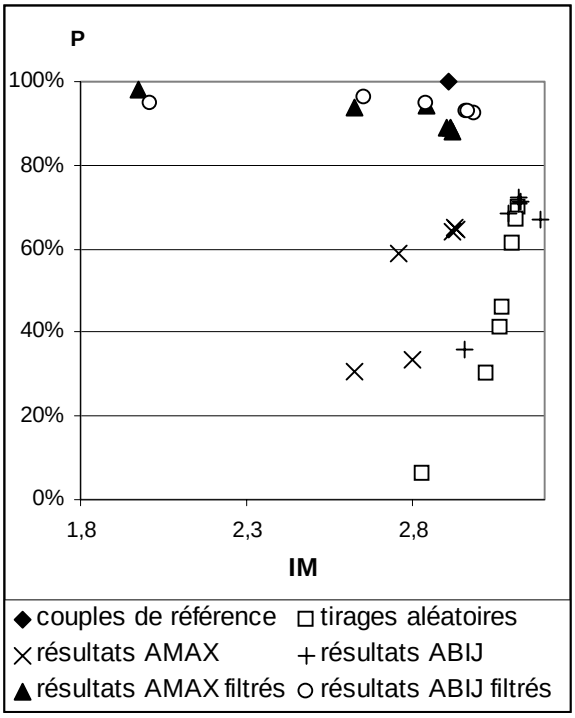


figure 63 : *précision des extractions en fonction de l'information mutuelle*

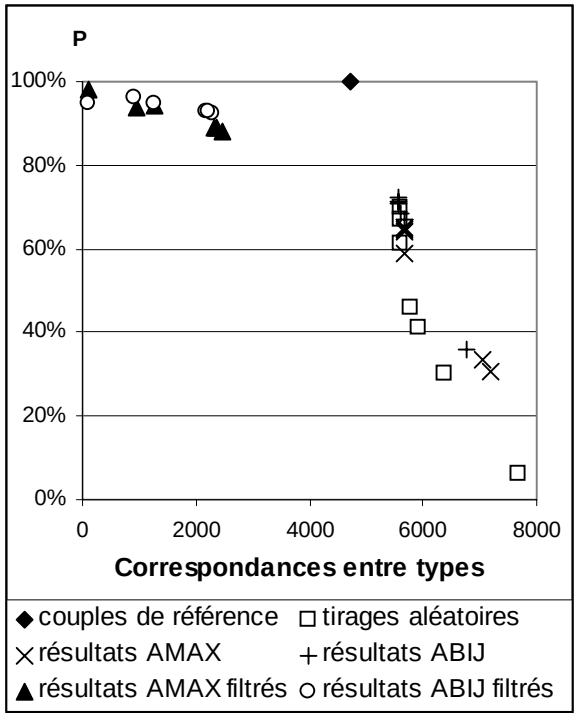


figure 64 : *précision des extractions en fonction des correspondances entre types*

Il semblerait que l'information mutuelle ne soit pas directement liée à la validité d'une extraction, à la différence du calcul élémentaire du nombre de correspondances types. Cet insuccès de l'information mutuelle était prévisible puisque l'indice donne des résultats médiocres dans la tâche d'extraction. Les correspondances entre unités peu fréquentes y pèsent de tout leur poids : or, comme on l'a vu précédemment, les correspondances entre hapax sont nombreuses et n'ont qu'une faible pertinence statistique.

Afin de quantifier plus précisément les observations précédentes, nous avons calculé, entre chaque mesure et les valeurs de précision, le coefficient de corrélation linéaire (cf. tableau 79).

Corrélations	***Cooccurrences entre vocables***	***Information mutuelle***	***Entropie conditionnelle maximum***
Toutes les extractions	-0,86	-0,27	-0,95
Extractions complètes	-0,94	0,44	-0,91

tableau 79 : *corrélations linéaires entre précision et mesures d'évaluation a priori*

III.3.10 Retour à l'alignement

Dans la deuxième partie de cette recherche, nous avons vu que de nombreux algorithmes d'alignement reposaient sur une forme de cercle vertueux, l'alignement grossier de segments textuels permettant d'extraire des correspondances lexicales, réutilisables ensuite pour extraire un alignement un peu moins grossier, et ainsi de suite jusqu'à stabilité.

A l'issue des développements précédents, il serait intéressant de « boucler la boucle », et d'examiner la possibilité d'améliorer les résultats de l'alignement au niveau des phrases en intégrant une mise en œuvre fine de l'extraction des correspondances.

Pour effectuer une telle vérification nous avons choisi le corpus *Verne*, pour lequel les méthodes « simples » n'ont pas donné entière satisfaction (car la traduction anglaise comporte de nombreuses contractions et omissions).

Les étapes de cette dernière évaluation sont les suivantes :

1. Alignement → calcul des cooccurrences :

A partir du meilleur alignement obtenu (cf. les résultats du tableau 38, p.331), une table de cooccurrence a été extraite, et l'indice P0 a été calculé pour tous les couples d'unités (formes simples) cooccurrentes.

2. Recalcul de l'alignement avec un nouvel indice basé sur PC avec ABIJ

Dans l'algorithme d'alignement, lors de la comparaison des segments, un nouvel indice est créé : l'idée est d'estimer, pour chaque couple de phrases candidat à l'alignement, le nombre de correspondances lexicales fiables qu'on peut en extraire. Le chemin correct sera alors probablement celui qui totalise le plus grand nombre de ces correspondances. Notons que le chemin optimal n'est plus celui qui minimise une mesure de distance, mais celui qui maximise le score de similarité.

Ainsi, pour chaque couple de phrases alignables, on effectue une extraction des correspondances sur la base de l'indice PC avec l'algorithme ABIJ. Ces correspondances sont ensuite filtrées, par l'application successive de trois filtrages : absolu, différentiel, et en fonction de la fréquence des unités. Avec ce troisième type de filtrage, ne sont retenus que les couples d'unités dont les fréquences f sont situées dans une certaine fourchette : $f_{min} \leq f < f_{max}$. On conjugue ces trois filtrages afin d'obtenir une précision maximale : si l'on peut garantir une majorité d'appariements corrects au niveau lexical, alors il est très probable que le nombre de correspondances extraites pour un couple de phrases équivalentes sera en moyenne supérieur à celui obtenu pour un couple de phrases quelconques. En outre, ces filtrages présentent des profils complémentaires :

- Le filtrage par fréquence garantit la pertinence de l'indice P0, qui se révèle peu significatif pour les unités peu fréquentes, cf. la figure 54, p. 483. Par ailleurs il permet d'éliminer les mots outils les plus fréquents, eux aussi générateurs de bruit.

- Le filtrage différentiel assure une grande précision lorsqu'il est appliqué à deux phrases équivalentes. Mais entre des phrases quelconques il perd de son efficacité, puisqu'il permet d'élire, pour chaque unité, l'appariement correct qui s'élève au-dessus des appariements erronés : si aucun appariement n'est correct, il risque de conserver le meilleur des appariements erronés.

- Le filtrage absolu pallie les défaillances du filtrage différentiel puisque son pouvoir discriminant reste identique quelles que soient les phrases comparées.

Notons que ces extractions sont lancées au niveau des couples de phrases (et non des couples de segments, pouvant inclure plus de deux phrases) : cela permet d'en sauvegarder facilement le résultat dans une table indexée par les coordonnées des deux phrases. De la sorte, l'extraction des correspondances n'est lancée qu'une seule fois par couple (lors du calcul du chemin optimal, deux mêmes phrases peuvent être comparées jusqu'à neuf fois, suivant les types de transition considérés). Dans le cas des transitions multiples, (1:2), (2:1), (1:3), (3:1), les comparaisons sont faites deux à deux : par exemple, pour évaluer le score de l'appariement $(P_i,P_{i+1},P_{i+2};P'_j)$ on effectue trois extractions de correspondance séparées, dont on additionne les scores en appliquant la distributivité :

$$Score_{PC}(P_i,P_{i+1},P_{i+2};P'_j) = Score_{PC}(P_i;P'_j) + Score_{PC}(P_{i+1};P'_j) + Score_{PC}(P_{i+2};P'_j)$$

Il s'agit certes d'une approximation, puisque de cette manière on ne respecte plus stricto sensu le critère d'affectation biunivoque (dans le cas précédent, une même unité peut apparaître dans trois couples différents) : mais il en résulte une importante économie de calcul[218]. Enfin, nous avons écarté les transitions de type (2:2), qui sont injustement favorisées par ce mode de calcul.

Pour le calcul de P0, nous avons mis en œuvre deux modèles de cooccurrences :

- Avec le modèle 1, les cooccurrences sont calculées simplement entre les phrases alignées issues de l'extraction précédente (cf. tableau 38).
- Le modèle 2 est plus tolérant : dans la mesure où l'extraction précédente présente un rappel d'environ 76 %, 24 % des binômes corrects ne sont pas pris en compte dans le précédent modèle. Or la plupart de ces binômes sont dus à des appariements erronés avec des segments situés immédiatement avant ou immédiatement après les segments de l'appariement correct. Ainsi, pour chaque segment S_i du texte source, on étend l'aire de cooccurrence au segment aligné S'_i,

[218] Avec nos transitions, il faudrait effectuer environ 9 fois plus d'extractions de correspondances lexicales. En effet, si le corpus comporte *n* x *m* phrases, il faut considérer en outre *n-1* couples et *n-2* triplets de phrases en langue source, ainsi que *m-1* couples et *m-2* triplets en langue cible. Il faut donc comparer : $(n + n\text{-}1 + n\text{-}2)\cdot(m + m\text{-}1 + m\text{-}2) \approx 9\ n*m$.

ainsi qu'aux segments précédents et suivant S'_{i-1} et S'_{i+1}. Dans le calcul de P0, les nombres d'occurrences des unités du texte cible sont donc multiplié par 3 (dans la mesure ou chaque segment cible est compté trois fois). En élargissant ainsi l'aire de cooccurrence, on génère un surcroît de bruit : mais peut-être ce bruit pourra-t-il être éliminé en appliquant un filtrage plus sélectif.

Dans une première série d'extractions, nous avons testé le score de similitude $Score_{PC}$ avec et sans filtrage. Le filtrage appliqué est une combinaison de filtrage différentiel de rapport $r = 2$ et de filtrage absolu de seuil $s = 3{,}5$, sans filtrage par fréquence : ces deux paramètres permettaient de dégager une précision voisine de 90 %, lorsqu'ils étaient appliqués indépendamment aux binômes alignés du corpus JOC (pour une étude détaillée des meilleurs paramètres, cf. infra). Les résultats de ces extractions figurent dans le tableau 80 :

	Modèle 1			***Modèle 2***		
Pas de filtrage	39,5 %	57,0 %	46,68 %	31,3 %	45,7 %	37,14 %
Filtrage ($r = 2$, $s = 3{,}5$)	74,75 %	76,20 %	75,47 %	75,10 %	83,17 %	78,93 %

tableau 80 : *résultat de l'indice* Score_{PC} *avec et sans filtrage pour les deux modèles de cooccurrence*

Ces résultats semblent confirmer la supériorité du modèle 2, à condition que les correspondances obtenues subissent un filtrage rigoureux (à l'intérieur de l'espace de recherche, plus de 95 % des couples extraits sont éliminés lors du filtrage, cf. tableau 132 de l'annexe).

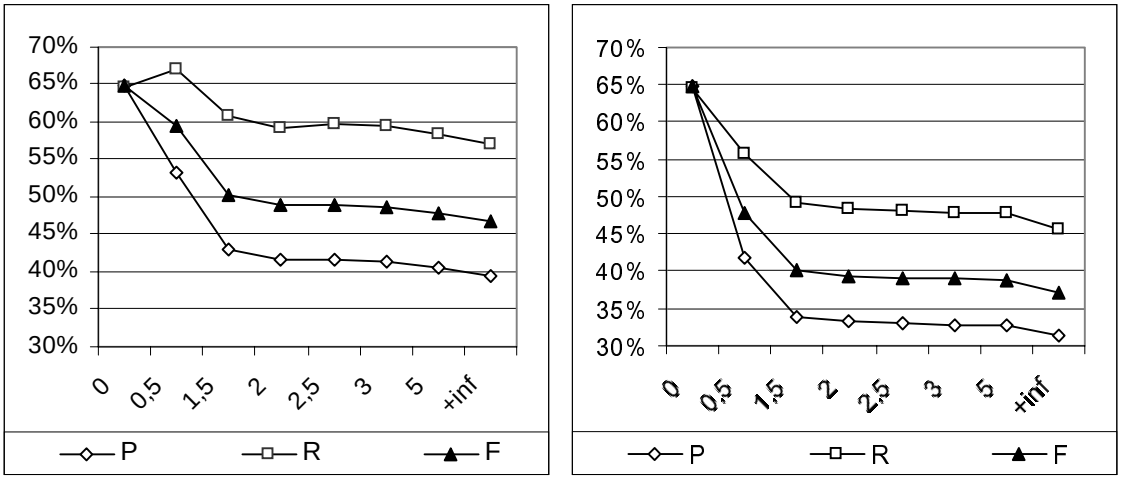
Sans filtrage, l'indice n'apporte pas une information assez discriminante, les binômes erronés obtenant pratiquement autant de correspondances que les binômes corrects : et le surcroît de bruit apporté par le modèle 2 aggrave cette tendance. En revanche, lorsqu'on applique un filtrage capable de garantir qu'une forte proportion des correspondances extraites sont correctes (les correspondances qui figurent dans le tableau 133 de l'annexe sont correctes à plus de 98 %), l'indice $Score_{PC}$ devient opérationnel et dépasse même les résultats précédemment obtenus avec la distance $Distance_{combinée}$ (cf. équation (36), p. 326).

Le modèle 2 se comporte mieux car il est conçu pour avoir un meilleur rappel dans l'extraction des correspondances – et le surcroît de bruit apporté par ce modèle de cooccurrence plus lâche est éliminé au filtrage.

On peut cumuler $Score_{PC}$ avec la distance combinée $Distance_{GC}$, afin de tirer parti simultanément des différentes sources d'information (longueurs et distributions) :

$$Score_{combiné} = k \cdot Score_{PC} - Distance_{GC} \quad (118)$$

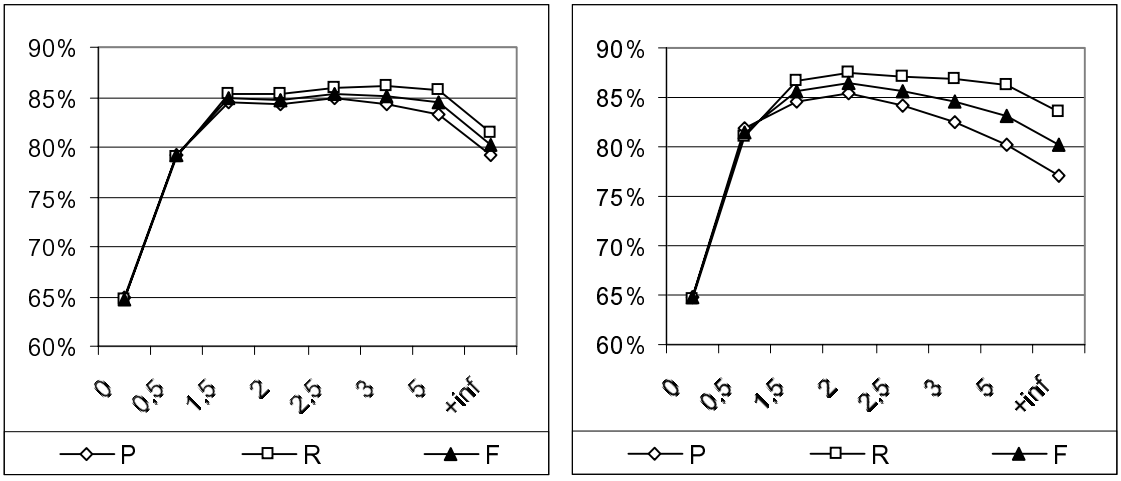
Les résultats de ce nouveau score, représentés figure 65 (cf. le tableau 126 de l'annexe) pour différentes valeurs du coefficient *k*, montrent que sans filtrage $Score_{PC}$ est moins performant que $Distance_{GC}$:



Modèle 1 *Modèle 2*

figure 65 : *résultats liés à* Score$_{\text{combiné}}$ (*sans filtrage*), *en fonction du coefficient* k (+*inf indique que seul* $Score_{PC}$ *intervient*)

Mais lorsque $Score_{PC}$ est mis en œuvre avec le filtrage précédent, on constate non seulement que cet indice obtient seul de meilleurs résultats que $Distance_{GC}$, mais en outre que les informations liées à ces deux indices sont complémentaires et cumulatives, puisque les courbes sont convexes et qu'un maximum est atteint aux alentours de $P = 87$ % pour le modèle 2 (cf. la figure 66 ci-après, et le tableau 127 de l'annexe). Notons que ce dernier modèle se distingue essentiellement, comme on pouvait s'y attendre, par un meilleur rappel.



Modèle 1 *Modèle 2*

figure 66 : *résultats liés à* $\text{Score}_{\text{combiné}}$ (*avec filtrage*) *en fonction du coefficient* k

Les meilleurs résultats sont atteints avec le modèle 2 et un coefficient $k = 2$: on obtient une F-mesure de 86,88 %.

Au filtrage précédent, nous avons ajouté un filtrage par fréquence, afin d'en déterminer les effets. Si l'on élimine les correspondances incluant des unités de basse fréquence, on constate une dégradation progressive des résultats.

	P	***R***	***F***
Pas de filtrage	85,57%	88,24%	86,88%
Elimination si f < 2	85,57%	88,24%	86,88%
Elimination si f < 3	85,46%	88,07%	86,75%
Elimination si f < 5	85,23%	87,78%	86,49%
Elimination si f < 7	84,53%	86,92%	85,71%
Elimination si f < 10	84,36%	87,25%	85,78%
Elimination si f < 20	81,35%	84,55%	82,92%

tableau 81 : *élimination des unités de basse fréquence cumulée avec un filtrage de paramètre* r = 2, s = 3,5 (*modèle 2*, $\text{Score}_{\text{combiné}}$ *avec* k = *2*)

L'élimination des hapax ne change rien aux résultats : tout se passe comme si les couples intégrant des unités peu fiables avaient déjà été éliminés par les filtrages différentiels et absolus, et que les couples restant étaient tous corrects. Cela explique la

dégradation des résultats : l'information utile s'appauvrit progressivement et l'indice perd de son efficacité.

Lorsqu'on effectue un filtrage éliminant les unités les plus fréquentes, dont beaucoup sont des signes de ponctuation et des mots outils, le constat est différent. La figure 67 (cf. le tableau 128 de l'annexe) montre l'évolution des résultats en fonction du seuil de fréquence maximum (nous noterons $s_{fréqmax}$), au-delà duquel les couples sont éliminés : aux alentours de 150, on obtient une amélioration d'environ 2 % pour *F1* (modèle 1) et 1 % pour *F2* (modèle 2).

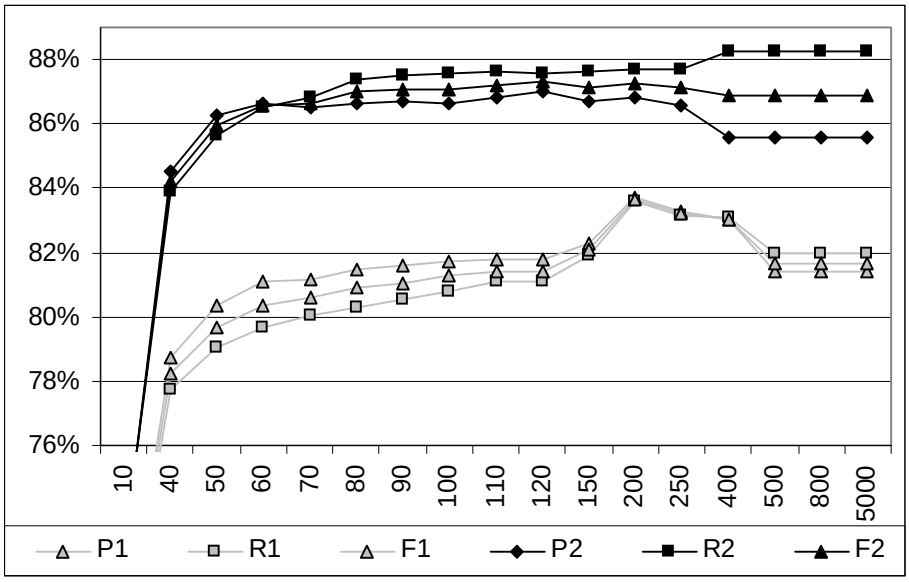


figure 67 : *élimination des unités de fréquence f > x, cumulée avec un filtrage de paramètre* r = 2, s = 3,5 (*modèles 1 et 2*, $\text{Score}_{\text{combiné}}$ *avec* k = *2*)

Le filtrage des unités de haute fréquence permet donc d'éliminer des correspondances erronées qui ont résisté aux filtrages différentiel et absolu.

Jusqu'à présent, nous avons choisi arbitrairement les paramètres fixés pour ces deux derniers filtrages. Nous avons cherché à évaluer l'impact de ces paramètres pris indépendamment.

– pour le filtrage différentiel, nous avons testé les valeurs suivantes de *r* :

1 ; 1,25 ; 1,5 ; 1,75 ; 2; 2,25 ; 2,5 ; 2,75 ; 3,25 ; 3,5 ; 3,75 ; 4 ; 4,25 ; 4,5.

– pour le filtrage absolu, nous avons testé les valeurs suivantes de *s* :
0 ; 1 ; 2 ; 3 ; 3,5 ; 4 ; 5 ; 6 ; 7 ; 8 ; 9 ; 10.

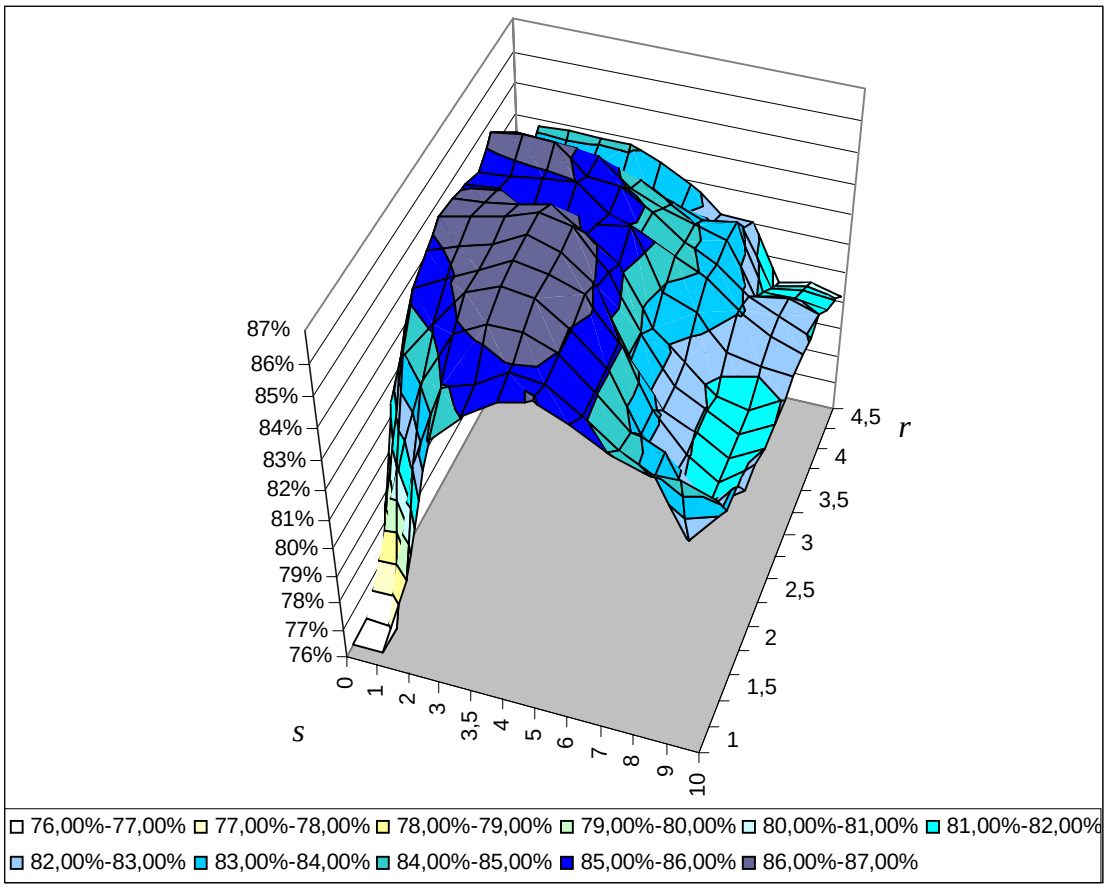


figure 68 : *F en fonction des paramètres de filtrage* r *et* s *pris indépendamment* (*modèle 2*, $\text{Score}_{\text{combiné}}$ *avec* k = *2*)

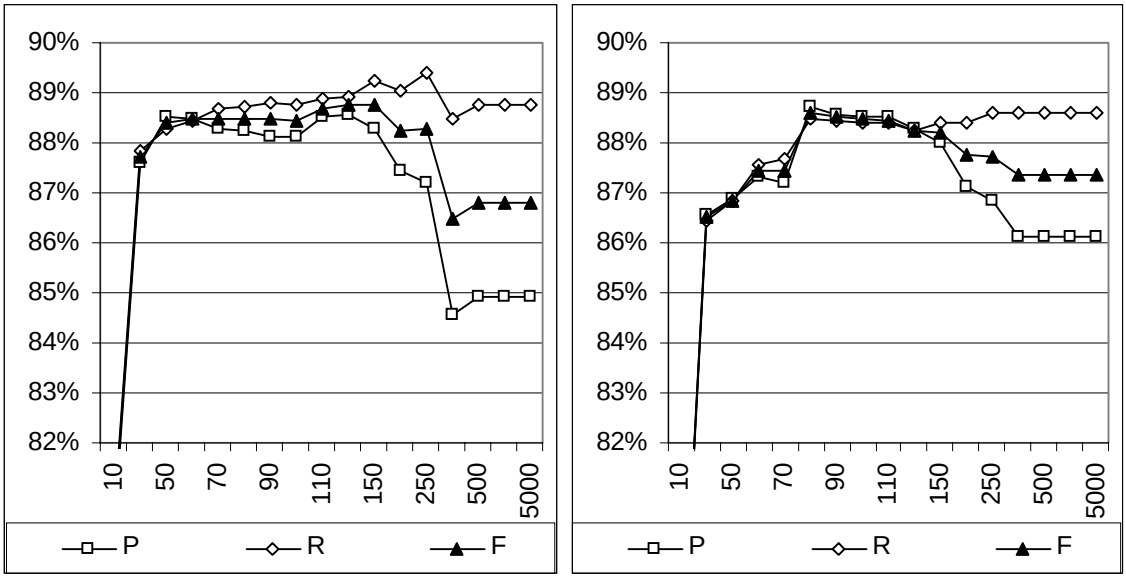
La figure 68 donne une représentation tridimensionnelle des résultats : *F* en fonction des deux paramètres de filtrage (*r*,*s*) (cf. le tableau 130 de l'annexe pour les valeurs numériques de précision et de rappel).

On constate que la courbe, globalement convexe, présente toutefois des irrégularités et comporte deux maxima. Le maximum absolu (pour les points calculés) est atteint avec *r* = 2,25 et *s* = 3,5. Les irrégularités sont principalement dues à l'évolution de la précision (cf. figure 82 et figure 83 de l'annexe), celle du rappel présentant un profil plus proche de la convexité.

– *Nouvelle extraction*

Le filtrage précédemment étudié (r = 2 et s = 3,5 sans filtrage par fréquence) permet d'atteindre un rappel assez élevé, de 88,24 %. Ces résultats ne sont pas optimaux, mais dans la mesure où ils réalisent une progression, par rapport à l'alignement d'après lequel on a calculé les cooccurrences, ils sont suffisants pour entreprendre une nouvelle itération[219].

Nous avons donc utilisé l'alignement résultant pour recalculer les cooccurrences et les valeurs de P0, en utilisant les deux modèles de cooccurrence 1 et 2. Dans cette deuxième étape, nous avons à nouveau appliqué un filtrage combiné, avec différents seuils pour les unités de haute fréquence : on obtient les résultats de la figure 69 (cf. tableau 129 de l'annexe).



Modèle 1 *Modèle 2*

figure 69 : *deuxième étape - élimination des unités de fréquence f > x*
(+ *filtrage de paramètre* r = 2, s = 3,5 $Score_{combiné}$ *avec* k = *2*)

On constate que le modèle 1 donne de meilleurs résultats, sauf lorsqu'on n'applique pas de filtrage par fréquence. Cette fois, le rappel de l'alignement dont on a extrait les

[219] Notre but est de montrer qu'il est possible d'améliorer les résultats, étapes après étapes, sans nécessairement « guider » le processus en employant des paramètres *ad hoc* : c'est pourquoi nous ne cherchons pas à utiliser les paramètres optimaux, dont on ne peut présumer avec exactitude en l'absence de l'alignement de référence.

cooccurrences est suffisant, et l'élargissement de l'aire de cooccurrence ne présente plus d'intérêt.

Si l'on compare l'évolution des résultats en fonction du seuil de fréquence, pour les différentes séries obtenues, on constate que le maximum de F-mesure n'est pas toujours atteint pour les mêmes seuils. Etrangement, comme le montre la figure 70, il semblerait que meilleur est le modèle de cooccurrence, et plus le filtrage gagne à être sélectif : pour le modèle 1 de l'étape 2, on a $F = 88{,}75$ %, avec $s_{fréqmax} = 150$ occurrences, alors que pour l'étape 1, le maximum s'établit aux alentours de $s_{fréqmax} = 200$. Dans tous les cas, on constate que le filtrage des unités de haute fréquence permet une amélioration des résultats.

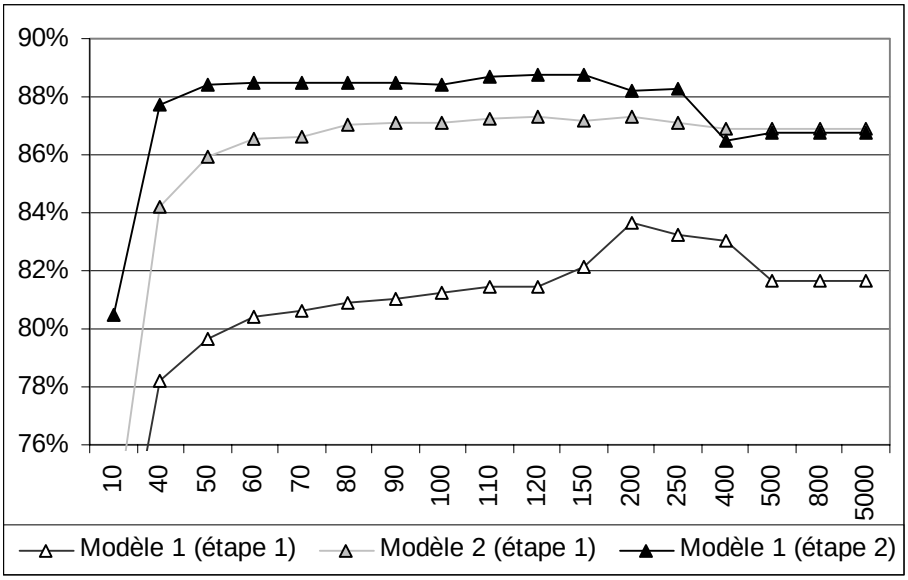


figure 70 : *évolutions comparées de F avec les filtrages par fréquence pour trois différentes séries*[220]

L'interprétation de ce phénomène est délicate. Il semblerait que l'on soit en présence de deux effets antagonistes :

- d'une part, comme nous l'avons démontré sur le JOC, la fiabilité des correspondances augmente avec la fréquence des unités concernées (sauf pour les mots outils très fréquents).
- d'autre part, les unités fréquentes sont susceptibles d'apparaître dans des phrases voisines, et cette dispersion peut éventuellement générer des appariements de

phrases erronés. Les unités peu fréquentes étant moins dispersées, la présence d'une correspondance correcte avec ces unités est plus informative pour l'alignement[221]. Cette hypothèse remet en question la signification de $Score_{PC}$: le chemin correct ne serait pas nécessairement celui qui contiendrait le plus de correspondances lexicales fiables *dans l'absolu*, mais celui qui contiendrait le plus de correspondances fiables *discriminant les phrases les unes des autres.*[222] Comme le montre le tableau 132 de l'annexe, suite aux filtrages absolus et différentiels, il reste en moyenne entre 2 et 3 correspondances par binôme, sur le chemin ; et pour un grand nombre de binômes, le filtrage laisse une seule correspondance, ou pas du tout. Avec une densité si faible, la dispersion de certaines correspondances sur des phrases voisines devient une source de bruit non-négligeable.

Ces deux phénomènes antagonistes expliqueraient le fait que les meilleurs résultats se situent au niveau des fréquences intermédiaires.

Pour l'étape 2, nous avons également étudié les variations des résultats en fonction des paramètres de filtrage *s* et *r*. L'évolution de *F* est représentée figure 71 (cf. tableau 131 de l'annexe).

Par rapport à l'étape précédente, nous avons cherché à « grossir » la zone de F-mesure optimale :

- pour le filtrage différentiel, nous avons testé les valeurs suivantes de *r* :

1,5 ; 1,75 ; 2 ; 2,25 ; 2,5 ; 2,75 ; 3 ; 3,5 ; 3,75 ; 4 ; 4,25 ; 4,5 ; 4,75.

- pour le filtrage absolu, nous avons testé les valeurs suivantes de *s* :

1,2 ; 1,5 ; 1,75 ; 2 ; 2,25 ; 2,5 ; 2,75 ; 3 ; 3,5 ; 4.

---

[220] Pour des raisons de clarté, nous n'avons pas représenté les résultats du modèle 2 de l'étape 2

[221] Ce principe est d'ailleurs mis en œuvre par Chen (1993), pour détecter les zones correspondantes consécutives à un décrochement important du parallélisme, dans le cas d'omission ou d'insertion massive.

[222] Sur cette base, on pourrait concevoir une autre méthode de filtrage, qui éliminerait tous les appariements ambigus, i.e. les correspondances qui apparaîtraient dans des binômes immédiatement voisins, et donc concurrents.

Là encore la courbe comporte des irrégularités. Au vu de ces résultats, on constate qu'il est difficile d'élaborer une méthode pour atteindre les paramètres optimaux. D'une étape à l'autre, nous avons procédé de manière empirique, en appliquant à chaque fois un filtrage déterminé arbitrairement, avec $r = 2$ et $s = 3{,}5$: il est très probable que les résultats puissent encore être améliorés en affinant le réglage des paramètres efficaces (r, s, mais aussi k et $s_{fréqmax}$).

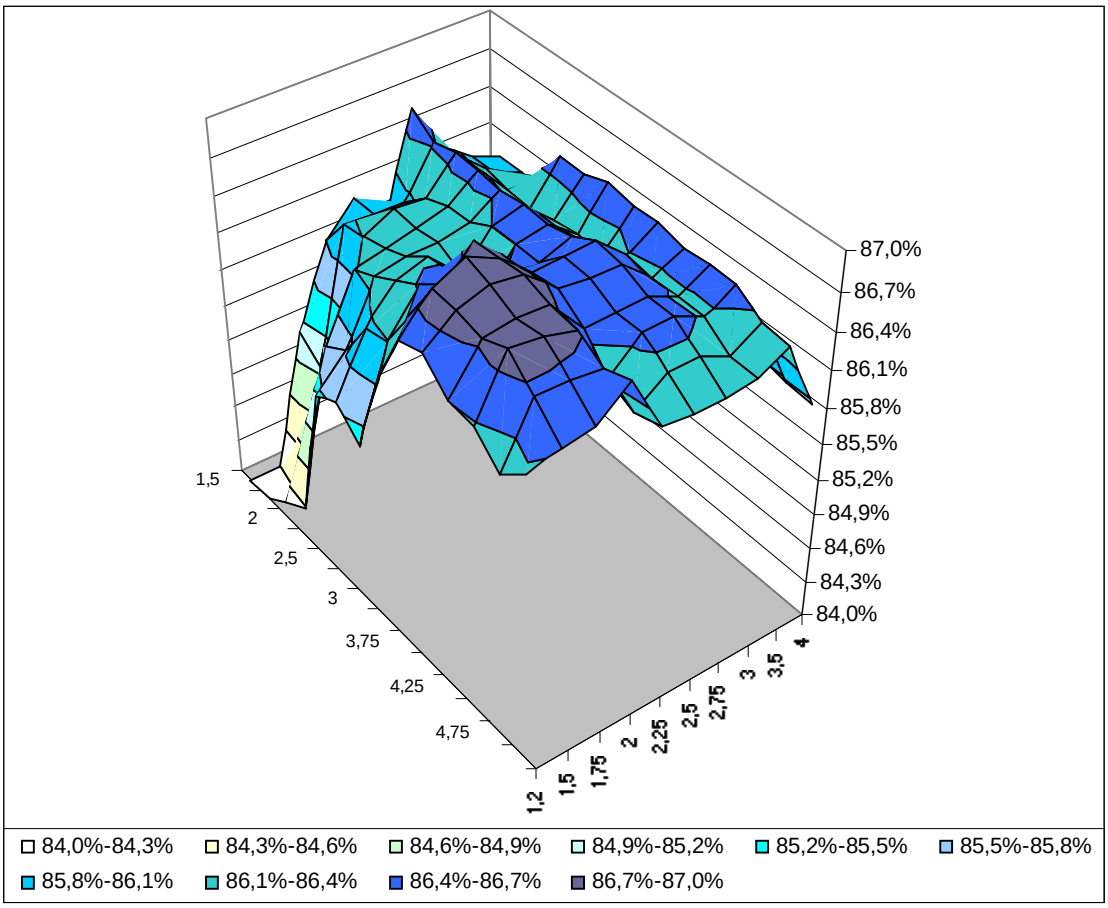


figure 71 : *deuxième étape - F en fonction des paramètres de filtrage* r *et* s *pris indépendamment* (*modèle 2*, $\text{Score}_{\text{combiné}}$ *avec* k = *2*)

En outre, rien ne nous permet d'affirmer que les paramétrages que nous avons progressivement affinés puissent être transposables à un autre corpus. Dès lors, le problème qui se pose est le suivant : quelle méthode permettrait de déterminer ces paramétrages « à l'aveugle », c'est-à-dire sans disposer de la possibilité d'évaluer les

résultats obtenus ? Dans le cas présent, il est clair que nous avons été guidé par les différentes étapes de notre évaluation : mais lorsqu'on aligne un corpus automatiquement, on ne peut évaluer en cours de route la précision et le rappel obtenu, en l'absence d'alignement de référence.

Pour mettre en œuvre automatiquement les méthodes que nous avons décrites, il faudrait pouvoir fixer les paramètres en fonction des caractéristiques du corpus. Les données pertinentes, pour la mise en œuvre d'un indice distributionnel tel que P0, sont vraisemblablement : 1/ la taille des textes, et 2/ secondairement la richesse du vocabulaire et la longueur moyenne des phrases à aligner.

Plus le corpus est grand, et plus P0 devient fiable pour l'extraction des correspondances, *k* doit donc augmenter afin de diminuer la part relative de $Distance_{GC}$; dans le même temps, les taux de filtrage peuvent être revus à la baisse, puisque le risque d'erreur diminue. Il est également possible de moduler automatiquement les paramètres de filtrage, afin d'assurer la conservation d'une proportion constante de couples d'unités correspondantes dans chaque binôme[223]. Il existe peut-être une proportion optimale qui assure de bons résultats quels que soient les textes. Pour vérifier ces hypothèses, il faudrait entreprendre des études complémentaires sur d'autres corpus.

Malgré tout, nous pensons que la méthode d'alignement présentée garde une certaine généralité, et qu'un réglage approximatif des paramètres permet néanmoins d'obtenir des résultats intéressants.

[223] par exemple, l'extraction qui obtient les meilleurs résultats (étape 2, modèle 1), implique un filtrage de paramètre $r = 2$, $s = 3{,}5$ et $s_{fréqmax} = 60$. Sur l'espace de recherche, 98,06% des couples sont éliminés, et il reste une moyenne de 1,65 correspondances par binôme sur le chemin optimal. Nous pensons que sur un corpus plus long, la proportion optimale peut être supérieure.

III.4 Conclusion de la troisième partie

Dans cette dernière partie, nous avons pris soin de distinguer l'extraction de correspondances lexicales et l'alignement : dans le premier cas, on s'intéresse à des unités de niveau lexical (mots, phrasèmes, termes, etc.) identifiées indépendamment des caractéristiques particulières du bi-texte, et pour lesquelles on ne peut présupposer qu'elles aient une contrepartie dans la portion équivalente ; dans le deuxième cas, on apparie des segments textuels (phrases, groupes de phrases, paragraphes, etc.) que l'on cherche à regrouper afin de satisfaire au mieux la propriété de compositionnalité traductionnelle. Cette distinction vise à mieux prendre en compte les phénomènes de divergence entre l'original et sa traduction, car comme nous l'avons montré, la traduction ne conserve pas toujours le découpage et le contenu sémantique des unités lexicales. Malgré cela, il n'est pas facile d'expliciter rigoureusement des critères linguistiques pour l'appariement de deux unités, même sur la base de leur équivalence dénotative. Du fait de la prégnance du contexte dans l'interprétation et la construction du sens, la notion de correspondance lexicale demeure problématique.

Mais ces difficultés théoriques n'empêchent pas, parallèlement, le développement des techniques destinées à automatiser l'extraction des correspondances.

Nous avons testé, dans une première phase, des algorithmes simples (AMAX et ABIJ) ne nécessitant qu'une seule itération sur les couples de phrases alignés. Nous avons constaté une nette supériorité de l'algorithme de meilleure affectation biunivoque, qui permet de contrôler les effets des correspondances indirectes.

Vis-à-vis des trois tâches que nous avons évaluées, il est apparu que l'identification préalable des unités polylexicales pouvait améliorer l'extraction des correspondances. Ceci s'explique simplement par la non-compositionnalité sémantique de certaines de ces unités, qui ne peuvent être traduites mot à mot. Le critère de biunivocité des correspondances est alors mieux vérifié lorsque ces unités sont considérées d'un bloc.

Par ailleurs, l'élimination de certaines variations morphologiques, au moyen d'une lemmatisation sommaire, renforce la régularité des cooccurrences parallèles, et améliore les résultats des indices basés sur les distributions.

En ce qui concerne l'indice basé sur les transfuges et cognats, nous en avons implémenté deux versions, COa et COb, la dernière produisant plus de bruit et moins de silence. Il s'est avéré que ce bruit est correctement filtré par les algorithmes étudiés, du fait de la faible concurrence entre couples de cognats au sein d'un même binôme. Ainsi, la version COb a apporté de meilleurs résultats, tant au plan du rappel que de la précision. En appliquant un léger filtrage différentiel, les effets du bruit sont pratiquement annulés, puisqu'on arrive à une précision de 90 % pour un rappel d'environ 25 %. L'information apportée par les cognats est donc utilisée à son maximum, puisque environ 21 % des couples extraits manuellement sont des candidats cognats ou transfuges. Si le rappel obtenu est supérieur à ce pourcentage, c'est qu'une partie des couples correctement extraits peut être imputée au hasard (cf. les résultats de l'indice AL, p. 455).

D'autre part, on note que les indices P0 et RV aboutissent à des résultats pratiquement identiques, avec un léger avantage toutefois pour RV. Dans la mesure où le calcul de P0 est plus complexe, le seul avantage de cet indice réside dans sa signification : P0 étant une probabilité estimant les chances d'un tirage aléatoire, il est plus facile de le combiner de manière cohérente avec d'autres probabilités, dans la modélisation d'un processus aléatoire plus complexe. Ainsi, nous avons considéré deux tirages indépendants : le tirage de deux vecteurs d'occurrence pour deux unités lexicales, et le tirage des lettres qui les constituent. Le calcul de probabilité d'un tel événement à été désigné par PC, ce qui nous a permis d'évaluer, en passant à l'opposé du logarithme, l'invraisemblance de l'hypothèse d'indépendance pour des lexies effectivement correspondantes. De la même manière, on peut concevoir une combinaison de P0 avec d'autres sources d'information, par exemple issues d'un dictionnaire bilingue.

Globalement, les résultats sont satisfaisants, même si des améliorations sont encore possibles, puisque avec ou sans mot outils, l'extraction LEM avec PC et ABIJ se situe à environ 10 % de précision et 13 % de rappel des résultats optimaux (définis par rapport aux contraintes de notre implémentation).

Par ailleurs, nous avons étudié différentes méthodes de filtrage permettant de réduire le taux de correspondances erronées. La méthode de filtrage absolu, éliminant toutes les

correspondances dont la valeur de l'indice est inférieure à un certain seuil, s'est révélée intéressante si l'on désire maintenir le rappel au-dessus de 50 %.

Par exemple, pour la F-mesure, les meilleurs résultats sont obtenus pour l'extraction LEM, avec l'algorithme ABIJ et l'indice PC, en appliquant un filtrage absolu de coefficient $x = 0{,}5$:

PC avec ABIJ pour LEM	***P***	***R***	***F***
avec filtrage absolu (*x = 0,5*)	82,2 %	55,5 %	66,3 %

tableau 82 : *meilleure F-mesure*

Le filtrage différentiel, basé sur la comparaison entre la valeur de l'indice d'un couple et la valeur du meilleur couple concurrent, semble en revanche plus adapté pour obtenir des correspondances avec un pourcentage d'erreur très faible : par exemple, on obtient une précision de 97 % avec un rappel de 32,5 % pour la tâche LEM avec l'algorithme ABIJ et l'indice PC.

Dans la perspective de l'extraction d'un glossaire bilingue, nous avons montré que le filtrage différentiel était plus adapté, puisqu'il fournit des correspondances plus variées (donc avec un rappel plus important du point de vue des correspondances types) tout en assurant une précision élevée.

Nous avons également montré que les correspondances lexicales manifestent, dans une certaine mesure, une propriété objective du bi-texte : il semble que le jeu de correspondances manuellement extraites coïncide, peu ou prou, avec la distribution qui atteint le minimum d'entropie conditionnelle. Cette propriété formelle laisse cependant une part d'indétermination : pour une distribution de correspondances donnée, il est toujours possible de diminuer l'entropie en éliminant les correspondances les moins régulières. En d'autres termes, l'entropie est aussi liée à la proportion d'unités résiduelles qui ont été éliminées par le filtrage, et cette proportion n'est pas rigoureusement inscrite dans la structure bi-textuelle : certes, pour la plupart des traductions elle se situe à l'intérieur d'une certaine fourchette – cette « norme » correspondant au degré de littéralité habituel de la plupart des traductions – mais elle n'est pas totalement déterminée par la totalité des distributions lexicales du corpus. Peut-être cette indétermination constitue-t-elle la limite

des méthodes formelles, au-delà de laquelle elles ne peuvent plus progresser, faute d'information : mais cette limite peut toujours être repoussée avec l'augmentation de la taille des corpus – car plus les régularités deviennent nettes, dans l'appariement des unités avec leurs équivalents, plus il est facile d'en écrémer les unités résiduelles.

Enfin, nos dernières expérimentations ont confirmé, s'il en était besoin, le bien fondé du principe circulaire, déjà mis en œuvre par les plus anciennes méthodes, de la consolidation réciproque des appariements phrastiques et subphrastiques : alignement au niveau des phrases et correspondances lexicales se renforcent mutuellement, car ils reflètent tous deux la même propriété de parallélisme. Les bons résultats (avec *F* voisine de 90 %) finalement obtenus sur le corpus Verne, pourtant réputé difficile, démontrent que les distributions lexicales constituent, et de loin, la source d'information la plus riche pour apparier les phrases : or cette information est aussi la plus générale, car son exploitation ne dépend pas du couple de langues envisagées (à la différence de la cognation).

Les indices tels que les cognats ou les transfuges n'en perdent pas leur intérêt, car ils permettent dans de très nombreux cas une économie substantielle de calcul. Mais le parallélisme est avant tout la résultante d'un réseau de correspondances potentielles qui se manifestent à chacun des paliers de la traduction, du texte à l'unité lexicale, en passant par les sections, les paragraphes et les phrases : avec l'outil statistique seul, on peut dans une large mesure démêler cet entrelacs.

Conclusion

« En pratique, la détermination du sens linguistique et des frontières entre celui-ci et les sémantismes d'ordre extralinguistique est difficile à opérer. L'idéal des linguistes est évidemment de repousser le plus loin possible les limites de ce qu'il est possible d'atteindre en se situant au niveau du sens linguistique »

Catherine Fuchs, *La paraphrase*, 1981 : 59

Conclusion générale

Le bi-texte, ou plus généralement le multi-texte, est un objet nouveau pour les sciences du langage.

C'est un objet empirique : il constitue la manifestation la plus immédiate de l'activité traductionnelle. En ce sens c'est un lieu privilégié, où s'accumulent un grand nombre d'informations brutes, traces ou dépôts d'une activité communicative complexe aux multiples facettes.

Mais c'est aussi un objet *construit* : nous l'avons vu, l'appariement des unités équivalentes, qu'il s'agisse de paragraphes, de phrases ou même d'unités lexicales, nécessite un important effort de définition théorique. De nombreux problèmes en découlent, tant au plan du découpage des unités que de leur mise en relation. En outre, même lorsque des critères détaillés sont explicités, l'accord entre les annotateurs n'est jamais total : il reste une part d'interprétation subjective qui constitue encore un défi au travail de construction théorique, impliquant tous les aspects des phénomènes langagiers, morphologiques, syntaxiques, sémantiques, voire pragmatiques.

Entre donnée brute et artefact, le bi-texte affiche d'étonnantes propriétés structurales : l'hypothèse d'une relation bi-textuelle, établissant la correspondance entre les unités qui composent ses deux parties, la recherche d'un ordre dans la distribution de ces connexions, le postulat du parallélisme, toutes ces conjectures rencontrent une vérification expérimentale éclatante. Pour preuve : la possibilité d'extraire automatiquement, moyennant la mise en œuvre concrète de l'hypothèse du parallélisme, des structures bi-textuelles qui s'approchent avec une grande précision de celles « manuellement » construites, par la réflexion et l'interprétation des linguistes.

Cette docilité vis-à-vis du traitement automatique en fait un objet aux enjeux pratiques de plus en plus manifestes. En particulier, le bi-texte s'insère harmonieusement dans le paysage varié de l'aide à la traduction : il apparaît comme une pièce maîtresse de l'architecture des stations de travail pour les traducteurs, en tant que réserve d'exemples facilement accessibles et réutilisables dans d'autres contextes. La mémoire de traduction

ainsi constituée est complémentaire des ressources plus classiques comme les dictionnaires : elle a l'avantage de présenter une certaine souplesse dans la mise à jour, et de receler des informations *in vivo* et *in situ*, qui ne figurent pas toujours dans les ouvrages lexicographiques ou les bases de données terminologiques. En outre l'alignement de traductions permet l'application d'outils de vérification, signalant les anomalies flagrantes telles que les omissions, l'emploi de faux amis, l'inconsistance terminologique, etc. Par ailleurs, nous avons vu que les bi-textes peuvent se révéler utiles pour alimenter les paramétrages de certains systèmes de TA, qu'ils soient basés sur des principes statistiques ou qu'ils soient mixtes, c'est-à-dire impliquant la mise en œuvre de règles formelles de transformation.

Certes, il faut admettre que l'usage des bi-textes est encore trop neuf pour que ces applications soient arrivées à maturité : pour en tirer le meilleur parti, il faudrait étudier concrètement les besoins réels des traducteurs et concevoir l'accès aux informations dans une perspective ergonomique, afin que l'utilisateur puisse naviguer efficacement au milieu d'une masse de données colossale qui peut rapidement devenir rédhibitoire.

Au-delà de l'aide à la traduction, les corpus bi-textuels présentent un intérêt grandissant dans des domaines connexes :

- en lexicographie et en terminographie, précisément, où le bi-texte est le lieu par excellence de l'observation à partir de corpus. Les filtres statistiques, comme l'ont montré nos expérimentations, permettent en outre d'extraire facilement une information brute d'une grande pertinence vis-à-vis de phénomènes massifs difficiles à évaluer manuellement.

- dans des domaines variés du TAL comme la reconnaissance vocale, la reconnaissance de caractère, la recherche d'information, etc. les bi-textes permettant d'alimenter des modèles paramétriques basés sur des jeux de probabilités.

- pour la didactique des langues étrangères, où les bi-textes sont utilisés depuis longtemps sous la forme d'éditions bilingues, etc.

Nos expérimentations, faisant suite à de très nombreux travaux dans le domaine, ont montré que la constitution automatique et massive de corpus bi-textuel n'est plus un objectif théorique à long terme. Le niveau de performance atteint autorise déjà l'exploitation industrielle et l'existence de débouchés concrets.

Les méthodes obtenant les meilleurs résultats sont le fruit de principes généraux et très simples : l'utilisation combinée de toutes les informations disponibles et facilement exploitables ; et la mise en œuvre heuristique d'un principe de précision d'abord commandant de procéder par raffinements successifs en partant des informations les plus fiables pour n'utiliser qu'en dernier recours les données les plus aléatoires.

Il est même un peu décevant, pour un œil de linguiste, de constater que les informations les plus triviales et les plus superficielles sont parfois celles qui donnent les meilleurs résultats : la correspondance des chiffres et des noms propres, l'observation des longueurs des phrases ou de certains signes de ponctuation, l'existence de mots graphiquement ressemblants – l'accumulation de tous ces indices superficiels suffit bien souvent à obtenir un alignement au niveau des phrases avec une certaine précision. L'ordinateur effectue à sa manière les mêmes hypothèses qu'un humain qui, ne connaissant pas les langues mises en présence, se bornerait à aligner au niveau de la forme sans comprendre le fond.

Heureusement (pour le linguiste) un autre type d'observation constitue fréquemment un complément indispensable, tant pour l'alignement des phrases que pour la mise en correspondances d'unités subphrastiques : la distribution des unités à travers les deux textes. Il s'agit là encore d'un indice formel, mais qui révèle des mécanismes profonds, situés au plan du contenu. Un phénomène d'importance est à la base de ces observations : la *cooccurrence*. Comme nous l'avons souligné, le concept de cooccurrence recouvre des acceptions différentes :

– d'une part, on peut s'intéresser à la cooccurrence de deux unités dans une même langue, le long de l'axe syntagmatique : il s'agit alors de *cooccurrence monolingue*. Ce type de cooccurrence nécessite la définition d'une portée, ou d'une fenêtre, dont les dimensions peuvent être fixes ou variables, à l'intérieur de laquelle on estime que les unités cooccurrent.

– d'autre part, et ceci est spécifique aux corpus de traduction, on observe la cooccurrence des unités de part et d'autres des textes équivalents : il s'agit alors de *cooccurrence parallèle*. Là aussi, il faut définir une certaine zone, ou *aire de cooccurrence*, correspondant à un sous-ensemble du produit cartésien des deux textes. L'aire de cooccurrence peut correspondre à des segments textuels en relation d'équivalence traductionnelle avérée, ou bien seulement présumée. Ce type de cooccurrence vise bien entendu à caractériser l'équivalence traductionnelle au niveau des unités lexicales : on peut supposer que deux unités qui cooccurrent très fréquemment dans des phrases équivalentes sont elles-mêmes équivalentes. Par suite, les distributions de deux unités peuvent être simplement comparées sur la base de leurs occurrences, séparément dans chaque côté du bi-texte, et de leurs cooccurrences, ensemble dans les zones correspondantes : un simple filtrage statistique permet alors d'évaluer si ces cooccurrences ont une pertinence, ou bien semblent découler de coïncidences fortuites. Cette comparaison est hors de portée de l'humain, qui peut difficilement établir le compte des occurrences et des cooccurrences pour des millions d'unités, tandis que ce genre de tâche est réalisable par n'importe quel ordinateur personnel récent.

L'observation combinée de tous ces indices, insérée dans des algorithmes à la complexité presque linéaire (souvent en $O(n\ log(n))$) donne des résultats satisfaisants. Pourtant, et nous espérons que notre travail empirique permet d'en entrevoir l'ampleur, le champ d'investigation reste encore immense, et de très nombreuses questions demeurent en suspens.

D'abord, les expérimentations présentées dans ce travail demandent à être étendues à d'autres corpus et à d'autres couples de langue. Certes, de nombreuses langues ont déjà été impliquées (entre autres l'anglais, le français, le chinois, le japonais, l'hébreu, le grec, l'allemand, le suédois), montrant une certaine généralité des méthodes étudiées (l'indice le plus universel étant la distribution des unités lexicales). Mais comme nous l'avons montré, la question du paramétrage des modèles mis en œuvre est centrale, le succès d'une technique dépendant largement de ce type de réglage. Or les paramètres sont nombreux :

moyenne et variance du rapport des longueurs des phrases correspondantes, probabilités de transition, probabilités liées à la cognation, largeur de l'espace de recherche centré sur la diagonale, seuil de déviation autorisé entre deux points d'ancrage, pondérations liées à la combinaison de différents indices, seuils de filtrages des correspondances lexicales, etc. Ce point a déjà été soulevé lors de discussions sur la liste de diffusion du projet ARCADE : si l'humain doit intervenir au niveau de chacun de ces réglages à chaque fois qu'un nouveau corpus se présente, on ne peut plus parler d'extraction totalement automatisée.

Pour minimiser cet interventionnisme et mieux maîtriser les phénomènes mis en jeu, il est essentiel de corréler les paramétrages optimaux avec les caractéristiques formelles des corpus : nous pensons qu'un certain nombre d'indices sont déjà inscrits dans le corpus parallèle, avant d'en extraire un alignement. Nous avons montré que certaines caractéristiques formelles étaient décisives : densité de transfuge et densité de cognats, longueur des textes, entropie conditionnelle, redondance. D'autres traits méritent l'attention : l'entropie au niveau monolingue, la richesse du vocabulaire, etc. Reste à réaliser une boucle de rétroaction, à étudier des méthodes permettant d'affiner le paramétrage des techniques, en fonction des traits textuels disponibles avant ou pendant la mise en œuvre de celles-ci.

De plus, l'étude approfondie de bi-textes déjà construits, pris comme référence, peut encore révéler des propriétés intéressantes. De telles observations conduiront peut-être à la découverte de nouveaux phénomènes, dont on pourra tirer parti pour l'extraction automatique. Par exemple, la mise en évidence de schémas de commutation, tels que ceux étudiés par Malavazos *et al*. (2000, cf. p. 204), permet d'obtenir des correspondances lexicales sans passer par les statistiques d'occurrences et de cooccurrences : mais, avant de les exploiter, il serait intéressant de déterminer empiriquement la fréquence de ce genre de schéma, et leur fiabilité.

Pour réaliser ce genre d'étude, il faut disposer de bi-textes établis et manuellement vérifiés, ce qui soulève de nombreux problèmes pratiques et théoriques : coût prohibitif, format d'encodage, critères mis en œuvre, accord intersubjectif. On l'a vu, la définition linguistique de l'alignement, ou des correspondances lexicales, est épineuse. D'une part, la délimitation des unités (phrases, unités lexicales) est problématique : qu'on se situe sur un plan strictement intralinguistique, dans la définition des lexies polylexicales, ou sur un plan contrastif, nous avons montré que l'indépendance des unités est un phénomène scalaire

s'inscrivant dans un continuum. D'autre part, le niveau d'équivalence retenu est soumis à des fluctuations difficilement tenables. Comme le note Fuchs (1981) à propos du « jugement de paraphrase », le « jugement » d'équivalence est indissociable de mécanismes interprétatifs impliquant la subjectivité d'un locuteur. Pour assurer une consistance linguistique à ce jugement, il faudrait se situer au niveau d'une sémantique linguistique vérifiable expérimentalement. La route est encore longue avant d'y parvenir. En attendant des caractérisations rigoureuses et complètement explicites, il faut se contenter de définitions opératoires visant à simplifier les problèmes, en s'attachant à donner une réponse aux problèmes les plus courants : de toute façon, un corpus de référence ne constitue jamais un standard absolu, mais un étalon de mesure servant à la comparaison.

Au-delà de notre étude, les techniques abordées ont de nombreux prolongements. Comme le suggère Simard (in Véronis, 2000 §3), on peut attendre beaucoup des corpus multilingues impliquant plus de deux langues. La variété des traductions, la somme de leurs divergences et de leurs convergences, apporte un surcroît d'information utilisable dans la constitution des bi-textes deux à deux.

Par ailleurs, nous n'avons pas traité le problème de l'identification automatique des unités polylexicales. Dans certaines approches, comme chez Melamed (1998a), la segmentation et l'appariement s'appuient l'un sur l'autre. Celui-ci fait l'hypothèse que l'identification, comme unité, des segments dont la traduction n'est pas compositionnelle doit aboutir à un renforcement des régularités des correspondances bi-textuelles. En retenant comme candidat les unités dont la fusion entraîne une augmentation de l'information mutuelle générale d'un texte sur l'autre, l'auteur obtient des résultats encourageants. Il serait intéressant d'effectuer le même genre d'expérience en se basant sur l'entropie conditionnelle.

On pourrait aussi nous reprocher d'avoir mis de côté, tout le long de notre étude, les méthodes classiques de l'ingénierie linguistique : nous n'avons utilisé ni dictionnaire de transfert, ni étiqueteur morphologique, ni grammaire formelle permettant d'effectuer des analyses syntaxiques superficielles ou profondes. Faut-il en déduire que ces outils doivent être proscrits dans la constitution et l'utilisation des corpus bi-textuels ? Bien évidemment non. Au contraire, nous pensons qu'il y a beaucoup à attendre d'un couplage des approches

statistiques et linguistiques : l'identification des cognats, par exemple, pourrait bénéficier de systèmes de conversion graphémiques adaptés pour les langues apparentées ; la plupart des méthodes présentées tireraient un parti avantageux de dictionnaires de transfert plus ou moins complets contenant les équivalences pour le vocabulaire courant ; la lemmatisation des unités, on l'a vu, est très utile pour dégager certaines régularités ; lors de la constitution d'une mémoire de traduction, les exemples de traduction peuvent être codés sous la forme d'arbre étiquetés (Sumita & Iida, 1992 ; Watanabe, 1992), dont la comparaison permet ensuite d'effectuer des généralisations, comme chez Malavazos *et al*. (2000) ; enfin l'observation statistique de phénomènes de surface permet parfois de restreindre les arbres de choix lors de l'application de systèmes de règles : les statistiques peuvent aussi fournir un complément heuristique aux grammaires formelles. Si nous n'avons pas développé ces idées, c'est parce que l'économie de notre de sujet nous l'imposait, avec la conviction que les méthodes statistiques « nues » offrent un champ d'exploration encore très ouvert.

Signalons enfin une autre piste intéressante, que nous n'avons pas empruntée : l'étude des corpus comparables (non parallèles). D'après Fung (in Véronis, 2000 §11) on peut caractériser les corpus comparables sur la base du domaine et de la période : « 1/ pour les mêmes sujets, les mots ont des contextes comparables d'une langue à l'autre, 2/ les mots d'un même domaine et d'une même période temporelle ont des usages comparables »[224] Tirant parti de ce constat, Fung expose une méthode, nommée *Convec*, permettant de déterminer des équivalents traductionnels à travers des textes non-parallèles anglais / chinois, en se basant sur la similarité des contextes. Cette méthode nécessite le recours à un dictionnaire bilingue, qu'il s'agit de compléter automatiquement. Tous les mots inconnus du dictionnaire anglais / chinois sont caractérisés en fonction de leurs co-textes immédiats (on ne tient compte que des co-textes qui mettent en jeu des mots connus du dictionnaire) : ces constellations contextuelles sont représentées sous la forme de vecteurs représentant les fréquences relatives (par rapport à la taille du corpus) des cooccurrences avec les mots du vocabulaire. La comparaison des vecteurs d'une langue à l'autre se fait grâce au dictionnaire de transfert, qui donne les correspondances entre les coordonnées des vecteurs (les dimensions des vecteurs contextuels correspondent aux mots

[224] "1. For the same topic, words have comparable contexts across languages 2. Words in the same domain and the same time period have comparable usage patterns"

connus du dictionnaire). Le score de similarité utilisé pour la comparaison des vecteurs est inspiré des techniques de recherche d'information (mesure du cosinus). Dans son évaluation, Fung montre que cette méthode permet de trouver, dans son corpus, la traduction correcte de la plupart des mots inconnus du dictionnaire.

Ces développements montrent que la notion de parallélisme peut être étendue à un niveau local, sur le plan des relations conceptuelles[225] qu'entretiennent des unités lexicales, pourvu qu'on se situe dans des domaines similaires. Là encore, c'est l'observation des cooccurrences, combinée à des méthodes de filtrage statistique, qui permet la comparaison entre les langues : il y a donc une certaine continuité avec les techniques mises en œuvre sur les corpus parallèles. La seule différence tient dans la nature des cooccurrences : il ne s'agit plus de cooccurrences parallèles, mais de cooccurrences monolingues. L'aspect paradigmatique des cooccurrences parallèles, c'est-à-dire le fait que les équivalents traductionnels commutent ensemble dans des contextes différents, est remplacé par une caractérisation syntagmatique : cette fois les équivalents traductionnels apparaissent à des endroits différents dans des contextes similaires.

Cette nouvelle approche suggère la mise en œuvre d'une méthode combinée : dans l'extraction des correspondances lexicales à partir d'un corpus bi-textuel, les deux angles d'attaque peuvent être pris en compte simultanément. En d'autres termes, il serait intéressant, pour comparer les unités des deux textes, de confronter à la fois leurs cooccurrences monolingues et leurs cooccurrences parallèles : on peut s'attendre à un renforcement mutuel des deux sources d'information. Cette piste reste encore à explorer.

On aura peut-être aperçu une contradiction profonde dans notre approche de la traduction :

– d'une part nous n'avons cessé d'affirmer que la traduction n'est pas un système de transformations appliqué directement de langue à langue : la traduction des *messages* implique nécessairement une plongée dans l'extralinguistique où se

[225] Il faut noter que les contextes pertinents doivent inclure des mots pleins, les cooccurrences avec les mots outils ne présentant pas une information assez discriminante.

construit la recherche d'équivalence, à travers des contraintes pragmatiques, fonctionnelles, culturelles, encyclopédiques, conceptuelles, etc., la strate linguistique fournissant en dernier lieu un système de contraintes propres largement surdéterminé par la constellation des données interprétatives. Nous ne nions pas que certaines traductions peuvent se situer au seul niveau du transcodage : c'est là tout l'objet de la Traduction automatique depuis ces origines. Seulement, la réalité empirique de la traduction ne peut être réduite à ce cas particulier.

– d'autre part, nous appliquons des techniques fondées sur la recherche, au niveau le plus superficiel, des similarités entre un texte et sa traduction, en présumant l'existence d'un réseau serré de correspondances au niveau de ses unités, voire de ses mots. Et nous affirmons que cette propriété structurale des textes en relation de traduction, que nous nommons *parallélisme*, est assez largement répandue parmi les textes issus de la pratique concrète de la traduction en tant qu'activité de communication.

En somme, quand nous stigmatisons une vision simpliste de la traduction (comme traduction des mots), on pourrait nous reprocher de lui fermer la porte au nez pour la faire rentrer discrètement par la fenêtre.

Cette contradiction n'est qu'apparente : elle est déjà sous-jacente à la fonction ancillaire de la traduction, qui ne peut remplir sa mission que si le traducteur conquiert une certaine liberté. Israël (in Lederer & Israël, 1991 : 27) cite une réflexion de André Gide :

> « Dans les premiers temps, je demandais que les traductions de mes œuvres me fussent soumises, et celle-ci me paraissait la meilleure qui suivait de plus près le texte français ; j'ai vite reconnu mon erreur et, à présent, je recommande à mes traducteurs de ne jamais se croire esclaves de mes mots, de ma phrase, de ne pas rester trop penchés sur leur travail... Mais, encore une fois, ce conseil n'est bon que si le traducteur connaît admirablement les ressources de sa propre langue et qu'il est capable de pénétrer l'esprit et la sensibilité de l'auteur qu'il entreprend de traduire jusqu'à s'identifier à lui. »[226]

[226] A. Gide (1931) *Les Essais*, t. 3, Gallimard, Paris, p. 196

Bien entendu, l'équivalence est échafaudée dans et par la langue d'arrivée : mais aucune nécessité ne guide cette construction. Il y a *détermination linguistique* mais pas *déterminisme linguistique*. C'est pourquoi l'observation d'un corpus de traduction permet de faire émerger des *régularités*, et non des *règles*. Pour capter c'est régularités, l'outil statistique est particulièrement adapté, puisqu'il confère la distance nécessaire à l'apparition de phénomènes généraux non pertinents sur le plan local. En s'éloignant des variations liées à la singularité des phénomènes traductionnels qui déterminent les choix locaux, l'observation statistique fait apparaître les contraintes générales qui régissent le passage d'un code à un autre : derrière le « bruit » de la liberté en traduction, la loi des grands nombres permet de dessiner, avec une netteté croissante à mesure que les corpus prennent de l'ampleur, les lignes de force du transcodage.

Par l'observation empirique, on fait ainsi apparaître l'espace du contrastif : à des unités lexicales de la langue source, dans un certain contexte, correspondent d'autres unités dans la langue cible. De manière plus générale, on peut dégager la correspondance d'*unités de traduction*, au sens où nous les avons définies, cristallisant des préférences et des habitudes idiomatiques divergentes dans les deux langues. Mais rien n'oblige à s'arrêter au niveau lexical : pourquoi ne pas étudier de même la correspondance des phénomènes morphosyntaxiques ? Si le bi-texte permet de mettre en regard des structures complexes, la détection de régularités peut concerner tout phénomène : diathèse, temps et modes verbaux, parties du discours, fonctions grammaticales, cas, etc. Le fait que les correspondances ne soient pas biunivoques révèle sans doute la complexité du niveau contrastif : mais cette complexité n'est pas synonyme de chaos, sinon en apparence, et l'on sera toujours récompensé d'y rechercher un ordre sous-jacent. Les irrégularités mêmes peuvent être riches d'enseignement, comme le suggère Santos (in Véronis, 2000 §8), et ce qui peut apparaître localement comme du bruit est susceptible de révéler, à un niveau général, des divergences profondes entre les codes :

> « Plutôt que de rechercher des règles fiables ou des correspondances, de considérer les données qui s'éloignent de ces normes comme du bruit résiduel (Church & Gale, 1991), ou encore de rejeter la création stylistique hors du champ de la sémantique (Dyvi, 1998), je pense que toutes les paires de traductions – incluant les simples erreurs et les réécritures complètes – mettent en lumière les systèmes des deux langues. En effet, le plus souvent, les erreurs de traduction

> sont liées aux difficultés mêmes qui découlent des différences entre les langues (…) »[227]

Ces phénomènes qui apparaissent au niveau macroscopique ont une autre vertu : ils permettent d'examiner, par le détour d'un autre système, les rapports de la langue au monde. Comme le dit A. Greimas, la traduction est déjà une amorce d'explicitation du sens. Le texte traduit, par rapport au texte source, présente l'embryon d'un métalangage décrivant son sens : par exemple, la variété des équivalents correspondant à une même unité est une manifestation observable de sa polysémie – et l'usage des bi-textes constitue un détour précieux pour les tâches de désambiguïsation. Lorsque la constellation de ces correspondances est confirmée par des millions d'occurrences, on voit apparaître une configuration stable d'acceptions (les équivalents les plus fréquents) esquissant la structure d'une description de son signifié.

Là où le structuralisme échouait à faire sortir la langue d'une certaine forme d'autarcie, l'observation massive des phénomènes de correspondance permet de généraliser le test de commutation aux transformations d'une langue dans une autre. Lorsqu'on remarque que deux unités cooccurrent (parallèlement) dans des contextes différents, on ne fait rien d'autre que de commuter leur contexte pour examiner leur constance sémantique[228] : seulement, plutôt que de se fier au jugement du linguiste pour évaluer la constance ou la différence sémantique, on se base sur les centaines de traducteurs qui ont estimé que les deux unités étaient susceptibles de jouer des rôles équivalents. Les structures qui en résultent sont objectives, car issues d'un consensus intersubjectif.

Pour établir une sémantique, le métalangage constitué par la langue cible ne suffit pas, car il souffre de la même opacité et des mêmes ambiguïtés que la langue source : c'est une langue naturelle. La donnée des équivalents potentiels d'une même unité reste marquée du sceau de l'implicite, mais rejeté un peu plus loin : ce n'est pas encore les rapports de la langue au monde qui apparaissent, ce sont les rapports de signifiés à d'autres signifiés. Mais c'est déjà un pas.

---

[227] cf. note 173

[228] dans le test original, on procède de manière inverse : on observe les différences sémantiques engendrées par la commutation de deux unités dans un même contexte.

Les correspondances bilingues permettent de faire émerger des différences dans le rapport de deux systèmes. Pour observer des différences dans le rapport du linguistique et de l'extralinguistique, c'est-à-dire pour construire une véritable sémantique référentielle, on pourrait peut être tirer parti de corpus multilingues impliquant plus de deux langues, et des langues éloignées. Imaginons que les correspondances lexicales n'impliquent pas des couples, mais des séries d'unités : le couple (*I have noted, j'ai noté*) est ambigu (p. ex. on peut hésiter entre « j'ai couché par écrit » ou « j'ai remarqué ») ; mais si l'on rajoute une correspondance avec l'italien *ho notato*, l'ambiguïté est levée (à la faveur de l'acception « j'ai remarqué »). L'addition de nouvelles correspondances permettrait de converger vers un sens référentiel précis. L'observation massive de telles correspondances aboutirait vraisemblablement à la mise en évidence de configurations stables liées à des référents ou concepts précis : dès lors, on disposerait de fondations solides pour étiqueter les différents sens liés aux unités lexicales de chaque langue. Sur ces bases métalinguistiques, on pourrait reconstruire les significations des unités, relier les usages à leurs différents substrats référentiels, classer les acceptions, différencier les polysèmes et déterminer les points de contact entre synonymes.

Pour Walter Benjamin (in Nergaard, 1993 : 227), toutes les langues convergent sous la force d'une affinité profonde, car « dans chacune d'elles, prise comme un tout, est entendue une seule et même chose, qui toutefois n'est accessible à aucune d'elles prise dans sa singularité, mais seulement à la totalité de leurs intentions réciproquement complémentaires : la langue pure »[229]. Cette *Reine Sprache* est une conjecture – et il ne nous est pas besoin d'y recourir pour supposer un autre type de convergence : les différentes traductions d'un même message convergent toutes vers un même *au-delà* des langues. La forme de cet invariant, qui se dessine en négatif dans les fluctuations des multi-textes, en révèle autant sur chaque langue que sur le sens particulier qui y est enclos : cette forme là reste à étudier.

[229] Dans la traduction italienne : "Piuttosto, ogni affinità metastorica delle lingue si basa sul fatto che in ciascuna di esse, presa come un tutto, è intesa una sola e medesima cosa, che tuttavia non è accessibile a nessuna di esse presa singolarmente, ma solo alla totalità delle loro intenzioni reciprocamente complementari : la pura lingua."

Annexe

A-I Corpus de travail

Le corpus BAF

Le corpus BAF (ou Bi-texte Anglais Français) a été élaboré au CITI, un laboratoire public canadien, dans le cadre de l'ARC, ou « Action de recherche concertée », projet coopératif financé par l'AUPELF-UREF (Simard, 1997 :493). Le corpus mis en oeuvre dans nos recherches, issu du projet ARCADE, est composé de textes que l'on peut ranger en 4 grandes catégories :

– *Textes institutionnels*

D'après Véronis & Langlais (in Véronis, 2000 §19) « c'est probablement un des genres pour lesquels la demande en textes alignés, pour des tâches pratiques telles que la traduction, est la plus forte, tout spécialement dans l'Union Européenne et au Canada, du fait du contexte multilingue. C'est un type de texte traditionnellement ''facile'' à aligner au niveau des phrases, parce que les traducteurs ne prennent pas de risques dans la traduction et restent très proches du texte original. »[230]

Au sein de cette catégorie, ces auteurs distinguent les sous-genres suivants :

- *les traductions directes.* Il s'agit de documents issus de la Cour Suprême du Canada (corpus Cour) et de rapports des nations unies (corpus Onu).
- *les traductions indirectes.* Celles-ci concernent des traductions différentes issues d'un même texte original, comme il est fréquent entre les nombreuses langues de l'union européenne. Un texte se range dans cette catégorie : un rapport de l'Organisation Internationale du Travail (corpus Ilo).

[230] "This is the type of data which is the easiest to gather in large quantities: it consists in texts such as parliamentary debates, official reports, legal documents, etc. It is probably one of the genres for which the demand of aligned texts for practical tasks such as translation is the greatest, especially in the European Union and Canada, due to the multilingual situation. It is a type of text traditionally "easy" to align at sentence level, because the translators take no risk in the translation and stay very close to the original text."

- *les transcription de discours oraux*, telles que le corpus Hansard (corpus Hans), enregistrant des débats du parlement canadien.

– *Manuels techniques*

Comme pour les textes institutionnels, la traduction de manuel technique constitue un domaine favorable à la mise en œuvre des ressources bi-textuelles. Même si ce type de traduction est généralement très proche de la littéralité, il présente des caractéristiques différentes du précédent, notamment par la quantité et l'importance des terminologies techniques, et par la présence d'éléments spécifiques en rupture avec la linéarité du discours : les tables, les figures, les glossaires, les listes, etc. Le texte choisi est un guide d'utilisation de logiciel de la société Xerox.

– *Articles scientifiques*

Selon Véronis & Langlais (in Véronis, 2000 §19), ces textes sont similaires aux manuels techniques, mais d'une prose plus linéaire. En outre, les traducteurs sont susceptibles de prendre plus de liberté par rapport à l'original, dans le but de respecter certains aspects rhétoriques de l'argumentation de l'auteur. Cinq articles ont été choisis (corpus CITI1, CITI2, TAO1, TAO2, TAO3).

– *Corpus littéraire*

Dans ce type de texte, la traduction est totalement linéaire. En revanche, le traducteur est susceptible de prendre plus de distance par rapport à la littéralité du texte originale. Dans le texte choisi, une traduction anglaise de *De la terre à la lune* de Jules Verne, des passages ont été contractés voire supprimés, comme quelques dialogues et certaines longues descriptions. Le parallélisme n'étant plus strict, ce corpus présente des difficultés supplémentaires pour la tâche d'alignement.

Les composantes du Corpus BAF sont détaillées dans le tableau 1 ci dessous.

tableau 83 : références et sources des constituants du corpus BAF

Institutionnel	Hans	Hansard – Canadian Parliamentary Proceedings. March 14, 1994	House of Commons Publication service (Canada)
	Cour	Supreme court of Canada (1995). *Terrence Wayne Burlingham v. Her Majesty the Queen*	*Centre de recherche en droit public* de la Faculté de droit de l'Université de Montréal
	Ilo	UN International Labor Organization (1985). *241st and 242nd Reports of the Commitee on Freedom of Association*	*ECI Multilingual Corpus*
	Onu	UN (1993). *Report of the Secretary-General on the Work of the Organization*	UN translation services
Technique	Xerox	Xerox Corporation, *ScanWorX User's Guide*	*ECI Multilingual Corpus*. Note : ces documents contiennent en appendice un glossaire relativement large, qui ne pouvait être aligné, à cause des différences dans l'ordre des entrées.
Littéraire	Verne	Verne, Jules. *De la terre à la lune*	La version originale française a été obtenue sur le site Web de l'association des Bibliophiles Universels. La traduction anglais provient du projet Gütenberg.

tableau 83 (suite)

Genre	*Nom*	*Références*	*Sources*
	CIT1	Geoffroy, Catherine (1994). *Les technologies de communication de l'information de l'information et les aîné(e)s.* CITI technical report.	Rapports techniques du CITI
	CIT1	Lapointe, François (1995). *Changement technologique et organisation du travail.* CITI technical report.	
Scientifique	TAO1	L'analyse de traduction et l'automatisation de la traduction.	
	TAO2	Macklovitch, Elliott (1995), *Peut-on vérifier automatiquement la cohérence terminologique* ? in Actes des 4èmes journées scientifiques, Lexicommatiques et Dictionnairiques, Lyon, France	
	TAO3	Simard, Michel (1995), *Réaccentuation automatique de textes français,* CITI technical Report.	

Le corpus JOC

Le corpus JOC est constitué de questions écrites posées par des membres du Parlement européen, suivie des réponses données par la Commission. Ces questions concernent des sujets très variés : agriculture, économie, environnement, institutions, droits de l'homme, transports, etc. Elles ont été publiées en 1993, dans les Séries C du Journal officiel de la Communauté européenne, et collectées dans le cadre du projet MLCC-MULTEXT.

A-II Préalignement avec les transfuges

Paramétrages :

Rapport des longueurs des textes (français / anglais) : $R_{diag} = m/n = 1,203$
Critère de diagonalité : $d(I,J)=|I/n+J/m| < Seuil_{diag}=0,7$
Critère de continuité entre deux points (I_p,J_p) et (I,J) :
$D_I = (I-I_p)*R_{diag}$ $D_J = (J-J_p)$ $D_{min} = min(D_I,D_J)$ $D_{max}=max(D_I,D_J)$ $Dév=D_{max}/D_{min}$,
contraintes : si $D_{min} = 1$ alors $Dév < 4$, si $2 \leq D_{min} < 5$ alors $Dév < 3,5$, si $5 \leq D_{min} < 10$ alors $Dév < 3$, si $10 \leq D_{min} < 20$ alors $Dév < 2,5$, si $20 \leq D_{min}$ alors $Dév < 2$

tableau 84 : Résultats du préalignement en utilisant indépendamment puis successivement différents types de transfuges

Corpus	*Type de transfuge*	*Application indépendante des différents type*			*Application successive des différents type*		
		P	***R***	***F***	***P***	***R***	***F***
Cour	Alphanumériques	100,0 %	11,3 %	20,3 %	100,0 %	11,3 %	20,3 %
	Majuscules	100,0 %	22,7 %	36,9 %	100,0 %	24,2 %	39,0 %
	Transfuges	100,0 %	68,3 %	81,1 %	99,9 %	68,8 %	81,5 %
	Transfuges*	98,2 %	79,1 %	87,6 %	99,5 %	80,3 %	88,9 %
Hans	Alphanumériques	100,0 %	1,9 %	3,8 %	100,0 %	1,9 %	3,8 %
	Majuscules	100,0 %	10,3 %	18,6 %	99,2 %	14,0 %	24,5 %
	Transfuges	99,9 %	25,6 %	40,8 %	99,6 %	33,4 %	50,0 %
	Transfuges*	97,9 %	51,9 %	67,8 %	98,2 %	51,9 %	67,9 %
Ilo	Alphanumériques	100,0 %	16,7 %	28,6 %	100,0 %	16,7 %	28,6 %
	Majuscules	100,0 %	34,8 %	51,7 %	100,0 %	46,0 %	63,0 %
	Transfuges	99,9 %	60,9 %	75,7 %	99,8 %	66,2 %	79,6 %
	Transfuges*	99,0 %	74,2 %	84,8 %	99,6 %	75,3 %	85,7 %
Onu	Alphanumériques	100,0 %	9,9 %	18,1 %	100,0 %	9,9 %	18,1 %
	Majuscules	100,0 %	28,0 %	43,8 %	100,0 %	38,0 %	55,0 %
	Transfuges	99,9 %	57,1 %	72,7 %	100,0 %	66,6 %	80,0 %
	Transfuges*	98,6 %	78,0 %	87,1 %	99,3 %	78,7 %	87,8 %
Tao1	Alphanumériques	100,0 %	2,3 %	4,4 %	100,0 %	2,3 %	4,4 %
	Majuscules	100,0 %	15,8 %	27,3 %	100,0 %	17,9 %	30,3 %
	Transfuges	99,9 %	68,7 %	81,4 %	100,0 %	71,6 %	83,5 %
	Transfuges*	99,8 %	88,9 %	94,1 %	99,7 %	87,0 %	92,9 %
Tao2	Alphanumériques	100,0 %	3,0 %	5,7 %	100,0 %	3,0 %	5,7 %
	Majuscules	100,0 %	10,6 %	19,2 %	100,0 %	11,8 %	21,1 %
	Transfuges	100,0 %	72,8 %	84,3 %	100,0 %	75,3 %	85,9 %
	Transfuges*	98,1 %	91,0 %	94,4 %	99,1 %	88,9 %	93,7 %
Tao3	Alphanumériques	100,0 %	0,4 %	0,8 %	100,0 %	0,4 %	0,8 %
	Majuscules	100,0 %	7,8 %	14,4 %	95,3 %	7,1 %	13,2 %
	Transfuges	97,8 %	46,8 %	63,3 %	97,9 %	48,7 %	65,0 %
	Transfuges*	98,4 %	70,8 %	82,3 %	99,1 %	69,7 %	81,8 %

tableau 84 (suite)

Citi1	Alphanumériques	99,2 %	1,7 %	3,4 %	99,2 %	1,7 %	3,4 %
	Majuscules	97,1 %	8,1 %	15,0 %	97,7 %	8,4 %	15,5 %
	Transfuges	98,0 %	45,4 %	62,1 %	97,8 %	47,8 %	64,2 %
	Transfuges*	92,8 %	73,5 %	82,0 %	90,7 %	72,9 %	80,8 %
Citi2	Alphanumériques	100,0 %	4,1 %	7,9 %	100,0 %	4,1 %	7,9 %
	Majuscules	100,0 %	25,1 %	40,2 %	100,0 %	28,2 %	44,0 %
	Transfuges	99,8 %	55,7 %	71,5 %	99,8 %	59,5 %	74,6 %
	Transfuges*	98,7 %	76,2 %	86,0 %	98,9 %	76,0 %	85,9 %
Verne	Alphanumériques	100,0 %	0,1 %	0,2 %	100,0 %	0,1 %	0,2 %
	Majuscules	100,0 %	8,1 %	15,0 %	100,0 %	8,2 %	15,2 %
	Transfuges	100,0 %	28,3 %	44,1 %	100,0 %	33,2 %	49,9 %
	Transfuges*	94,9 %	40,6 %	56,9 %	93,4 %	39,8 %	55,8 %
Xerox	Alphanumériques	100,0 %	0,02 %	0,04 %	100,0 %	0,0 %	0,0 %
	Majuscules	99,66 %	0,34 %	0,67 %	100,0 %	0,4 %	0,8 %
	Transfuges	99,83 %	2,38 %	4,64 %	99,8 %	2,5 %	5,0 %
	Transfuges*	97,04 %	3,13 %	6,06 %	98,7 %	3,3 %	6,4 %
Moyenne BAF	Alphanumériques	99,9 %	4,7 %	8,5 %	99,9 %	4,7 %	8,5 %
	Majuscules	99,7 %	15,6 %	25,7 %	99,3 %	18,6 %	29,2 %
	Transfuges	99,6 %	48,4 %	62,0 %	99,5 %	52,2 %	65,4 %
	Transfuges*	97,6 %	66,1 %	75,4 %	97,8 %	65,8 %	75,3 %
Moyenne BAF *	Alphanumériques	99,9 %	5,1 %	9,3 %	99,9 %	5,1 %	9,3 %
	Majuscules	99,7 %	17,1 %	28,2 %	99,2 %	20,4 %	32,1 %
	Transfuges	99,5 %	53,0 %	67,7 %	99,5 %	57,1 %	71,4 %
	Transfuges*	97,6 %	72,4 %	82,3 %	97,8 %	72,0 %	82,1 %

Transfuges* = prise en compte des transfuges quelconques, sans appliquer la condition de surdétermination des points.
BAF * = corpus BAF hors Xerox.

tableau 85 : corrélations entre les résultats et la densité de transfuge, pour les différents types de transfuge (alphanumériques, majuscules, transfuges quelconques).

		NbPhrases	*NbAnc*	*DPhrase*	*P*	*R*	*F*
Cour	Alphanumériques	1429	467	0,33	100,0 %	11,3 %	20,3 %
	Majuscules	1429	1013	0,71	100,0 %	22,7 %	36,9 %
	Transfuges	1429	3251	2,28	100,0 %	68,3 %	81,1 %
Hans	Alphanumériques	3119	193	0,06	100,0 %	1,9 %	3,8 %
	Majuscules	3119	1426	0,46	100,0 %	10,3 %	18,6 %
	Transfuges	3119	3891	1,25	99,9 %	25,6 %	40,8 %
Ilo	Alphanumériques	7426	3622	0,49	100,0 %	16,7 %	28,6 %
	Majuscules	7426	3831	0,52	100,0 %	34,8 %	51,7 %
	Transfuges	7426	15207	2,05	99,9 %	60,9 %	75,7 %
Onu	Alphanumériques	2619	1105	0,42	100,0 %	9,9 %	18,1 %
	Majuscules	2619	1230	0,47	100,0 %	28,0 %	43,8 %
	Transfuges	2619	5173	1,98	99,9 %	57,1 %	72,7 %
Tao1	Alphanumériques	371	55	0,15	100,0 %	2,3 %	4,4 %
	Majuscules	371	371	1,00	100,0 %	15,8 %	27,3 %
	Transfuges	371	987	2,66	99,9 %	68,7 %	81,4 %
Tao2	Alphanumériques	320	34	0,11	100,0 %	3,0 %	5,7 %
	Majuscules	320	237	0,74	100,0 %	10,6 %	19,2 %
	Transfuges	320	852	2,67	100,0 %	72,8 %	84,3 %
Tao3	Alphanumériques	192	14	0,07	100,0 %	0,4 %	0,8 %
	Majuscules	192	105	0,55	100,0 %	7,8 %	14,4 %
	Transfuges	192	379	1,98	97,8 %	46,8 %	63,3 %
Citi1	Alphanumériques	649	116	0,18	99,2 %	1,7 %	3,4 %
	Majuscules	649	459	0,71	97,1 %	8,1 %	15,0 %
	Transfuges	649	1372	2,11	98,0 %	45,4 %	62,1 %
Citi2	Alphanumériques	1595	482	0,30	100,0 %	4,1 %	7,9 %
	Majuscules	1595	1811	1,14	100,0 %	25,1 %	40,2 %
	Transfuges	1595	4024	2,52	99,8 %	55,7 %	71,5 %
Verne	Alphanumériques	3819	20	0,01	100,0 %	0,1 %	0,2 %
	Majuscules	3819	1273	0,33	100,0 %	8,1 %	15,0 %
	Transfuges	3819	3449	0,90	100,0 %	28,3 %	44,1 %

NbPhrases = moyenne du nombre de phrases en anglais et en français
NbAnc = moyenne des nombres d'ancrage identifiés en anglais et en français
DPhrase = moyenne des densités phrastiques d'ancrage en anglais et en français

A-III Préalignement avec les cognats

Paramétrages :

Distance minimale requise entre deux points d'ancrage (Id,Jd) et (If,Jf) issus de l'étape précédente : LargeurSection = min(If-Id,Jf-Jd), LargeurSection > 10
Critère de diagonalité dans une section encadrée par (Id,Jd) et (If,Jf)
Contrainte : d(I,J)=|(I-Id)/(If-Id)+(J-Jd)/(Jf-Jd)| < Seuildiag = 0,2
Critère de continuité entre deux points (Ip,Jp) et (I,J) :
*DI = (I-Ip)*Rdiag DJ = (J-Jp) Dmin = min(DI,DJ) Dmax=max(DI,DJ) Dév=Dmax/Dmin,*
Contraintes : si Dmin = 1 alors Dév < 4; si 2 ≤ Dmin < 5 alors Dév < 2,5; si 5 ≤ Dmin < 10 alors Dév < 2; si 10 ≤ Dmin < 20 alors Dév < 1,7; si 20 ≤ Dmin ≤ 50 alors Dév < 1,5; si 50 ≤ Dmin alors Dév < 1,3

tableau 86 : repérage des 3-grammes+ et des transfuges comme candidats cognats

	Modèle symétrique			***Modèle dissymétrique***		
	P	***R***	***F***	***P***	***R***	***F***
Cour	99,1 %	67,3 %	80,2 %	99,4 %	72,6 %	83,9 %
Hans	97,3 %	58,0 %	72,7 %	97,7 %	58,9 %	73,5 %
Ilo	98,2 %	66,2 %	79,1 %	99,6 %	73,2 %	84,4 %
Onu	99,2 %	75,7 %	85,9 %	99,7 %	84,2 %	91,3 %
Tao1	100,0 %	89,8 %	94,6 %	100,0 %	82,8 %	90,6 %
Tao2	99,6 %	86,8 %	92,7 %	99,9 %	82,5 %	90,4 %
Tao3	100,0 %	75,8 %	86,2 %	99,2 %	79,8 %	88,4 %
Citi1	96,2 %	60,7 %	74,4 %	98,9 %	74,3 %	84,9 %
Citi2	99,4 %	76,2 %	86,2 %	99,7 %	78,2 %	87,7 %
Verne	94,3 %	57,3 %	71,3 %	95,8 %	53,5 %	68,7 %
Xerox	99,1 %	2,9 %	5,7 %	99,1 %	2,8 %	5,4 %
Moyenne BAF	98,4 %	65,1 %	75,4 %	99,0 %	67,5 %	77,2 %
Moyenne BAF *	98,3 %	71,4 %	82,3 %	99,0 %	74,0 %	84,4 %

tableau 87 : résultats pour différents seuil du rapport r (longueur de la SCM rapporté à la longueur de la plus courte des formes comparées, cf. équation 30).

r_{seuil}	***Identification des cognats***			***Préalignement subséquent*** *Corpus Cour*		
	P_c	R_c	F_c	P_a	R_a	F_a
0,8	83,3 %	61,5 %	70,7 %	99,8 %	82,3 %	90,2 %
0,69	82,9 %	70,9 %	76,5 %	99,8 %	86,0 %	92,4 %
0,66	75,0 %	73,3 %	74,1 %	99,7 %	86,2 %	92,5 %
0,63	74,5 %	73,8 %	74,2 %	99,8 %	85,9 %	92,3 %
0,6	73,2 %	75,6 %	74,4 %	99,9 %	85,4 %	92,1 %
0,55	61,0 %	79,5 %	69,0 %	99,7 %	85,4 %	92,0 %
0,5	59,3 %	80,3 %	68,2 %	99,8 %	85,8 %	92,3 %
0,4	24,3 %	80,2 %	37,3 %	99,6 %	82,6 %	90,3 %

tableau 88 : résultats pour une combinaison d'indice mêlant n-grammes et SCM

Paramétrages :

Sont considérés comme candidats cognats :
- les transfuges numériques
- les transfuges quelconques de longueur supérieure ou égale à 3
- les couples de mots de longueur inférieure ou égale à 6 avec un 4-gramme commun.
- les couples intégrant les SCM de longueur supérieure ou égale à 4, avec r = 0,66.

	Modèle symétrique			***Modèle dissymétrique***		
	P	***R***	***F***	***P***	***R***	***F***
Cour	99,3 %	80,6 %	89,0 %	99,9 %	80,8 %	89,4 %
Hans	98,8 %	66,3 %	79,4 %	98,8 %	71,7 %	83,1 %
Ilo	99,9 %	85,7 %	92,3 %	100,0 %	87,9 %	93,5 %
Onu	99,9 %	89,5 %	94,4 %	99,8 %	93,2 %	96,4 %
Tao1	100,0 %	85,8 %	92,3 %	100,0 %	91,1 %	95,3 %
Tao2	100,0 %	85,3 %	92,1 %	100,0 %	92,8 %	96,2 %
Tao3	100,0 %	83,2 %	90,9 %	100,0 %	87,0 %	93,0 %
Citi1	98,9 %	80,5 %	88,8 %	99,0 %	82,2 %	89,8 %
Citi2	99,5 %	82,8 %	90,4 %	99,8 %	81,8 %	89,9 %
Verne	97,9 %	60,5 %	74,8 %	97,5 %	59,2 %	73,7 %
Xerox	99,0 %	2,6 %	5,0 %	99,1 %	2,7 %	5,3 %
Moyenne BAF	99,4 %	73,0 %	80,8 %	99,44 %	75,5 %	82,3 %
Moyenne BAF*	99,4 %	80,0 %	88,4 %	99,47 %	82,8 %	90,0 %

tableau 89 : résultats pour une combinaison d'indice mêlant n-grammes et SCM, avec une pondération des différents cas de figure.

Paramétrages :

Modèle symétrique. Sont considérés comme candidats cognats :
les transfuges de longueur supérieure ou égale à 2
les couples de mots de longueur inférieure ou égale à 6 avec un 4-gramme commun.
les couples intégrant les SCM de longueur supérieure ou égale à 4, avec r = 0,66.

Cas	*transfuges*	*4-gram$^+$ l<7*	*SCM 4*	*SCM 5*	*SCM 6*	*SCM 7*	*SCM 8*	*SCM 9*	*SCM ≥10*
Pondération 1	10	5	1	8	6	6	3	8	10
Pondération 2	10	6	2	6	7	8	8	9	10

	Pondération 1			***Pondération 2***		
	P	***R***	***F***	***P***	***R***	***F***
Cour	99,6 %	83,5 %	90,9 %	99,7 %	86,2 %	92,5 %
Hans	99,5 %	73,9 %	84,8 %	99,1 %	73,1 %	84,1 %
Ilo	99,9 %	88,8 %	94,0 %	100,0 %	87,8 %	93,5 %
Onu	99,9 %	92,7 %	96,1 %	99,8 %	93,2 %	96,4 %
Tao1	99,9 %	92,1 %	95,9 %	100,0 %	91,0 %	95,2 %
Tao2	100,0 %	94,9 %	97,4 %	100,0 %	93,8 %	96,8 %
Tao3	100,0 %	86,8 %	92,9 %	100,0 %	85,7 %	92,3 %
Citi1	99,0 %	84,7 %	91,3 %	99,2 %	83,0 %	90,4 %
Citi2	99,8 %	85,8 %	92,3 %	99,8 %	82,7 %	90,5 %
Verne	98,6 %	62,0 %	76,1 %	97,2 %	59,7 %	74,0 %
Xerox	99,2 %	2,7 %	5,2 %	98,9 %	2,8 %	5,4 %
Moyenne BAF	99,6 %	77,1 %	83,3 %	99,4 %	76,3 %	86,3 %
Moyenne BAF*	99,6 %	84,5 %	91,2 %	99,5 %	83,6 %	90,6 %

tableau 90 : corrélation entre la densité de cognats et les résultats.

Paramétrages identiques à ceux du tableau 89 (pondération 2).

	Mesures issues des comparaisons			***Résultats***		
	NbComp	***NbCognats***	d_{Cognat}	***P***	***R***	***F***
Cour	26 738 260	25 000	0,09 %	99,7 %	86,2 %	92,5 %
Hans	32 808 436	28 926	0,09 %	99,1 %	73,1 %	84,1 %
Ilo	1,04E+08	108 924	0,10 %	100,0 %	87,8 %	93,5 %
Onu	31 694 853	40 898	0,13 %	99,8 %	93,2 %	96,4 %
Tao1	4 469 875	6 503	0,15 %	100,0 %	91,0 %	95,2 %
Tao2	5 200 048	8 453	0,16 %	100,0 %	93,8 %	96,8 %
Tao3	1 988 832	2 873	0,14 %	100,0 %	85,7 %	92,3 %
Citi1	8 653 652	10 072	0,12 %	99,2 %	83,0 %	90,4 %
Citi2	11 603 884	15 730	0,14 %	99,8 %	82,7 %	90,5 %
Verne	22 955 239	13 190	0,06 %	97,2 %	59,7 %	74,0 %
Xerox	15 716 994	44 065	0,28 %	98,9 %	2,8 %	5,4 %
Moyenne BAF			0,13 %	99,43 %	76,26 %	82,82 %
Moyenne BAF*			0,12 %	99,48 %	83,61 %	90,56 %

NbComp : nombre de comparaison effectuées
NbCognats : nombre de cognats identifiés lors des comparaisons
d_{cognat} : densité de cognat (NbCognats/NbComp)

A-IV Alignement

tableau 91 : probabilités empiriques des transitions T1=1:1, T2=2:1, T3=1:2, T4=0:1, T5=1:0, T6=2:2, T7=3:1, T8=1:3.

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	*Autre*
Cour	92,7 %	0,9 %	3,2 %	0,2 %	1,2 %	1,2 %	0,0 %	0,2 %	0,5 %
Hans	84,4 %	4,9 %	2,9 %	3,3 %	3,3 %	0,4 %	0,4 %	0,1 %	0,3 %
Ilo	94,4 %	0,7 %	2,8 %	0,4 %	0,3 %	0,4 %	0,0 %	0,5 %	0,5 %
Onu	95,5 %	2,3 %	0,7 %	0,7 %	0,1 %	0,5 %	0,0 %	0,0 %	0,1 %
Tao1	92,8 %	2,2 %	0,8 %	0,3 %	2,5 %	0,8 %	0,0 %	0,0 %	0,6 %
Tao2	93,1 %	2,3 %	0,0 %	0,3 %	0,3 %	3,0 %	0,7 %	0,0 %	0,3 %
Tao3	88,6 %	5,1 %	1,7 %	0,0 %	0,6 %	0,6 %	0,0 %	0,0 %	3,4 %
Citi1	90,6 %	3,6 %	2,7 %	0,0 %	0,0 %	0,2 %	0,9 %	0,0 %	2,0 %
Citi2	83,7 %	1,6 %	4,1 %	0,5 %	0,2 %	7,8 %	0,0 %	0,1 %	2,0 %
Verne	63,8 %	3,3 %	3,9 %	0,7 %	23,2 %	2,9 %	0,3 %	0,5 %	1,4 %
Xerox	92,7 %	0,3 %	0,3 %	1,0 %	3,9 %	1,5 %	0,0 %	0,0 %	0,3 %

tableau 92 : résultats de la méthode GC (longueurs de phrases en nombre de mots).

Paramétrages :
$R_{diag} = m/n = 1,203 \quad s^2=1,025$
Critère de diagonalité dans une section encadrée par (I_d,J_d) *et* (I_f,J_f)*:*
LargeurSection = min(If-Id,Jf-Jd), si Largeur Section > 10 alors $Seuil_{diag} = 0,2$ *; si* $10 \geq$ *Largeur Section > 3 alors* $Seuil_{diag} = 1$ *; si* $3 \geq$ *Largeur Section 3 alors* $Seuil_{diag} = 2,5$
$p(T_1\text{-}T_6)$=(0,89 ; 0,0425 ; 0,0425 ; 0,00495 ; 0,00495 ; 0,011)
$p(T_1\text{-}T_8)$=(0,883; 0,0442; 0,0442; 0,0049; 0,0049; 0,01; 0,0044; 0,0044)

	Transitions T_1-T_6			*Transitions* T_1-T_8		
	P	***R***	***F***	***P***	***R***	***F***
Cour	98,3 %	97,1 %	97,7 %	98,3 %	97,2 %	97,7 %
Hans	93,8 %	95,0 %	94,4 %	94,3 %	95,9 %	95,1 %
Ilo	98,5 %	97,1 %	97,8 %	98,7 %	97,8 %	98,3 %
Onu	98,1 %	98,2 %	98,1 %	98,1 %	98,2 %	98,1 %
Tao1	98,4 %	98,2 %	98,3 %	98,4 %	98,2 %	98,3 %
Tao2	95,4 %	97,2 %	96,3 %	97,4 %	98,8 %	98,1 %
Tao3	93,1 %	88,5 %	90,8 %	93,2 %	88,6 %	90,8 %
Citi1	95,8 %	92,6 %	94,2 %	95,6 %	92,6 %	94,0 %
Citi2	96,2 %	95,1 %	95,6 %	96,2 %	95,2 %	95,7 %
Verne	65,6 %	64,5 %	65,0 %	62,5 %	63,1 %	62,8 %
Xerox	96,5 %	3,9 %	7,5 %	96,4 %	3,9 %	7,5 %
Moyenne BAF	93,6 %	84,3 %	85,1 %	93,5 %	84,5 %	85,1 %
Moyenne BAF*	93,3 %	92,3 %	92,8 %	93,3 %	92,6 %	92,9 %

tableau 93 : résultats de la méthode GC (longueurs de phrases en nombre de caractères).

Paramétrages :

Identiques à ceux du tableau 92, sauf pour R_{diag} = m/n = 1,18 et s^2=4

	Transitions T_1-T_6			***Transitions T_1-T_8***		
	P	***R***	***F***	***P***	***R***	***F***
Cour	99,3 %	97,4 %	98,4 %	99,3 %	97,4 %	98,4 %
Hans	94,8 %	95,4 %	95,1 %	95,0 %	96,4 %	95,7 %
Ilo	99,0 %	97,3 %	98,1 %	99,1 %	98,2 %	98,6 %
Onu	99,3 %	98,6 %	98,9 %	99,3 %	98,6 %	98,9 %
Tao1	98,9 %	97,8 %	98,3 %	98,9 %	97,8 %	98,3 %
Tao2	96,6 %	96,6 %	96,6 %	99,2 %	98,8 %	99,0 %
Tao3	98,0 %	91,5 %	94,6 %	98,0 %	91,6 %	94,7 %
Citi1	97,4 %	92,7 %	95,0 %	97,4 %	92,9 %	95,1 %
Citi2	97,4 %	95,1 %	96,3 %	97,3 %	95,1 %	96,2 %
Verne	64,8 %	66,0 %	65,4 %	63,7 %	65,6 %	64,7 %
Xerox	97,0 %	3,9 %	7,5 %	96,6 %	3,9 %	7,5 %
Moyenne BAF	94,8 %	84,8 %	85,8 %	94,9 %	85,1 %	86,1 %
Moyenne BAF*	94,6 %	92,8 %	93,7 %	94,7 %	93,2 %	94,0 %

tableau 94 : corrélations entre les longueurs exprimées en mots ou en nombre de caractères.

	Longueurs en nombre de caractères			***Longueurs en nombre de mots***		
	Moy L'/L	***Var L'/L***	***Corrélation***	***Moy N'/N***	***Var N'/N***	***Corrélation***
Cours	1,078	0,046	0,988	1,175	0,069	0,983
Hans	1,131	0,124	0,958	1,173	0,151	0,946
Ilo	1,047	0,048	0,988	1,092	0,097	0,980
Onu	1,084	0,042	0,982	1,169	0,079	0,957
TAO1	1,173	0,047	0,977	1,199	0,066	0,970
TAO2	1,086	0,022	0,978	1,101	0,031	0,964
TAO3	1,068	0,026	0,944	1,101	0,053	0,923
CITI1	1,132	0,986	0,937	1,228	1,006	0,927
CITI2	1,111	0,459	0,981	1,175	0,608	0,968
Verne	1,123	0,230	0,835	1,173	0,247	0,816
Xerox	1,246	0,062	0,978	1,203	0,079	0,970
BAF	1,119	0,123	0,973	1,161	0,160	0,962

Pour chaque binôme comparé, L et L' sont les longueurs des segments anglais et français en nombre de caractères, et N et N' sont les longueurs des segments anglais et français en nombre de mots.

tableau 95 : résultats de la méthode GC (longueurs de phrases en nombre de caractères), avec les probabilités de transitions empiriques

Paramétrages :

Identiques à ceux du tableau 93, sauf pour p(T_1-T_8)=(91,73 ; 1,69; 2,15 ; 0,94 ; 1,44 ; 1,19 ; 0,11 ; 0,21 ; 0,54)

	Transitions T_1-T_8 empiriques		
	P	***R***	***F***
Cour	99,3 %	97,4 %	98,4 %
Hans	95,2 %	96,2 %	95,7 %
Ilo	99,0 %	97,4 %	98,2 %
Onu	99,1 %	98,4 %	98,8 %
Tao1	98,7 %	97,4 %	98,0 %
Tao2	99,2 %	98,8 %	99,0 %
Tao3	95,9 %	90,2 %	93,0 %
Citi1	97,3 %	92,4 %	94,8 %
Citi2	97,3 %	95,0 %	96,2 %
Verne	64,8 %	65,9 %	65,3 %
Xerox	96,9 %	3,9 %	7,5 %
Moyenne BAF	94,8 %	84,8 %	85,9 %
Moyenne BAF*	94,6 %	92,9 %	93,7 %

tableau 96 : résultats avec une mesure de distance combinant les longueurs de phrase et la cognation.

Paramétrages :

Pour la distance de la méthode GC, on reprend les paramétrages du tableau 93, avec T1-T8. On étudie alors l'évolution des résultats en fonction la pondération entre D_{GC} et D_{cognat} (cf. équation 34 et 35) : Distance = (1- k_{co}) DGC + k_{co} D_{cognat}

Moyenne BAF *		***$D_{cognat2}$***			***$D_{cognat1}$***		
(1-k_{co})	***k_{co}***	***P***	***R***	***F***	***P***	***R***	***F***
1	0	94,7 %	93,2 %	94,0 %	94,7 %	93,2 %	94,0 %
0,75	0,25	95,3 %	93,9 %	94,6 %	95,3 %	93,8 %	94,5 %
0,5	0,5	95,9 %	94,2 %	95,1 %	95,9 %	94,1 %	95,0 %
0,25	0,75	96,5 %	94,8 %	95,6 %	96,3 %	94,5 %	95,3 %
0,2	0,8	96,3 %	94,7 %	95,5 %	96,1 %	94,4 %	95,2 %
0	1	95,3 %	92,0 %	93,6 %	94,9 %	91,2 %	93,0 %

figure 72 : Evolution de P et R en fonction de kco

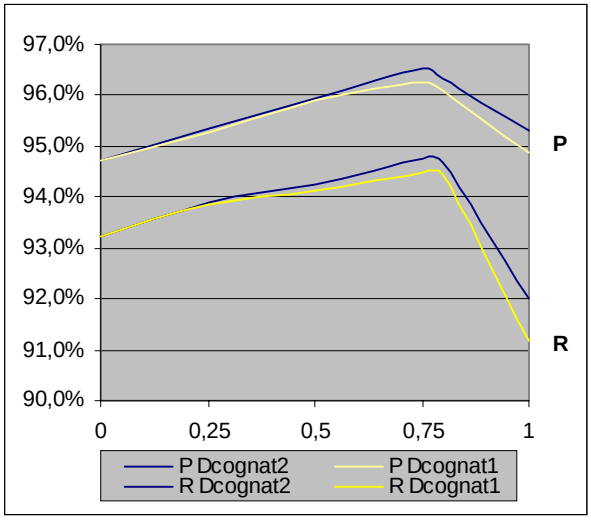


tableau 97 : résultats détaillés avec une mesure de distance combinant les longueurs de phrase et la cognation, de paramètre kco=0,75

	k_{co}=0,75		
	P	***R***	***F***
Cour	99,8 %	97,4 %	98,6 %
Hans	98,6 %	97,0 %	97,8 %
Ilo	99,9 %	98,7 %	99,3 %
Onu	99,6 %	98,9 %	99,2 %
Tao1	99,3 %	98,1 %	98,7 %
Tao2	99,6 %	98,8 %	99,2 %
Tao3	99,6 %	96,0 %	97,8 %
Citi1	98,9 %	93,5 %	96,1 %
Citi2	99,4 %	96,4 %	97,9 %
Verne	70,6 %	72,7 %	71,7 %
Xerox	97,0 %	3,9 %	7,6 %
Moyenne BAF	96,6 %	86,5 %	87,6 %
Moyenne BAF*	96,5 %	94,8 %	95,6 %

tableau 98 : résultats détaillés avec une mesure de distance combinant les longueurs de phrase et la cognation, de paramètre* k_{co}=0,75 *(probabilités de transition empiriques).

k_{co}=0,75 Transitions T_1-T_8 empiriques	*P*	*R*	*F*
Cour	99,8 %	97,4 %	98,6 %
Hans	98,5 %	96,8 %	97,6 %
Ilo	99,9 %	98,7 %	99,3 %
Onu	99,8 %	98,8 %	99,3 %
Tao1	99,7 %	98,1 %	98,9 %
Tao2	99,6 %	98,8 %	99,2 %
Tao3	99,6 %	96,1 %	97,8 %
Citi1	98,5 %	92,9 %	95,6 %
Citi2	99,6 %	96,3 %	97,9 %
Verne	77,0 %	75,2 %	76,1 %
Xerox	97,5 %	3,9 %	7,5 %
Moyenne BAF	97,2 %	86,6 %	88,0 %
Moyenne BAF*	97,2 %	94,9 %	96,0 %

A-V Exemples de correspondances tirées du corpus de référence

Correspondances extraites manuellement du corpus JOC, après alignement automatique.

00000327

The Community and its Member States will continue to make their views known to the authorities of Malawi as they did in the case of the intimidation of Catholic Bishops after which they received assurances that the Bishops would be in no danger and free to move about and to conduct Church services .

La Communauté et ses États membres continueront à faire connaître leur position aux autorités du Malawi comme ils l ' ont fait dans le cas de l ' intimidation des évêques catholiques , à la suite de quoi ils ont reçu l ' assurance que les évêques ne seraient pas en danger , qu ' ils bénéficieraient de la liberté de mouvement et de celle de célébrer les offices religieux .

(The, La) (Community, Communauté) (and, et) (its, ses) (Member States, États membres) (continue, continueront) (make, faire) (their, leur) (views, position) (known, connaître) (to, aux) (authorities, autorités) (of, du) (Malawi, Malawi) (as, comme) (they, ils) (of, des) (did, fait) (in danger, en danger) (in the case of, dans le cas de) (the, l ') (intimidation, intimidation) (Catholic, catholiques) (the, les) (Bishops, évêques) (after which, à la suite de) (which, quoi) (they, ils) (received, reçu) (assurances, assurance) (that, que) (Bishops, évêques) (would be, seraient) (and, et) (free, liberté) (move about, mouvement) (conduct, célébrer) (Church services, les offices religieux)

00002120

Given the historic nature of Arachova , will the Commission intervene to prevent the Athens - Delphi road being widened by pulling down listed buildings in the town ?

La Commission a - t - elle , compte tenu du caractère traditionnel de l ' architecture d ' Arahova , l ' intention de manifester son intérêt afin qu ' il ne soit pas procédé à l ' élargissement de la route Athènes - Delphes au détriment des édifices classés de la ville ?

(Given, compte tenu du) (historic nature, caractère traditionnel) (of, de) (Arachova, Arahova) (will, a l ' intention de) (the, La) (Commission, Commission) (intervene, manifester son intérêt) (Athens - Delphi, Athènes - Delphes) (road, route) (widened, élargissement) (listed, classés) (buildings, édifices) (the, la) (town, ville)

00002211

Will the Commission take steps with a view to decentralization in Greece ?

La Commission entend - elle agir afin que la décentralisation soit mise en oeuvre en Grèce également ?

(Will, entend) (the, La) (Commission, Commission) (take steps, agir) (with a view to, afin que) (decentralization, décentralisation) (in, en) (Greece, Grèce)

00004116

However , applicant countries are required to acknowledge the existence of foreign , security and defence policies as an ` acquis communautaire '

Toutefois , au nombre des conditions que les pays candidats devront accepter comme acquis communautaire , figurent les politiques extérieure , de sécurité et de défense

(However, Toutefois) (applicant, candidats) (countries, pays) (are required to, devront) (acknowledge, accepter) (foreign, extérieure) (security, sécurité) (and, et) (defence, défense) (policies, politiques) (as, comme) (acquis communautaire, acquis communautaire)

00005985

Thus the United States applies the reduced rate as a matter of course , i.e . 15 % instead of 30 % , if the US company indicates in its tax return the address in Germany of the recipient of the dividends .

Les États - Unis d ' Amérique accordent ainsi directement l ' application du taux réduit , c ' est - à - dire 15 % au lieu de 30 % , si la société américaine , lors de sa déclaration fiscale , communique l ' adresse du destinataire des dividendes en république fédérale d ' Allemagne .

(Thus, ainsi) (the, Les) (United States, États - Unis d ' Amérique) (applies, application) (the, du) (reduced rate, taux réduit) (i.e, c ' est - à - dire) (15, 15) (%, %) (instead of, au lieu de) (30, 30) (%, %) (if, si) (the, la) (US, américaine) (company, société) (indicates, communique) (in, lors de) (its, sa) (tax return, déclaration fiscale) (the, l ') (address, adresse) (in, en) (Germany, république fédérale d ' Allemagne) (of the, du) (recipient, destinataire) (of the, des) (dividends, dividendes)

00008072

On 22 June , the Council communicated the above to the Board of Governors of the EIB , which may authorize the Bank to launch these new activities as soon as the Council and Parliament have decided that the loans in question should be guaranteed under the Community budget .

Le 22 juin 1992 , le Conseil a communiqué les conclusions ci - dessus au Conseil des gouverneurs de la BEI , qui pourra autoriser la Banque à démarrer ces nouvelles activités dès que le Conseil des ministres et le Parlement européen auront décide de la prise en charge de la garantie des prêts en question par le budget communautaire .

(On, Le) (22, 22) (June, juin) (the, le) (Council, Conseil) (communicated, a communiqué) (the, les) (above, ci - dessus) (to, au) (Board of Governors, Conseil des gouverneurs) (of, de) (the, la) (EIB, BEI) (which, qui) (may, pourra) (authorize, autoriser) (the, la) (Bank, Banque) (to, à) (launch, démarrer) (these, ces) (new, nouvelles) (activities, activités) (as soon as, dès que) (the, le) (Council, Conseil) (and, et) (Parliament, Parlement) (have, auront) (decided that, décidé de) (the, des) (loans, prêts) (in question, en question) (guaranteed, garantie) (under, par) (the, le) (Community, communautaire) (budget, budget)

00009545

The centenary of the discovery of America by Christopher Columbus has been marked by a great number of festivities and events , many of which have taken place with the support of the European Community

Le cinquième centenaire de la découverte de l ' Amérique par Christophe Colomb a été marqué par de nombreuses festivités et manifestations ¤ Bon nombre de celles - ci ont bénéficié du concours de la Communauté européenne

(The, Le) (centenary, centenaire) (of, de) (the, la) (discovery, découverte) (of, de) (America, l ' Amérique) (by, par) (Christopher, Christophe) (Columbus, Colomb) (marked, marqué) (number, nombreuses) (of, de) (festivities, festivités) (and, et) (events, manifestations) (many of, Bon nombre de) (have, ont) (taken place with the support of, bénéficié du concours de) (the, la) (European Community, Communauté européenne)

00010786

At the moment the Commission is waiting for them to reply .

À l ' heure actuelle , la Commission attend une réponse de leur part .

(At the moment, À l ' heure actuelle) (the, la) (Commission, Commission) (is waiting for, attend) (reply, réponse)

00010999

In a letter dated 14 May 1990 the Commission informed the Greek authorities that aid to a particular aquaculture project in the Amvrakikos area was suspended

Dans une lettre datée du 14 mai 1990 , la Commission a notifié aux autorités helléniques la suspension de l ' aide accordée à un projet spécifique d ' aquaculture dans le golfe d ' Ambracie

(In, Dans) (a, une) (letter, lettre) (dated, datée) (14, 14) (May, mai) (1990) (the, la) (Commission, Commission) (informed, a notifié aux) (Greek, helléniques) (authorities, autorités) (aid, aide) (to, accordée à) (a, un) (particular, spécifique) (aquaculture, aquaculture) (project, projet) (in, dans) (the, le) (Amvrakikos area, golfe d ' Ambracie) (suspended, suspension)

00011065

This is because there is good reason to believe that - at least within the EC - the successful introduction of terrestrial DAB is a precondition for the introduction of satellite DAB .

En effet , il semble justifié de l ' introduction de la radiodiffusion numérique terrestre est une condition préalable à l ' introduction de la radiodiffusion numérique par satellites .

(This is because, En effet) (there is good reason to believe that, il semble justifié de) (the, l ') (introduction, introduction) (of, de) (terrestrial, terrestre) (DAB, radiodiffusion numérique) (is, est) (a, une) (precondition for, condition préalable à) (the, l ') (introduction, introduction) (of, de) (satellite, satellites) (DAB, radiodiffusion numérique)

00011099

The project will be implemented by the Federación de Cámaras y Asociaciones de Exportadores de Centroamérica y el Caribe (FECAEXCA) , a private sector body which put forward an initial version of the scheme through the EEC - Central America Joint Committee

Ce projet va être mis en oeuvre par la Federación de Cámaras y Asociaciones de Exportadores de Centroamérica y el Caribe (FECAEXCA) , une association du secteur privé qui avait présenté une première version du projet dans le cadre de la Commission Mixte CEE - Amérique centrale

(project, projet) (will be, va être) (implemented, mis en oeuvre) (by, par) (the, la) (Federación de Cámaras y Asociaciones de Exportadores de Centroamérica y el Caribe FECAEXCA, Federación de Cámaras y Asociaciones de Exportadores de Centroamérica y el Caribe FECAEXCA) (a, une) (private sector, secteur privé) (body, association) (which, qui) (put forward, avait présenté) (an, une) (initial, première) (version, version) (of the, du) (scheme, projet) (through, dans le cadre de) (the, la) (EEC - Central America Joint Committee, Commission Mixte CEE - Amérique centrale)

00012098

N º 3084/92 by Mr José Valverde López (PPE) to the Commission (

N º 3084/92 de M . José Valverde López (PPE) à la Commission (

(N º, N º) (3084/92, 3084/92) (by, de) (Mr José Valverde López, M . José Valverde López) (PPE, PPE) (to, à) (the, la) (Commission, Commission)

00012733

Does the Commission intend to provide economic assistance for those farmers in the Prefecture of Rodopi whose crops have been badly affected by adverse weather conditions ? Does it intend to use the special programmes drawn up to cover cases of this kind in this instance ?

La Commission a - t - elle l ' intention d ' adopter des mesures destinées à venir financièrement en aide aux agriculteurs de la préfecture de Rhodope , dont la production a été victime des conditions climatiques défavorables ? Est - elle disposée à mettre en oeuvre à cette fin des programmes spéciaux destinés à faire face à ce genre de circonstances ?

(the, La) (Commission, Commission) (intend to, Est disposée à) (provide assistance for, venir en aide aux) (economic, financièrement) (farmers, agriculteurs) (in, de) (the, la) (Prefecture, préfecture) (of, de) (Rodopi, Rhodope) (whose, dont) (crops, production) (have been, a été) (badly affected by, victime des) (adverse, défavorables) (weather, climatiques) (conditions, conditions) (it, elle) (intend to, a l ' intention d ') (use, mettre en oeuvre) (the, des) (special, spéciaux) (programmes, programmes) (drawn up to, destinés à) (cover, faire face à) (cases of this kind, ce genre de circonstances)

00012837

Of the forty or so schemes envisaged , half will be managed by NGOs , likewise with a view to positive promotion of fundamental rights , involving training programmes and technical and data - processing support .

De la quarantaine de projets envisagés , la moitié sera gérée par des ONGs , toujours dans la perspective d ' une promotion positive des droits fondamentaux qui comporte aussi bien des programmes de formation que d ' appui technique et informatique .

(Of, De) (the, la) (forty or so, quarantaine de) (schemes, projets) (envisaged, envisagés) (half, la moitié) (will be, sera) (managed, gérée) (by, par) (NGOs, ONGs) (likewise, toujours) (with a view to, dans la perspective d ') (positive, positive) (promotion, promotion) (of, des) (fundamental, fondamentaux) (rights, droits) (training, formation) (programmes, programmes) (and, et) (technical, technique) (and, aussi bien que) (data - processing, informatique) (support, appui)

00014429

Efficiency and quality within the food industry will demand that note be taken of imperfections that will constantly arise within a market driven system .

L ' efficacité et la qualité dans le secteur alimentaire exigent que l ' on se préoccupe des imperfections se manifestant constamment dans un système axé sur le marché

(Efficiency, efficacité) (and, et) (quality, qualité) (within, dans) (the, le) (food industry, secteur alimentaire) (demand that, exigent que) (note be taken of, on se préoccupe des) (imperfections, imperfections) (constantly, constamment) (arise, se manifestant) (within, dans) (a, un) (market driven, axé sur le marché) (system, système)

00015063

2 . that nuclear scientists who may be left unemployed following nuclear disarmament in the Soviet Union will not offer the benefit of their experience to countries which are in the process of building a nuclear arsenal ?

b) l ' assurance que les experts nucléaires qui pourraient perdre leur emploi du fait de la réduction de l ' arsenal nucléaire soviétique n ' offriront pas leur expérience à des pays qui veulent se doter de l ' arme nucléaire ?

(that, que) (nuclear, nucléaires) (scientists, experts) (who, qui) (may, pourraient) (be left unemployed, perdre leur emploi) (following, du fait de) (nuclear, nucléaire) (disarmament, réduction de l ' arsenal) (in the Soviet

Union, soviétique) (will offer, offriront) (their, leur) (experience, expérience) (to, à) (countries, pays) (which, qui) (building a nuclear arsenal, se doter de l ' arme nucléaire)

00015472

Following the amendments by the European Parliament the 1993 budget adopted on 17 December 1992 provides for ECU 14 million for this budget heading .

Suite aux amendements du Parlement européen , le budget 1993 adoptée le 17 décembre dernier prévoit 14 millions d ' écus à cette ligne budgétaire .

(Following, Suite aux) (amendments, amendements) (by, du) (European Parliament, Parlement européen) (the, le) (1993, 1993) (budget, budget) (adopted, adoptée) (on, le) (17, 17) (December, décembre) (provides for, prévoit) (ECU, écus) (14, 14) (million, millions) (this, cette) (budget heading, ligne budgétaire)

00015652

The Community and its Member States will continue to follow closely the case of Mr Chihana and will decide what further action to take in the light of the decision on his appeal .

La Communauté et ses États membres continueront à suivre de près le cas de M . Chihana et décideront de ce qu ' il y a lieu de faire au vu de la décision qui sera prise en appel .

(The, La) (Community, Communauté) (and, et) (its, ses) (Member States, États membres) (will continue to, continueront à) (follow, suivre) (closely, de près) (the, le) (case, cas) (of, de) (Mr Chihana, M . Chihana) (and, et) (will decide, décideront de) (what action to take, ce qu ' il y a lieu de faire) (in the light of, au vu de) (the, la) (decision, décision) (appeal, appel)

00015698

In view of the importance of this matter , an ad hoc High - Level Working Party has been set up at the Council .

Eu égard à l ' importance de cette question , un groupe ad hoc à haut niveau a été créé au sein du Conseil .

(In view of, Eu égard à) (the, l ') (importance, importance) (of, de) (this, cette) (matter, question) (an, un) (ad hoc, ad hoc) (High - Level Working Party, groupe à haut niveau) (has been, a été) (set up, créé) (at, au sein du) (Council, Conseil)

00015941

Residence and employment in the Member States

Séjour et activité professionnelle dans les États membres

(Residence, Séjour) (and, et) (employment, activité professionnelle) (in, dans) (the, les) (Member States, États membres)

00017020

Will the Community be adding its (financial) support to the National Academy of Sciences ' study proposal , ` The Conservation of Egyptian Monuments ' ?

La Communauté est - elle disposée à apporter son appui (financier) à la proposition visant à préserver les monuments égyptiens contenue dans l ' étude réalisée à ce sujet par l ' Académie nationale des sciences ?

(Will, est disposée à) (the, la) (Community, Communauté) (adding its support, apporter son appui) (financial, financier) (to, à) (the, La) (National Academy of Sciences, Académie nationale des sciences) (study, étude) (proposal, proposition) (Conservation, préserver) (Egyptian, égyptiens) (Monuments, monuments)

00018750

In the last financial year , however , payment appropriations earmarked for the Structural Fund operations in the new Länder were so low that it was not possible to grant all applications for the second advance

Les prévisions de crédits de paiement relatifs à ces interventions des fonds structurels dans les nouveaux Länder allemands furent , déjà pour le dernier exercice budgétaire , si basses qu ' il fut impossible d ' autoriser toutes les demandes au titre de la deuxième avance

(the, les) (last, dernier) (financial year, exercice budgétaire) (payment, paiement) (the, la) (Structural Fund, fonds structurels) (operations, interventions) (in, dans) (the, les) (new, nouveaux) (Länder, Länder) (were, furent) (so, si) (low, basses) (that, qu ') (it, il) (was, fut) (not possible to, impossible d ') (grant, autoriser) (all, toutes) (applications for, demandes au titre de) (second, deuxième) (advance, avance)

00018819

The first is Speech Analytic Hearing Aids for the Profoundly Deaf in Europe (Stride) , and is concerned with the development of a wearable prototype aid , providing optimal auditory support for lip - reading to the many people in Europe with such profound impairment that conventional aids are of little assistance

Le premier est dénommé Speech Analytic Hearing Aids for the Profoundly Deaf in Europe (Stride) et vise le développement d ' un prototype d ' appareil portable , assurant une aide auditive optimale à la lecture labiale pour les nombreuses personnes en Europe qui présentent un handicap tel que les moyens conventionnels se révèlent peu adéquats

(The, Le) (first, premier) (is, est dénommé) (Speech Analytic Hearing Aids for the Profoundly Deaf in Europe Stride, Speech Analytic Hearing Aids for the Profoundly Deaf in Europe Stride) (and, et) (is concerned with, vise) (the, le) (development, développement) (of, d ') (a, un) (wearable prototype aid, prototype d ' appareil portable) (providing, assurant) (optimal, optimale) (auditory, auditive) (support for, aide à) (lip - reading, lecture labiale) (to, pour) (the, les) (many, nombreuses) (people, personnes) (in, en) (Europe, Europe) (with, qui présentent) (such, tel) (impairment, handicap) (that, que) (conventional, conventionnels) (aids, moyens) (are, se révèlent) (of little assistance, peu adéquats)

00020053

5 ¤ In most cases there are no guarantees that the companies undertaking projects are able satisfactorily to carry out the training measures as part of the programmes pursued by local government and non - profitmaking companies .

5) dans la plupart des cas , il n ' existe aucune garantie que les sociétés commanditées ont les capacités requises pour exécuter convenablement , en ce qui concerne les programmes des collectivités locales , mais aussi des associations à but non lucratif , les actions de formation dont elles ont la charge .

(5, 5) (In most cases, dans la plupart des cas) (there are no, il n ' existe aucune) (guarantees, garantie) (that, que) (the, les) (companies, sociétés) (undertaking projects, commanditées) (are able to, ont les capacités pour) (satisfactorily, convenablement) (to, pour) (carry out, exécuter) (the, les) (training, formation) (measures, actions) (as part of, en ce qui concerne) (programmes, programmes) (local government, collectivités locales) (and, mais aussi) (non - profitmaking companies, associations à but non lucratif)

00020399

The reduction in the funding of TIDE from ECU 8 million to ECU 2 million is undermining the excellent potential of this project as it is sending out negative signals to the many industrial and academic interests stimulated by the successful launch of the pilot action

La diminution des subventions affectées au programme Tide , dont la dotation est passée de 8 millions d ' écus à 2 millions d ' écus , rétrécit le champ des possibilités extraordinaires offertes par ce projet étant donné

qu ' elle transmet une image négative aux nombreux groupes industriels et universitaires qui se réjouissaient du succès du lancement de cette action - pilote

(The, La) (reduction in, diminution des) (the, des) (funding, subventions) (of, au) (TIDE, Tide) (from, de) (8, 8) (million, millions) (ECU, écus) (to, à) (ECU, écus) (2, 2) (million, millions) (is undermining, rétrécit) (the, le) (excellent, extraordinaires) (potential, champ des possibilités) (this, ce) (project, projet) (as, étant donné qu ') (it, elle) (is sending out negative signals, transmet une image négative) (to, aux) (many, nombreux) (industrial, industriels) (and, et) (academic, universitaires) (successful, succès) (launch, lancement) (of, de) (pilot action, action - pilote)

00020903

A preliminary examination of the provisions on the pursuit of the profession of statutory auditor in Greece (Presidential Decree No 226/1992) (1) , which have now been adopted but not yet formally notified to the Commission , gives rise to the following observations :

Un premier examen de la législation sur l ' exercice de la profession d ' expert compable en Grèce [décret présidentiel n º 226/1992 (1)] qui a été promulguée entre - temps , mais qui n ' a pas encore été notifié à la Commission , permet de faire les constatations suivantes .

(A, Un) (preliminary, premier) (examination, examen) (of, de) (the, la) (provisions, législation) (on, sur) (the, l ') (pursuit of, exercice de) (the, la) (profession, profession) (of, d’) (statutory auditor, expert comptable) (in, en) (Greece, Grèce) (Presidential Decree, décret présidentiel) (No, n º) (226/1992, 226/1992) (1, 1) (which, qui) (have been, a été) (now, entre - temps) (adopted, promulguée) (but, mais) (not yet, pas encore) (notified to, été notifié à) (the, la) (Commission, Commission) (gives rise to, permet de faire) (the, les) (following, suivantes) (observations, constatations)

00021198

The Erasmus Programme was set up to facilitate exchanges between the higher education systems of the Member States , and applies both to teaching staff and students , particularly students who have completed at least two years of higher education

Le programme Erasmus a été conçu pour faciliter les échanges entre les systèmes d ' Enseignement supérieur des pays de la Communauté et concerne , aussi bien les enseignements que les étudiants ¤ Il s ' applique plus particulièrement aux étudiants de niveau supérieur à BAC + ¤ 2

(The, Le) (Erasmus Programme, programme Erasmus) (was, a été) (set up to, conçu pour) (facilitate, faciliter) (exchanges, échanges) (between, entre) (the, les) (higher, supérieur) (education, Enseignement) (systems, systèmes) (of the, des) (Member States, pays la Communauté) (and, et) (applies to, concerne) (both and, aussi bien que) (teaching, enseignements) (students, étudiants) (particularly, particulièrement) (students, étudiants) (at least, supérieur) (two years of higher education, BAC + 2)

00022119

A commitment appropriation of Ecu 3 929 million was entered against Item B3 - 4012 of the budget for the 1992 financial year .

La ligne budgétaire B3 - 4012 prévoyait , pour l ' exercice 1992 , un crédit d ' engagement de 3 929 000 écus .

(A, un) (commitment appropriation, crédit d ' engagement) (Ecu, écus) (3 929, 3 929) (million, 000) (Item of the budget, ligne budgétaire) (B3 - 4012, B3 - 4012) (for, pour) (1992, 1992) (financial year, exercice)

00023281

The Commission has also set up specific schemes outside the Structural Funds involving , in particular , the conduct of individual assessment exercises to prepare those concerned for a new job ;

De plus , la Commission a décidé la mise en place d ' actions spécifiques hors Fonds structurels permettant notamment la réalisation de bilans individuels préventifs , en vue de la préparation des personnes concernées pour un nouvel emploi .

(The, la) (Commission, Commission) (has, a) (set up, décidé la mise en place d ') (also, De plus) (specific, spécifiques) (schemes, actions) (outside, hors) (Structural Funds, Fonds structurels) (in particular, notamment) (the, la) (conduct, réalisation) (of, de) (individual, individuels) (assessment exercises, bilans préventifs) (to, en vue de) (prepare, préparation) (those, personnes) (concerned, concernées) (for, pour) (a, un) (new, nouvel) (job, emploi)

00023359

In the wetland of Alyki on Kos , in particular , the populations of flamingoes , herons , marbled ducks , little egrets and gulls

Dans le biotope d ' Alyki , en particulier , on a constaté une diminution spectaculaire du nombre des flamants roses , hérons sarcelles marbrées , aigrettes garzettes , et même des mouettes

(In, Dans) (the, le) (wetland, biotope) (of, d ') (Alyki on Kos, Alyki) (in particular, en particulier) (flamingoes, flamants roses) (herons, hérons) (marbled ducks, sarcelles marbrées) (little egrets, aigrettes garzettes) (and, et) (gulls, mouettes)

00024410

given the drop in employment which has resulted in the loss of 44 000 jobs

que la montée du chomâge a entraîné la perte de 44 000 emplois

(given, que) (the, la) (drop in employment, montée du chomâge) (has, a) (resulted in, entraîné) (the, la) (loss of, perte de) (44 000, 44 000) (jobs, emplois)

00025524

Although it is protected by the International Berne Convention , no special measures have been taken in Greece for its protection

Bien que le chacal soit protégé par la Convention de Berne , aucune mesure de protection particulière n ' est prise en Grèce

(Although, Bien que) (is, soit) (protected, protégé) (by, par) (the, la) (International Berne Convention, Convention de Berne) (no, aucune) (special, particulière) (measures for its protection, mesure de protection) (have been, est) (taken, prise) (in, en) (Greece, Grèce)

00026777

Will it provide the necessary funding and urge the Greek authorities to speed up the completion of slaughterhouses under construction and the building of new slaughterhouses ?

2) si elle compte prendre les mesures de fincancement indispensables et intervenir auprès des autorités grecques pour accélérer la réfection des abattoirs existants et la construction de nouveaux ?

(Will, compte) (it, elle) (provide funding, prendre les mesures de financement) (necessary, indispensables) (and, et) (urge, intervenir auprès des) (Greek authorities, autorités grecques) (to, pour) (speed up, accélérer) (the, la) (completion, réfection) (of, des) (slaughterhouses, abattoirs) (and, et) (the, la) (building, construction) (of, de) (new, nouveaux)

00026862

For 40 years the Greek Automobile and Touring Club (ELPA) has been providing its members , currently numbering 150 000 , with breakdown services

Voici maintenant 40 ans que l ' « Automobile et Touring Club de Grèce » (Elpa) fournit à ses membres , dont le nombre s ' élève aujourd ' hui à 150 000 , une assistance routière

(For, Voici maintenant) (40, 40) (years, ans) (then l') (Greek, Grèce) (Automobile and Touring Club, Automobile et Touring Club) (ELPA, Elpa) (has been providing with, fournit à) (its, ses) (members, membres) (currently, aujourd ' hui) (numbering, dont le nombre s ' élève à) (150 000, 150 000) (breakdown services, assistance routière)

00027248

However , the situation is still under review under Section 301 of the US trade law , which authorized the United States to reinstate the procedure and introduce countermeasures should it see fit .

Toutefois , le cas est encore sous surveillance au titre de la section 301 de la loi commerciale américaine ¤ elle autorise les États - Unis d ' Amérique à relancer la procédure et à introduire de mesure de rétorsion s ' ils l ' estiment nécessaire .

(However, Toutefois) (the, le) (situation, cas) (is, est) (still, encore) (under review, sous surveillance) (under, au titre de) (Section, section) (301, 301) (of, de) (the, la) (US, américaine) (trade law, loi commerciale) (which, elle) (authorized to, autorise à) (the, les) (United States, États - Unis d ' Amérique) (reinstate, relancer) (the, la) (procedure, procédure) (and, et) (introduce, introduire) (countermeasures, mesure de rétorsion) (should it see fit, s ' ils l ' estiment nécessaire)

00027577

Does the Commission agree that democracy in Europe is being undermined by racism ? Is the Commission prepared to provide funding for the European youth organization opposed to racism , which was set up in the summer of 1992 in response to the race riots in the Federal Republic of Germany , and its Brussels centre , which runs campaigns , provides information for young people and is holding a demonstration of European young people opposed to racism on 24 October 1992 in Brussels ? These are activities which could be eligible for funding under budget Item A 3030 (subsidies for the defence of human rights) .

La Commission ne partage - t - elle pas l ' avis selon lequel le racisme qui sévit en Europa sape la démocratie ? Aussi , ne serait - elle pas disposée à soutenir dans ses activités l ' organisation européenne de jeunes antiracistes qui s ' est créée au cours de l ' été 1992 en réaction aux violences racistes qui ont secoué l ' ex - République démocratique allemande , activités dont témoigne la base bruxelloise du mouvement qui mène des campagnes , diffuse des informations parmi les jeunes et organise par exemple le 24 octobre 1992 à Bruxelles une manifestation de jeunes Européens contre le racisme ? Ces activités pourraient relever de la ligne budgétaire A - 3030 relative aux aides en faveur des droits de l ' homme .

(the, La) (Commission, Commission) (agree that, partage l ' avis selon lequel) (democracy, démocratie) (in, en) (Europe, Europa) (undermined, sape) (racism, racisme) (Is, serait) (prepared to, disposée à) (provide, soutenir) (the, l ') (European, Européens) (youth, jeunes) (organization, organisation) (opposed to, contre) (racism, racisme) (which, qui) (was set up, s ' est créée) (in, au cours de) (summer, été) (1992, 1992) (in response to, en réaction aux) (race riots, violences racistes) (Federal Republic of Germany, République démocratique allemande) (and, et) (Brussels centre, base bruxelloise) (which, qui) (runs, mène) (campaigns, campagnes) (provides, diffuse) (information, informations) (young people, jeunes) (and, et) (is holding, organise) (a, une) (demonstration, manifestation) (of, de) (European, européenne) (young people, jeunes) (opposed to racism, antiracistes) (on, le) (24, 24) (October, octobre) (1992, 1992) (in, à) (Brussels, Bruxelles) (These, Ces) (activities, activités) (could, pourraient) (be eligible for, relever de) (budget Item, ligne budgétaire) (A 3030, A 3030) (subsidies, aides) (for the defence of, en faveur des) (human rights, droits de l ' homme)

00030286

The income of retired people in Moscow , for example , who are among the least well off , has risen by 50 % over the programme ' s period of operation .

Par exemple , durant sa mise en oeuvre , le revenu des retraités à Moscou , qui sont parmi les plus démunis , a été augmenté de 50 % .

(The, le) (income, revenu) (of, des) (retired people, retraités) (in, à) (Moscow, Moscou) (for example, Par exemple) (who, qui) (are, sont) (among, parmi) (the, les) (least well off, plus démunis) (has risen by, a été augmenté de) (50 %, 50 %) (over, durant)

00030594

What steps will the current presidency take to coordinate measures to combat unemployment and poverty in the Community ?

Quelles initiatives l ' actuelle présidence envisage - t - elle de prendre pour contribuer , au niveau communautaire , à la lutte contre le chômage et la pauvreté ?

(What, Quelles) (take steps, prendre initiatives) (will, envisage) (the, l’) (current, actuelle) (presidency, présidence) (to, pour) (to combat, lutte contre) (unemployment, chômage) (and, et) (poverty, pauvreté) (in the Community, au niveau communautaire)

00030900

The Commission manages these funds through larger - scale Operational Programmes in the context of Community Support Frameworks agreed jointly with the Member States

La Commission gère ces fonds en les affectant à de vastes programmes opérationnels en application des Cadres communautaires d ' appui (CCA) convenus d ' un commun accord avec les États membres

(The, La) (Commission, Commission) (manages, gère) (these, ces) (funds, fonds) (larger - scale, vastes) (Operational, opérationnels) (Programmes, programmes) (in the context of, en application des) (Community Support Frameworks, Cadres communautaires d ' appui CCA) (agreed, convenus) (jointly, d ' un commun accord) (with, avec) (the, les) (Member States, États membres)

00031517

Will the Commission not agree that it is entirely unsatisfactory for the European Community to be run as a partnership between the Commission and the press , thus excluding the only properly democratically elected institutions ?

La Commission ne convient - elle pas qu ' il est hautement insatisfaisant pour la Communauté européenne de se voir gérée comme une association de la Commission et de la presse , d ' où est donc exclue la seule institution issue d ' élections démocratiques appropriées ?

(the, La) (Commission, Commission) (agree that, convient qu ') (it, il) (is, est) (entirely, hautement) (unsatisfactory, insatisfaisant) (for, pour) (the, la) (European Community, Communauté européenne) (to, de) be, se voir) (run, gérée) (as, comme) (a, une) (partnership between, association de) (the, la) (Commission, Commission) (and, et) (the, la) (press, presse) (thus, donc) (excluding, exclue) (the, la) (only, seule) (properly, appropriées) (democratically, démocratiques) (elected, élections) (institutions, institution)

00031647

In the meantime the Commission is proceeding with the preparation of a guide to the special advantages available to older people throughout the Community

La Commission a entrepris entre - temps , la rédaction d ' un guide des avantages spéciaux accordés aux personnes âgées dans la Communauté

(In the meantime, entre - temps) (the, La) (Commission, Commission) (is proceeding with, a entrepris) (the, la) (preparation, rédaction) (of, d ') (a, un) (guide, guide) (to the, des) (special, spéciaux) (advantages, avantages) (available to, accordés au) (older people, personnes âgées) (throughout, dans) (the, la) (Community, Communauté)

00032024

This has also led to a number of complications , better known as computer viruses which have , with good reason , caused alarm and horror in computer circles .

Il est regrettable que cette évolution soit accompagnée de divers effets secondaires , connus sous le nom de virus informatique ¤ À juste titre , ces virus n ' ont pas manqué de susciter une certaine panique dans les milieux informatisés .

(number of, divers) (complications, effets secondaires) (better known as, connus sous le nom de) (computer, informatique) (viruses, virus) (have,ont) (caused, n ' pas manqué de susciter) (with good reason, À juste titre) (alarm and horror, panique) (in, dans) (computer, informatisés) (circles, milieux)

00036142

To illustrate how little sport receives in the way of EC aid , I recently heard that total Community aid to sport corresponds to the amount it costs to run the Community ' s olive oil publicity office in Copenhagen .

À titre d ' illustration de la modicité du subventionnement du sport par la Communauté , je viens d ' apprendre que l ' ensemble des crédits octroyés par la Communauté au sport équivaut aux frais de gestion du Bureau de l ' huile d ' olive de la l ' Communauté à Copenhague .

(To illustrate, À titre d ' illustration) (little, modicité) (sport, sport) (EC, Communauté) (aid, subventionnement) (I, je) (recently, viens d') (heard, apprendre) (that, que) (total, ensemble) (Community, Communauté) (aid, crédits) (to, au) (sport, sport) (corresponds to, équivaut aux) (amount it costs to run, frais de gestion) (Community, Communauté) (olive oil, huile d ' olive) (office, Bureau) (in, à) (Copenhagen, Copenhague)

00036211

When applicable , Community competition law can , to an extent , contribute to these objectives .

Dans son domaine d ' application , le droit communautaire de la concurrence peut contribuer pour une part à la réalisation de ces objectifs .

(applicable, application) (Community law, droit communautaire) (competition, concurrence) (can, peut) (to an extent, pour une part) (contribute to, contribuer à) (these, ces) (objectives, objectifs)

00036486

Under the pretext that crystal objects (Baccarat , Waterford , Val Saint - Lambert etc.) contain more lead than permitted by US , and especially Californian standards , new restrictions are being imposed or are apparently about to be imposed on imports of such products from Europe .

Sour le prétexte que les objets de cristal (Baccarat , Waterford , Val Saint - Lambert , . . .) comportent une composante de plomb supérieure aux normes des États - Unis d ' Amérique , dont particulièrement celles de la Californie , de nouvelles restrictions s ' imposent ou seraient en voie de s ' imposer aux importations européennes de ce type de produit .

(Under, Sous) (the, le) (pretext, prétexte) (that, que) (crystal, cristal) (objects, objets) (Baccarat, Baccarat) (Waterford, Waterford) (Val Saint - Lambert, Val Saint - Lambert) (contain, comportent) (more, supérieure)

(lead, plomb) (US, États - Unis d ' Amérique) (especially, particulièrement) (Californian, Californie) (standards, normes) (new, nouvelles) (restrictions, restrictions) (are being imposed, s’imposent) (or, ou) (about to, en voie de) (be imposed on, s ' imposer aux) (imports, importations) (of, de) (such, ce type de) (products, produit) (Europe, européennes)

00036791

Danish enterprise zones

Zones d ' activités économiques au Danemark

(Danish, Danemark) (enterprise zones, Zones d ' activités économiques)

00036867

Italian Law No 102 earmarked Lit 2 400 billion for the restoration of the Valtellina following the major flooding which took place in 1987 .

La loi italienne n º 108 avait affecté 2 400 milliards de lires au réaménagement de la Valtellina , ravagée par les inondations de 1987 .

(Italian, italienne) (Law, loi) (No, n º) (earmarked for, avait affecté au) (Lit, lires) (2 400, 2 400) (billion, milliards) (restoration, réaménagement) (of, de) (the, la) (Valtellina, Valtellina) (flooding, inondations) (in, de) (1987, 1987)

00037095

It also resolves the question of the funding of the Financial Mechanism set up by the Agreement with a view to providing financial assistance to the development and structural adjustment of certain Community Member States

Il règle aussi la question de la dotation du mécamisme financier institué par l ' Accord en vue de fournir une assistance financière pour le développement et l ' ajustement structurel de certains États membres de la Communauté

(It, Il) (also, aussi) (resolves, règle) (the, la) (question, question) (of, de) (the, la) (funding, dotation) (of, du) (Financial Mechanism, mécamisme financier) (set up, institué) (by, par) (the, l ') (Agreement, Accord) (with a view to, en vue de) (providing, fournir) (financial, financière) (assistance, assistance) (to, pour) (the, le) (development, développement) (and, et) (structural, structurel) (adjustment, ajustement) (of, de) (certain, certains) (Community Member States, États membres de la Communauté)

00037302

However , the Turkish authorities are well aware of the importance which the Community and its Member States attach to the Rule of Law and the full respect of human rights .

Cependant , les autorités turques n ' ignorent pas l ' importance que la Communauté et ses États membres attachent à la primauté du droit et au respect intégral des droits de l ' homme .

(However, Cependant) (the, les) (Turkish, turques) (authorities, autorités) (are well aware of, n ' ignorent pas) (the, l ') (importance, importance) (which, que) (the, la) (Community, Communauté) (and, et) (its, ses) (Member States, États membres) (attach to, attachent à) (the, la) (Rule of Law, primauté du droit) (and, et) (full respect, respect intégral) (of, des) (human rights, droits de l ' homme)

00037889

2 ¤ What practical measures will it take to ensure that the Council resolution is properly implemented in Greece in order to end discrimination against AIDS sufferers ?

2) quelles mesures concrètes elle compte adopter pour que la résolution du Conseil soit effectivement appliquée en Grèce , et qu ' il soit ainsi mis un terme dans ce pays aux traitements discriminatoires à l ' encontre des malades du Sida ?

(2, 2) (What, quelles) (practical, concrètes) (take measures, adopter mesures) (will, compte) (it, elle) (to ensure that, pour que) (the, la) (Council, Conseil) (resolution, résolution) (is, soit) (properly, effectivement) (implemented, appliquée) (in, en) (Greece, Grèce) (in order to, ainsi) (end, mis un terme aux) (discrimination, discriminatoires) (against, à l ' encontre des) (AIDS, Sida) (sufferers, malades)

00041979

Are the 500 or so people who were imprisoned following the uprisings in Burundi in late 1991 still in gaol ? Have they been informed of the charges against them or sent for trial ?

Les quelque 500 personnes incarcérées après les attaques lancées par des rebelles à la fin de l ' année 1991 sont - elles encore , à ce jour , détenues dans les prisons du Burundi ? Ces personnes ont - elles été informées des charges qui pèsent contre elles ou traduites en justice ?

(Are, sont) (the, Les) (500, 500) (or so, quelque) (people, personnes) (imprisoned, incarcérées) (following, après) (the, les) (uprisings, attaques) (in, du) (Burundi, Burundi) (in late, à la fin de l ' année) (1991, 1991) (still, encore à ce jour) (in, dans) (gaol, prisons) (Have been, ont été) (informed of, informées des) (charges, charges) (against, contre) (them, elles) (or, ou) (sent for trial, traduites en justice)

00042113

Issues under discussion include desertification , environmental management , data collection , training , maritime pollution and emergency response preparedness and waste management

Les questions abordées incluent la désertification , la gestion de l ' environnement , la collecte de données , la formation , la pollution marine et la faculté de faire face à des situations d ' urgence ainsi que la gestion des déchets

(Issues, questions) (under discussion, abordées) (include, incluent) (desertification, désertification) (environmental, environnement) (management, gestion) (data, données) (collection, collecte) (training, formation) (maritime, marine) (pollution, pollution) (and, et) (emergency, urgence) (response preparedness, la faculté de faire face à) (and, ainsi que) (waste, déchets) (management, gestion)

00044891

Does not the Commission consider that the seriousness of the situation calls for both the stepping up of Community aid and a strong impetus from the United Nations which coordinates aid to people in distress ? What does the Commission intend to do in cooperation with the ACP countries to promote the restoration of peace in Somalia and , initially , to ensure that emergency aid is commensurate with requirements ?

La Commission n ' estime - t - elle pas que la gravité de la situation appelle à la fois un renforcement de l ' aide de la Communauté et une très forte impulsion au niveau des Nations unies qui coordonnent l ' aide apportée aux populations en détresse ? Que compte faire la Commission en relation avec les pays d ' Afrique , des Caraïbes et du Pacifique (ACP) pour favoriser le retour de la paix dans ce pays et , dans l ' immédiat , pour assurer que l ' aide d ' urgence soit à la mesure des besoins ?

(the, La) (Commission, Commission) (consider, estime) (that, que) (the, la) (seriousness, gravité) (of, de) (the, la) (situation, situation) (calls for, appelle) (both, à la fois) (the, un) (stepping up, renforcement) (of, de) (Community, Communauté) (aid, aide) (and, et) (a, une) (strong, forte) (impetus, impulsion) (from, au niveau des) (United Nations, Nations unies) (which, qui) (coordinates, coordonnent) (aid, aide) (to, pour) (people, populations) (in distress, en détresse) (What, Que) (the, la) (Commission, Commission) (intend to, compte) (do, faire) (in cooperation with, en relation avec) (the, les) (ACP, Afrique des Caraïbes et du Pacifique ACP) (countries, pays) (to, pour) (promote, favoriser) (the, le) (restoration, retour) (of, de) (peace, paix) (in, dans)

(and, et) (initially, dans l ' immédiat) (to, pour) (ensure, assurer) (that, que) (emergency, urgence) (aid, aide) (is, soit) (commensurate with, à la mesure des) (requirements, besoins)

00047262

It is therefore essential for the EFTA countries concerned to be able to participate from the outset

C ' est pourquoi il est capital que les pays de l ' AELE concernés aient , dès le début , la possibilité de participer à l ' exécution de l ' accord

(It, il) (is, est) (therefore, C ' est pourquoi) (essential, capital) (the, les) (EFTA, AELE) (countries, pays) (concerned, concernés) (to be able to, aient la possibilité de) (participate, participer) (from the outset, dès le début)

00047326

The question arises whether the value of the supplies , works and services of all operational units have to be aggregated , or the value of each operational unit has to be taken separately for the purpose of determining whether they exceed the thresholds laid down in the directives on public procurement .

La question se pose de savoir si la valeur des fournitures , des travaux et des services de toutes les unités opérationnelles doit être agrégée ou si la valeur des contrats de chaque unité doit être prise séparément pour déterminer si elle dépasse la valeur des seuils prévus dans les directives relatives aux marchés publics .

(The question arises, La question se pose) (whether, si) (the, la) (value, valeur) (of the, des) (supplies, fournitures) (works, travaux) (and, et) (services, services) (of, des) (all, toutes) (operational, opérationnelles) (units, unités) (have to, doit) (be, être) (aggregated, agrégée) (or, ou) (the, la) (value, valeur) (of, de) (each, chaque) (unit, unité) (has to be, doit) (taken, prise) (separately, séparément) (for the purpose of, pour) (determining, déterminer) (whether, si) (they, elle) (exceed, dépasse) (the, la) (thresholds, seuils) (laid down, prévus) (in, dans) (the, les) (directives, directives) (on, relatives aux) (public, publics) (procurement, marchés)

00047864

- there has been a systematic attempt by the General Secretariat of the Council to make available to the press , at the pre - Council briefings , background notes giving an outline of the issues scheduled for debate in the Council

- lors des briefings susmentionnés , le Secrétariat général du Conseil s ' est efforcé systématiquement de mettre à la disposition de la presse des notes de synthèse qui présentent , dans leurs grandes lignes , les questions inscrites à l ' ordre du jour de la session du Conseil

(there has been a attempt to, s ' est efforcé de) (systematic, systématiquement) (the, le) (General Secretariat of the Council, Secrétariat général du Conseil) (make available to, mettre à la disposition de) (the, la) (press, presse) (at, lors des) (briefings, briefings) (background notes, notes de synthèse) (giving, présentent) (outline, grandes lignes) (the, les) (issues, questions) (scheduled for debate, inscrites à l ' ordre du jour) (Council, Conseil)

00048088

Is the Council aware that such a policy contradicts the letter and spirit of the resolutions adopted by Parliament for the protection of minorities and their cultural identities ?

2) Le Conseil n ' ignore - t - il pas que l ' instauration d ' une norme telle que celle - ci est en contradiction avec l ' esprit et la lettre des résolutions adoptées par le Parlement européen en faveur des minorités et de leur identité culturelle ?

(Is aware that, n ' ignore pas que) (the, Le) (Council, Conseil) (such, telle) (a, une) (policy, norme) (contradicts, est en contradiction avec) (the, la) (letter, lettre) (and, et) (spirit, esprit) (of the, des) (resolutions, résolutions) (adopted, adoptées) (by, par) (Parliament, Parlement) (for the protection of, en faveur des) (minorities, minorités) (and, et) (their, leur) (cultural, culturelle) (identities, identité)

00050032

In addition , the Commission would like to draw the attention of the Honourable Member to the fact that tree felling can be completely compatible with sound management of forest stands if they have reached maturity and their replacement is ensured .

Par ailleurs , la Commission voudrait attirer l ' attention de l ' honorable parlementaire sur le fait que des coupes d ' arbres peuvent être tout à fait conformes à une bonne gestion des peuplements forestiers s ' ils sont arrivés à maturité et que leur remplacement est assuré .

(In addition, Par ailleurs) (the, la) (Commission, Commission) (would like to, voudrait) (draw the attention to, attirer l ' attention sur) (of, de) (the, l ') (Honourable Member, honorable parlementaire) (the, le) (fact that, fait que) (tree felling, coupes d ' arbres) (can, peuvent) (be, être) (completely, tout à fait) (compatible with, conformes à) (sound, bonne) (management, gestion) (of, des) (forest stands, peuplements forestiers) (if, s ') (they, ils) (have reached maturity, sont arrivés à maturité) (and, et) (their, leur) (replacement, remplacement) (is, est) (ensured, assuré)

00050123

The external delegations to which the Honourable Member refers are delegations opened by the Commission in exercise of its independent rights to manage its own activities ¤ this is clear from the fact that they are called ` Delegation of the Commission of the European Communities ' in the headquarters agreements and letters of accreditation , and the basis is to be found in Article 30 (9) of the Single European Act

Les délégations extérieures auxquelles se réfère l ' honorable parlementaire sont des délégations ouvertes par la Commission dans le cadre de sa propre autonomie de gestion , ainsi qu ' il ressort de la dénomination « Délégation de la Commission des Communautés européennes » utilisée dans l ' accord de siège et la lettre de créance s ' appuyant en plus sur l ' Acte unique et des dispositions de l ' article 30 , paragraphe 9 de l ' Acte unique européen

(The, Les) (external, extérieures) (delegations, délégations) (to which, auxquelles) (the, l ') (Honourable Member, honorable parlementaire) (refers, se réfère) (are, sont) (delegations, délégations) (opened, ouvertes) (by, par) (the, la) (Commission, Commission) (in exercise of, dans le cadre de) (its, sa propre) (independent, autonomie) (to, de) (manage, gestion) (this is clear from the fact that, ainsi qu ' il ressort de) (called, dénomination) (Delegation of the Commission of the European Communities, Délégation de la Commission des Communautés européennes) (in, dans) (the, l ') (headquarters agreements, accord de siège) (and, et) (letters of accreditation, lettre de créance) (and, et) (Article, article) (30, 30) (9, 9) (of, de) (the, l ') (Single European Act, Acte unique européen)

00050386

Will the Commission fund a programme to investigate the adverse effects of aircraft emissions on the environment and humans ?

La Commission serait - elle disposée à financer un programme visant à évaluer les incidences négatives des gaz d ' échappement des aéronefs sur l ' homme et sur l ' écosystème ?

(Will, serait disposée à) (the, La) (Commission, Commission) (fund, financer) (a, un) (programme, programme) (to, visant à) (investigate, évaluer) (the, les) (adverse effects, incidences négatives) (of, des) (aircraft, aéronefs) (emissions, gaz d ' échappement) (on, sur) (the, l ') (environment, écosystème) (and, et) (humans, homme)

00052030

A group of experts from the Member States was convened by the Commission to report on the situation with regard to Community self - sufficiency in the supply of blood and blood products , and to devise measures to encourage the voluntary donation of blood and to attain self - sufficiency

Un groupe d ' experts des États membres a été réuni par la Commission pour faire rapport sur l ' état de l ' autosuffisance de la Communauté en sang et dérivés sanguins et pour élaborer des mesures de promotion du don du sang bénévole et des mesures visant à l ' autosuffisance

(A, Un) (group, groupe) (of, d ') (experts, experts) (from, des) (Member States, États membres) (was, a été) (convened, réuni) (by, par) (the, la) (Commission, Commission) (to report, pour faire rapport) (on, sur) (the, l ') (situation, état) (with regard to, de) (Community, Communauté) (self - sufficiency, autosuffisance) (in, en) (blood products, dérivés sanguins) (and, et) (blood, sang) (and, et) (to, pour) (devise, élaborer) (measures, mesures) (encourage, promotion) (the, du) (voluntary, bénévole) (donation of blood, don du sang) (and, et) (to attain, visant à) (self - sufficiency, autosuffisance)

00052980

Is the Commission aware that the UK Child Support Act (to be implemented in April 1993) discriminates against single mothers who refuse to name the father of their children for maintenance purposes ? Refusal to name the father could lead to a substantial reduction in social security benefits ¤ This could pose particular problems for women who have faced violence or other forms of abuse who may be unwilling to name the father .

La Commission sait - elle que la loi britanntique sur les allocations familiales , qui doit entrer en vigueur en avril 1993 , établit une discrimination au détriment des mères célibataires qui refusent d ' indiquer le nom du père de leurs enfants aux fins de pension alimentaire ? En effet , ce refus peut entraîner une réduction substantielle des prestations de sécurité sociale , ce qui est de nature à susciter des problèmes particuliers pour les femmes victimes de violence ou d ' autres formes d ' abus et ne souhaitant pas désigner le père .

(Is aware that, sait que) (the, La) (Commission, Commission) (the, la) (UK Act, loi britanntique) (Child Support, allocations familiales) (to be, qui doit) (implemented, entrer en vigueur) (in, en) (April, avril) (1993, 1993) (discriminates, établit une discrimination) (against, au détriment des) (single mothers, mères célibataires) (who, qui) (refuse to, refusent d ') (name, indiquer le nom) (the, du) (father, père) (of, de) (their, leurs) (children, enfants) (for purposes, aux fins de) (maintenance, pension alimentaire) (Refusal, refus) (name, désigner) (the, le) (father, père) (could, peut) (lead to, entraîner) (substantial, substantielle) (reduction in, réduction des) (social security benefits, prestations de sécurité sociale) (This could, ce qui est de nature à) (pose problems, susciter des problèmes) (particular, particuliers) (for, pour) (women, femmes) (who have faced, victimes de) (violence, violence) (or, ou) (other, autres) (forms, formes) (of, d ') (abuse, abus) (unwilling, ne souhaitant pas)

00053439

Does the Commission intend to ask that the Greek Tourist Board cease work on this hotel on the site of the Parliament of Psara ?

La Commission a - t - elle l ' intention de demander l ' arrêt de la construction de l ' hôtel de l ' EOT à l ' emplacement du Parlement de Psara ?

(the, La) (Commission, Commission) (intend to, a l ' intention de) (ask, demander) (Greek Tourist Board, EOT) (cease, arrêt) (work on, construction de) (hotel, hôtel) (on the site of, à l ' emplacement du) (Parliament of Psara, Parlement de Psara)

00054754

However , the Commission has not itself approved any project for the supply of shipping vessels ¤ or indeed for any other material or service which might be for military use .

Cela étant , la Commission en tant que telle n ' a approuvé aucun projet de fourniture de navires , ni d ' autres équipements ou de services , pouvant être utilisés à des fins militaires .

(However, Cela étant) (the, la) (Commission, Commission) (has approved, a approuvé) (not any, n’ aucun) (itself, en tant que telle) (project for, projet de) (supply, fourniture) (of, de) (shipping vessels, navires) (or, ou) (not any, ni) (other, autres) (material, équipements) (or, ou) (service, services) (might, pouvant) (be, être) (for, à des fins) (military, militaires) (use, utilisés)

00055821

On the question of cooperation between Community institutions , it was agreed at the Council meeting (Economic and Financial Affairs) of 23 November 1992 that in future the Council , Parliament and the Commission will act in concert to determine priorities for the annual programme of anti - fraud work .

En ce qui concerne la coopération entre les institutions communautaires , il y a lieu de noter qu ' à la suite du Conseil Ecofin du 23 novembre 1992 , le Conseil , le Parlement et la Commission établiront désormais de concert les priorités du programme de travail annuel relatif à la lutte contre la fraude .

(On the question of, En ce qui concerne) (cooperation, coopération) (between, entre) (Community, communautaires) (institutions, institutions) (at, à la suite du) (Council meeting Economic and Financial Affairs, Conseil Ecofin) (of, du) (23, 23) (November, novembre) (1992, 1992) (in future, désormais) (the, le) (Council, Conseil) (Parliament, Parlement) (and, et) (the, la) (Commission, Commission) (determine, établiront) (in concert, de concert) (priorities for, priorités du) (annual, annuel) (programme, programme) (of, relatif à) (anti - fraud work, lutte contre la fraude)

00056022

1 ¤ What view does the Commission take of holding strategic stocks of tinned meat available for food aid ?

1) La Commission ne pourrait - elle envisager de disposer d ' un stock stratégique de conserves de produits de viande de boeuf susceptible d ' être utilisé à des fins d ' aide alimentaire ?

(the, La) (Commission, Commission) (of, de) (holding, disposer) (strategic, stratégique) (stocks, stock) (tinned, conserves) (meat, viande) (available, susceptible d ' être utilisé) (for, à des fins d ') (food, alimentaire) (aid, aide)

00056851

This proposal also stipulated that all workers should be granted at least four weeks ' annual paid holiday .

Cette proposition prévoit en outre que tout travailleur bénéficie d ' un congé annuel payé d ' au moins quatre semaines .

(1, 1) (This, Cette) (proposal, proposition) (also, en outre) (stipulated, prévoit) (that, que) (all, tout) (workers, travailleur) (be granted, bénéficie) (at least, au moins) (four, quatre) (weeks, semaines) (annual, annuel) (paid holiday, congé payé)

00056902

In other western states majority voting exists in the context of a presidential republic or constitutional monarchy and Italy would therefore be the only country in Europe to change from proportional to first - past - the - post elections in the absence of a presidential republic , resulting in the abolition of individuals ' control through the representation of minority parties in parliament and transferring minority votes invalidated under the new law to the majority party , thereby infringing the principle of democracy and preventing millions of citizens from voting as they wish .

Par ailleurs , dans les autres États membres occidentaux , le système majoritaire coexiste avec le régime présidentiel ou la monarchie constitutionnelle et l ' Italie serait donc le seul État en Europe à remplacer la

représentation proportionnelle par le système majoritaire sans régime présidentiel ¤ Cela aurait pour conséquence , d ' une part , de retirer aux citoyens toute possibilité de contrôle à travers la représentation au Parlement des partis minoritaires , et ferait , d ' autre part , que les voix des minorités électorales , ignorées par la nouvelle loi , iront grossir les résultats du parti de la majorité en bafouant les principles de la démocratie et en empêchant de facto des millions de citoyens d ' être représentés .

(In, dans) (other, autres) (western states, États membres occidentaux) (majority voting, système majoritaire) (exists in the context of, coexiste avec) (presidential republic, régime présidentiel) (or, ou) (constitutional, constitutionnelle) (monarchy, monarchie) (and, et) (Italy, Italie) (would be, serait) (therefore, donc) (the, le) (only, seul) (country, État) (in, en) (Europe, Europe) (to, à) (change from to, remplacer par) (proportional, proportionnelle) (first - past - the - post elections, système majoritaire) (in the absence of, sans) (presidential republic, régime présidentiel) (resulting in, aurait pour conséquence de) (control, contrôle) (through, à travers) (the, la) (representation, représentation) (of, des) (minority, minorités) (parties, partis) (in, au) (parliament, Parlement) (and, et) (minority, minoritaires) (votes, voix) (invalidated, ignorées) (new, nouvelle) (law, loi) (majority, parti) (party, majorité) (infringing, bafouant) (the, les) (principle, principles) (democracy, démocratie) (and, et) (preventing from, empêchant d ') (millions, millions) (of, de) (citizens, citoyens)

00057504

- the new Convention on the protection of the marine environment of the Baltic Sea Area 1992 (Helsinki Convention) (the Community will soon accede to the 1974 Helsinki Convention) .

- la nouvelle convention sur la protection du milieu marin dans la zone de la mer Baltique de 1992 (convention d ' Helsinki) (la Commission adhèrera bientôt à la convention d ' Helsinki de 1974) .

(the, la) (new, nouvelle) (Convention, convention) (on, sur) (the, la) (protection, protection) (of the, du) (marine, marin) (environment, milieu) (of, de) (the, la) (Baltic Sea, mer Baltique) (Area, zone) (1992, 1992) (Helsinki Convention, convention d ' Helsinki) (the, la) (Community, Commission) (will accede to, adhèrera à) (soon, bientôt) (the, la) (1974, 1974) (Helsinki Convention, convention d ' Helsinki)

00058777

960 ¤ These shares corresponded to a capital injection of Bfrs 10,1 million (the company made a profit of Bfrs 57 million in 1984) .

Cet action correspondaient à l ' apport de capital de 10,1 million de francs (en 1984 , la société déclarait déjà un bénéfice de 57 millions de francs belges) ;

(These, Cet) (shares, action) (corresponded to, correspondaient à) (a, l’) (capital injection, apport de capital) (of, de) (Bfrs, francs) (10,1, 10,1) (million, million) (the, la) (company, société) (made, déclarait) (a, un) (profit, bénéfice) (of, de) (Bfrs, francs belges) (57, 57) (million, millions) (in, en) (1984, 1984)

00061718

The Commission was informed that Ente Ferrovie dello Stato had awarded to the company TAV a contract for the construction and operation of a high - speed rail network in Italy

La Commission a eu connaissance du fait que l ' Ente Ferrovie dello Stato a passé un contrat avec la société TAV ayant pour objet la construction et gestion économique d ' un réseau de chemins de fer à grande vitesse en Italie

(The, La) (Commission, Commission) (was informed that, a eu connaissance du fait que) (Ente Ferrovie dello Stato, Ente Ferrovie dello Stato) (had, a) (awarded a contract to, passé un contrat avec) (the, la) (company, société) (TAV, TAV) (for, ayant pour objet) (the, la) (construction, construction) (and, et) (operation, gestion économique) (of, d ') (a, un) (high - speed, à grande vitesse) (rail, chemins de fer) (network, réseau) (in, en) (Italy, Italie)

00062734

It is possible that they did not come from or even pass through Croatia , although they were presented at the Italian border with Croatian health certificates

Il se peut que ces animaux n ' aient pas été expédiés à partir de la Croatie , voire qu ' ils n ' aient même pas traversé ce pays , bien qu ' ils aient été présentés à la frontière italienne accompagnés de certificats sanitaires croates

(It is possible that, Il se peut que) (they, ils) (did not, n'aient pas) (or even, voire même) (pass through, traversé) (Croatia, pays) (although, bien qu ') (they, ils) (were, aient été) (presented, présentés) (at, à) (the, la) (Italian, italienne) (border, frontière) (with, accompagnés de) (Croatian, croates) (health certificates, certificats sanitaires)

00064550

In view of the importance attached by the Community to town planning and environmental matters and the fact that there is no provision for state funding programmes of this kind , will the Commission say how it views the idea of launching a new financial programme for the implementation of the award - winning project ?

Vu la priorité que donne la Communauté aux matières relatives à l ' urbanisme et à l ' environnement et le fait qu ' il n ' est pas prévu de financer pareil programme au moyen de ressources nationales , la Commission peut - elle indiquer ce qu ' elle pense de la création d ' un nouveau programme de financement pour la réalisation des projets de planification récompensés ?

(In view of, Vu) (the, la) (attached importance, donne priorité) (the, la) (Community, Communauté) (to, aux) (town planning, urbanisme) (and, et) (environmental, environnement) (matters, matières) (and, et) (the, le) (fact that, fait qu ') (there is no provision for, il n ' est pas prévu de) (state funding, financer au moyen de ressources nationales) (programmes, programme) (of this kind, pareil) (will, peut) (the, la) (Commission, Commission) (say, indiquer) (it, elle) (views the idea of, pense de) (launching, création) (a, un) (new, nouveau) (financial programme, programme de financement) (for, pour) (the, la) (implementation, réalisation) (of the, des) (award - winning, récompensés) (project, projets)

00065540

What steps will be taken by the Commission to encourage the development of the European Association of Sailing Federations , in view of the decision by the national sailing federations to cooperate in the organization of competitive events , thereby promoting Europe ' s exceptional coastline ?

Que compte faire la Commission des Communautés européennes pour encourager le développement de l ' Association européenne des fédérations de voile ? En effet , les Fédérations nationales de voile ont décidé d ' unir leurs efforts pour promouvoir ensemble des compétitions sportives , ce qui également valorise notre exceptionnel littoral européen .

(What steps will be taken, Que compte faire) (the, la) (Commission, Commission) (to, pour) (encourage, encourager) (the, le) (development, développement) (of, de) (the, l ') (European Association of Sailing Federations, Association européenne fédérations de voile) (in view of, En effet) (decision to, décidé d ') (the, les) (national, nationales) (sailing, voile) (federations, Fédérations) (cooperate in, unir leurs efforts pour) (competitive events, compétitions sportives) (thereby, également) (promoting, valorise) (Europe, européen) (exceptional, exceptionnel) (coastline, littoral)

00065615

Can this regulation be applied to all political data concerning the EC ? Can it , for instance , be applied to the harmonization of policy on asylum for political refugees so that a public debate in the European and national parliaments may be circumvented ?

2) Le règlement est - il applicable à toutes les matières politiques intéressant la Communauté ? Est - il , par exemple , applicable à l ' harmonisation de la politique du droit d ' asile adoptée à l ' égard des réfugiés politiques , qui pourrait ainsi échapper à un débat public au sein du Parlement européen et des parlements nationaux ?

(Can be applied to, est applicable à) (regulation, règlement) (all, toutes) (political, politiques) (data, matières) (concerning, intéressant) (the, la) (EC, Communauté) (Can be applied to, est applicable à) (it, il) (for instance, par exemple) (the, l ') (harmonization, harmonisation) (of, de) (policy on, politique du) (asylum, droit d ' asile) (for, à l ' égard des) (political, politiques) (refugees, réfugiés) (so that, ainsi) (a, un) (public, public) (debate, débat) (in, au sein du) (European and national parliaments, Parlement européen et des parlements nationaux) (may, pourrait) (be circumvented, échapper à)

00065791

The Commission ' s information visits programme has also been reinforced in recent months to enable it to focus on groups from Member States involved in ratification by referendum .

Les programmes de visites d ' information de la Commission ont également été renforcés au cours des derniers mois afin de permettre à la Commission de se concentrer sur certains groupes dans les États membres où la ratification était soumise à un référendum .

(The, Les) (Commission, Commission) (information, information) (visits, visites) (programme, programmes) (has been, ont été) (also, également) (reinforced, renforcés) (in recent months, au cours des derniers mois) (to, afin de) (enable to, permettre de) (focus on, se concentrer sur) (groups, groupes) (from, dans) (Member States, États membres) (ratification, ratification) (referendum, référendum)

00066141

Community aid to improve nuclear reactor safety in Eastern European countries

Aide communautaire pour l ' amélioration de la sécurité des réacteurs nucléaires en service dans les pays d ' Europe de l ' Est

(Community, communautaire) (aid, Aide) (to, pour) (improve, amélioration) (nuclear, nucléaires) (reactor, réacteurs) (safety, sécurité) (in, dans) (Eastern European countries, pays d ' Europe de l ' Est)

00066947

Joint answer to Written Questions Nos 665/93 , 1653/93 and 1722/93 given by Mr Matutes on behalf of the Commission (30 September 1993)

Réponse commune aux questions écrites n º 665/93 , n º 1653/93 et n º 1722/93 donnée par M . Matutes au nom de la Commission (30 septembre 1993)

(Joint answer, Réponse commune) (to, aux) (Written Questions, questions écrites) (Nos, n º) (665/93, 665/93) (1653/93, 1653/93) (and, et) (1722/93, 1722/93) (given, donnée) (by, par) (Mr Matutes, M . Matutes) (on behalf of, au nom de) (the, la) (Commission, Commission) (30, 30) (September, septembre) (1993, 1993)

00067643

In recent conferences held by various European Community film societies , mention has been made of the statement by the Assembly of film society federations stressing the need to create a new framework to facilitate , both legally and financially , the distribution throughout the Community Member States of films of cultural interest which have not been put on general release since they were made .

Des congrès , tenus récemment par un certain nombre de ciné - clubs de la Communauté européenne , ont donné lieu à l ' élaboration , par les fédérations de ces associations culturelles , d ' un procès - verbal mettant en lumière la nécessité de créer un nouveau cadre visant à faciliter , sur les plans législatif et financier , la

circulation parmi les pays de la Communauté européenne , de films présentant un intérêt culturel et qui , au bout d ' un certain temps suivant leur production , ne font pas l ' objet d ' une distribution commerciale dans les douze pays .

(recent, récemment) (conferences, congrès) (held, tenus) (by, par) (various, un certain nombre de) (European Community, Communauté européenne) (film societies, ciné - clubs) (the, un) (statement, procès - verbal) (by, par) (the, les) (of, de) (federations, fédérations) (society, associations) (stressing, mettant en lumière) (the, la) (need to, nécessité de) (create, créer) (a, un) (new, nouveau) (framework, cadre) (to, visant à) (facilitate, faciliter) (legally, législatif) (and, et) (financially, financier) (the, la) (distribution, circulation) (throughout, parmi) (the, les) (Community Member States, pays de la Communauté européenne) (de, of) (films, films) (of cultural interest, présentant un intérêt culturel) (which, qui)

A-VI Distributions des indices : couples correspondants *vs* couples quelconques

Les figures suivantes montrent l'évolution des distributions $y = p(indice \geq x)$ pour différentes tranches de fréquence :

$$f \leq 3,\ 3 < f \leq 10,\ 10 < f \leq 50,\ 50 < f \leq 500,\ 500 < f \leq 5000$$

où f est la fréquence la plus basse des deux unités comparée. Pour chaque figure, on a représenté deux distributions : « *Indice* » pour les couples d'unités quelconques et « *Indice corr* » pour les couples d'unités équivalentes.

figure 73 : évolution des distributions y = p(IM ≥ x) pour différentes tranches de fréquences (« IM » = couples quelconques « IM corr » = couples d'équivalents)

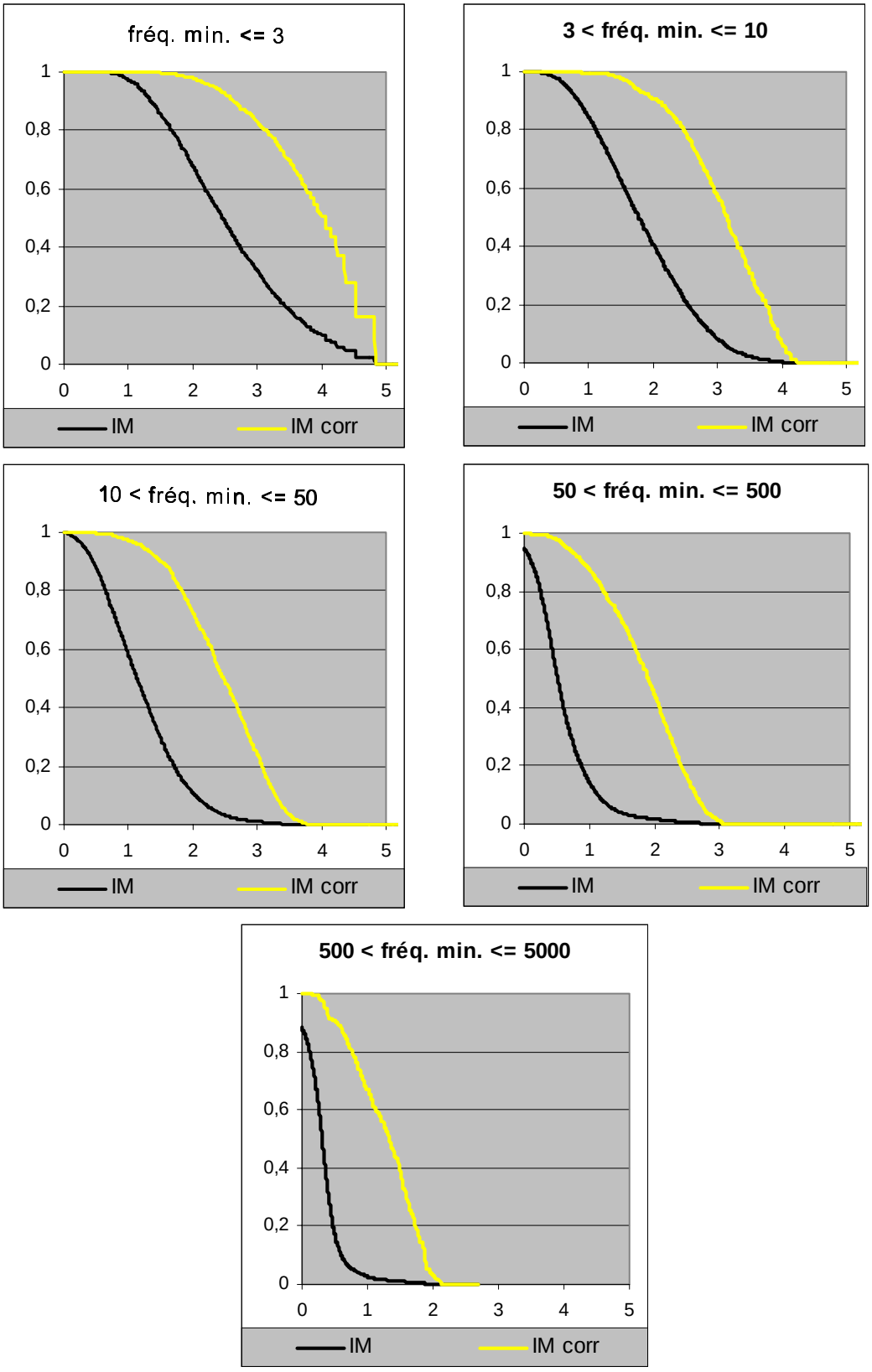


figure 74 : évolution des distributions y = p(TS ≥ x) pour différentes tranches de fréquences (« TS » = couples quelconques « TS corr » = couples d'équivalents)

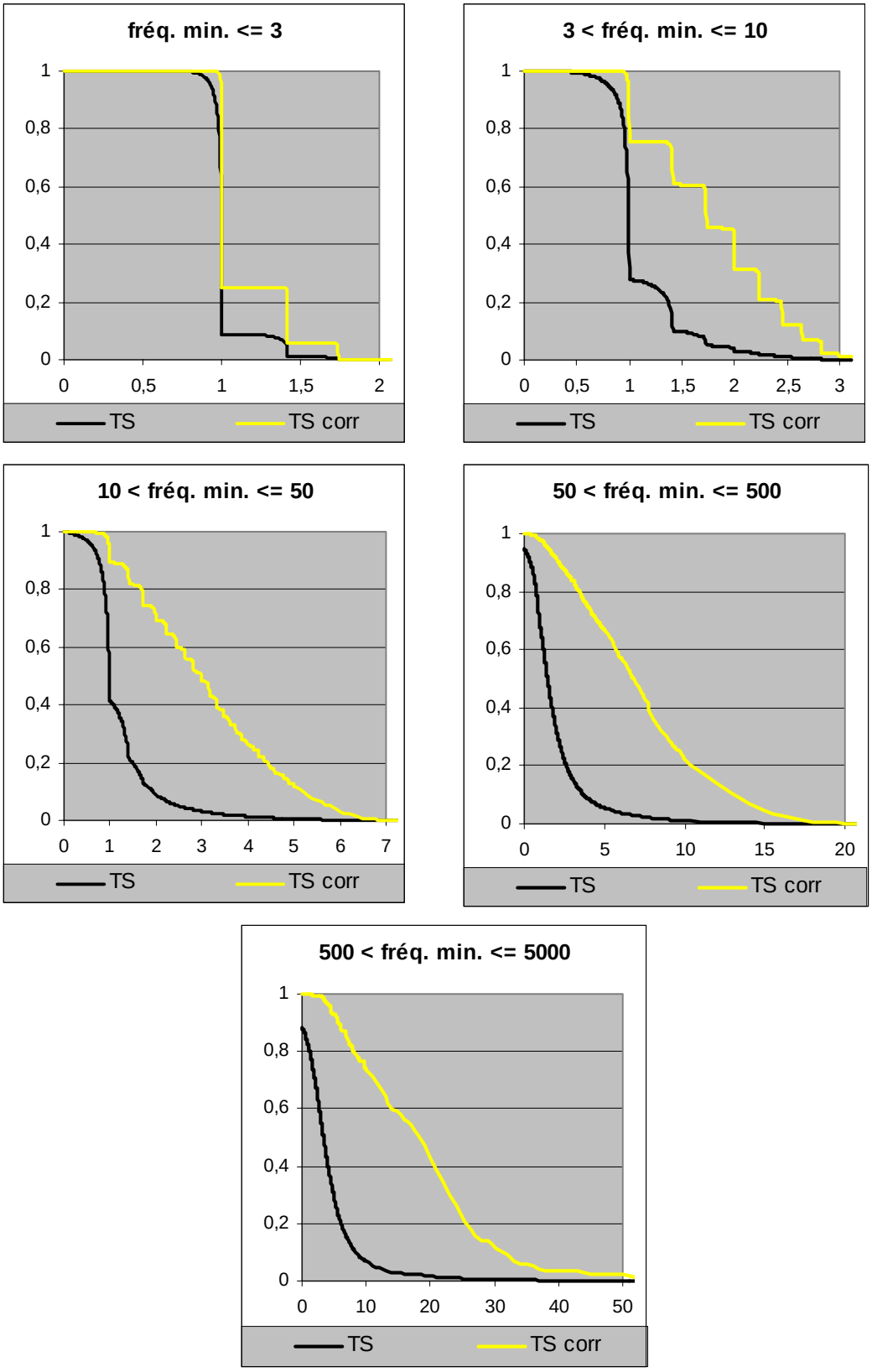


figure 75 : évolution des distributions y = p(RV ≥ x) pour différentes tranches de fréquences (« RV » =couples quelconques « RV corr » = couples d'équivalents)

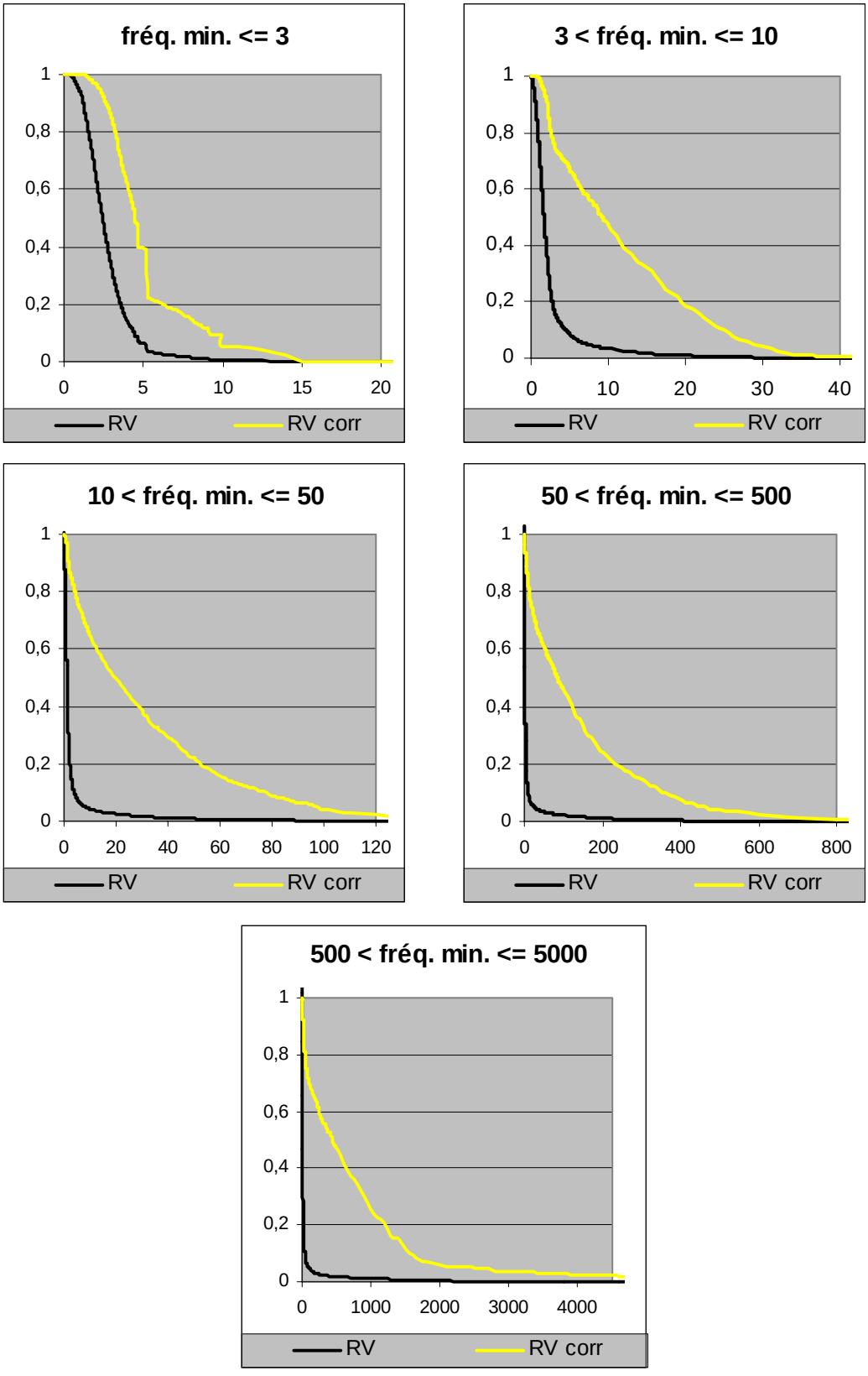


figure 76 : évolution des distributions y = p(CO ≥ x) pour différentes tranches de fréquences (« CO » =couples quelconques « CO corr » = couples d'équivalents)

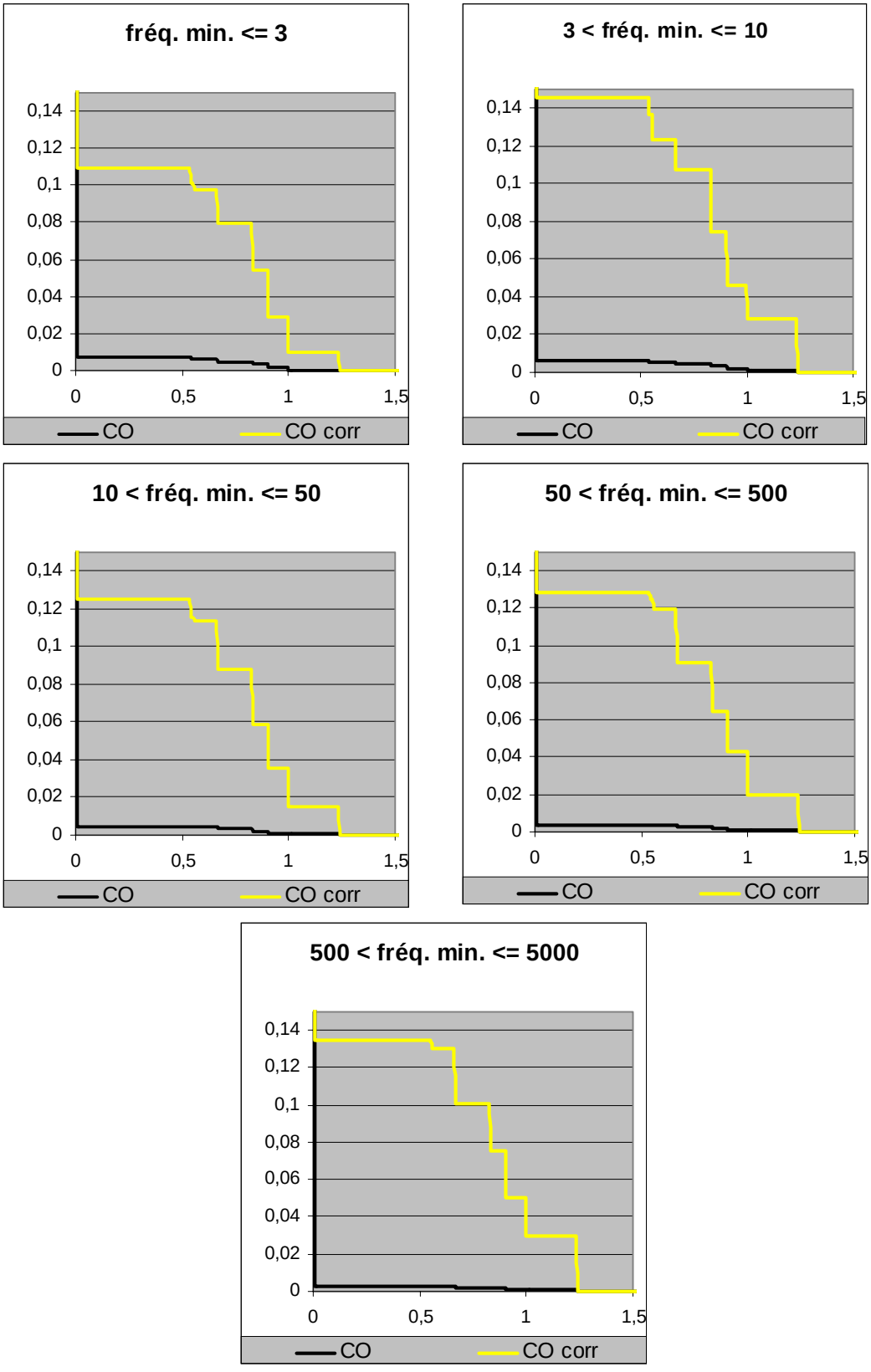


figure 77 : évolution des distributions y = p(PC ≥ x) pour différentes tranches de fréquences (« PC » =couples quelconques « PC corr » = couples d'équivalents)

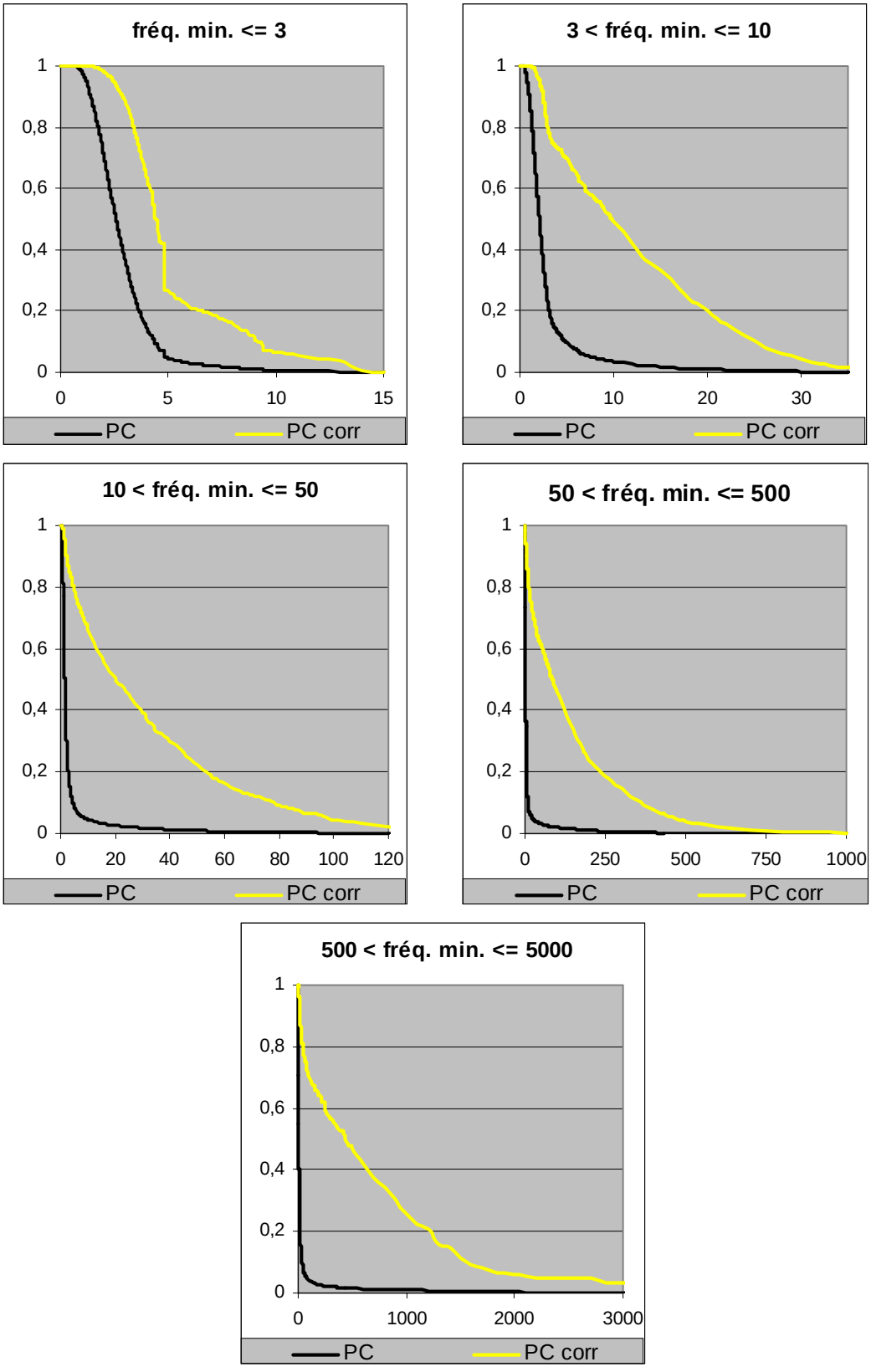
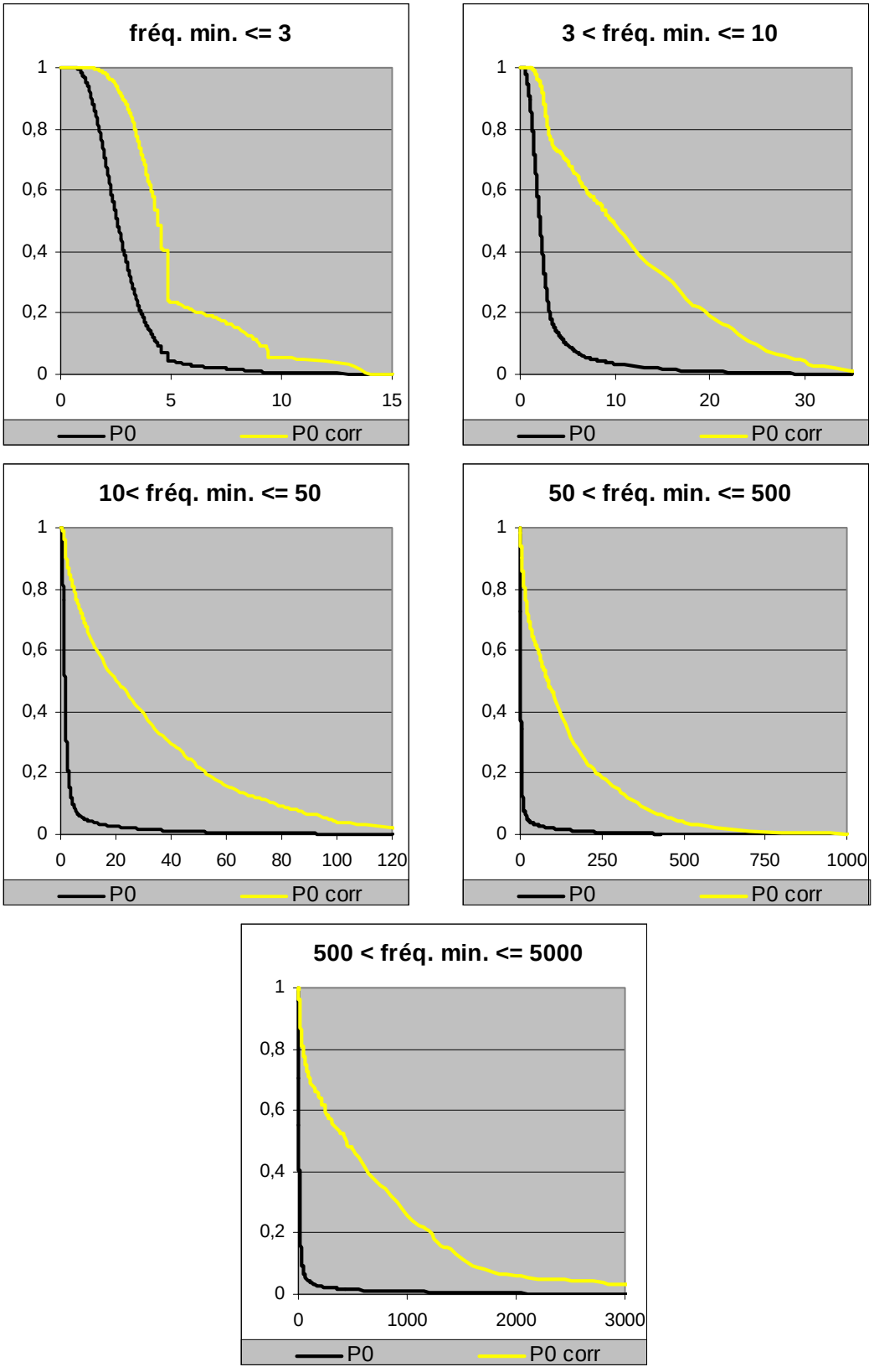


figure 78 : évolution des distributions y = p(P0 ≥ x) pour différentes tranches de fréquences (« P0 » =couples quelconques « P0 corr » = couples d'équivalents)



A-VII Résultats comparés des indices et des algorithmes

tableau 99 : résultats pour les différents indices avec l'algorithme AMAX

Indice	*Unités*	*P*	*R*	*F*
AL	formes	8,4 %	6,3 %	7,2 %
AL	lexies	6,8 %	5,7 %	6,2 %
AL	lemmes	7,5 %	6,0 %	6,7 %
COa	formes	33,7 %	25,3 %	28,9 %
COa	lexies	30,6 %	25,6 %	27,8 %
COa	lemmes	32,7 %	26,2 %	29,1 %
COb	formes	37,1 %	27,9 %	31,8 %
COb	lexies	33,6 %	28,1 %	30,6 %
COb	lemmes	35,5 %	28,4 %	31,6 %
IM	formes	40,8 %	30,7 %	35,0 %
IM	lexies	30,8 %	25,8 %	28,1 %
IM	lemmes	34,2 %	27,4 %	30,4 %
TS	formes	61,9 %	46,5 %	53,1 %
TS	lexies	58,7 %	49,1 %	53,4 %
TS	lemmes	62,1 %	49,8 %	55,3 %
RV	formes	67,1 %	50,4 %	57,6 %
RV	lexies	64,5 %	54,0 %	58,8 %
RV	lemmes	66,7 %	53,4 %	59,3 %
P0	formes	66,9 %	50,3 %	57,5 %
P0	lexies	64,2 %	53,7 %	58,5 %
P0	lemmes	66,5 %	53,3 %	59,2 %
PC	formes	67,7 %	50,9 %	58,1 %
PC	lexies	65,1 %	54,4 %	59,3 %
PC	lemmes	67,3 %	53,9 %	59,8 %

tableau 100 : résultats pour les différents indices avec l'algorithme ABIJ

Indice	*Unités*	*P*	*R*	*F*
AL	formes	8,4 %	6,1 %	7,1 %
AL	lexies	6,7 %	5,9 %	6,3 %
AL	lemmes	7,4 %	6,1 %	6,7 %
COa	formes	36,5 %	26,4 %	30,7 %
COa	lexies	33,3 %	27,1 %	29,9 %
COa	lemmes	36,2 %	27,9 %	31,5 %
COb	formes	39,9 %	28,9 %	33,5 %
COb	lexies	36,1 %	29,4 %	32,4 %
COb	lemmes	39,0 %	30,1 %	34,0 %
IM	formes	67,9 %	49,2 %	57,0 %
IM	lexies	66,8 %	54,4 %	60,0 %
IM	lemmes	69,6 %	53,7 %	60,6 %
TS	formes	69,6 %	50,4 %	58,5 %
TS	lexies	68,6 %	55,9 %	61,6 %
TS	lemmes	72,4 %	55,8 %	63,1 %
RV	formes	71,9 %	52,1 %	60,4 %
RV	lexies	71,1 %	57,9 %	63,8 %
RV	lemmes	74,1 %	57,1 %	64,5 %
P0	formes	71,6 %	51,9 %	60,2 %
P0	lexies	71,0 %	57,8 %	63,7 %
P0	lemmes	74,0 %	57,0 %	64,4 %
PC	formes	72,6 %	52,6 %	61,0 %
PC	lexies	72,2 %	58,8 %	64,8 %
PC	lemmes	75,0 %	57,8 %	65,3 %

Liste de mots outils et des mots très fréquents

– *Français*

un une Le Un Une le les l la La Les autre autres chaque Chaque dernière pas actuellement ailleurs assez aujourd Aujourd aussi Aussi autant avant bien Bien bientôt ci déjà delà dessus Deuxièmement effectivement également encore Enfin ensemble entretemps environ Environ hui ici longtemps maintenant Malheureusement même mieux moins non Non ores peu plus Plus presque souvent surtout tant très trop voire après Voici ainsi Ainsi alors Alors et afin Afin cependant désormais donc et/ou lorsqu lorsque mais malgré Néanmoins néanmoins ni Or ou parce pourquoi Pourquoi puisque si Si Toutefois toutefois outre ce ces cette Ce Ces cet Cet Cette n ne cinq cinquième deux dix dizaines douze Douze huit milliards milliers millions onze Premier premières quatre six trente trois à pour dans sur aux du au par À Au auprès avec chez contre Dans De depuis Depuis dès Du durant En entre grâce hors jusqu Jusqu lors Lors Par parmi Parmi Pour près quant Quant sans selon Selon sous Sous Sur vers comme Comme des d de Des leur leurs nos notre sa Sa ses son Son en qui que il qu ceci Cela celle celles Celles celui Celui ceux Comment comment dont duquel Elle elles eux II Il ils Ils je Je laquelle lequel lesquelles lesquels lui m on On où plusieurs Qu quand Que quel Quel Quelle Quelles quelles quels quoi se tel telle telles tels y aucun aucune Aucune certain certaine certaines Certaines certains chacun chacune quelque quelques tous tout toute toutes elle plupart est a sont été ont A ai aient ait aura auraient aurait auront avaient avait avoir ayant devait devra devraient devrait devront doit doivent dû due dues Est étaient était étant Étant être eu faire faisait faisant fait faite fallait fasse faut fera ferait font furent fut Ont peut peuvent plein pourra pourraient pourrait pouvant pu puisse puissent sera seraient serait seront soient soit va veulent voudrait t

– *Anglais*

The the a an An other last likely next great past better latter latest not No out up only more down already same further now well very far still yet even less fully soon together forward early clearly always Finally off too Further almost again just back often ago later ahead twice earliest hardly Far and as or also so if As whether but However therefore If however because thus while then Moreover i.e thereby although and/or Although e.g indeed Nevertheless Therefore moreover Thus otherwise Whether Because Whilst that this these This those These Those one two first third five six four One thousands eight Third ten Two Twelve First millions eleven sixth thousand thirds forty fourteenth to in by of on with for from at In under between into within against since than through about over On during above before after without According At For Under according To With towards throughout until outside near upon like despite around From via During beyond behind Of amongst instead Since which it its their What what It they there who where them when both how How his They There itself When he why themselves our Why Its ones None all any some each most many several another least much every few Such All Each be is are has will have been was does would should being were can Does may Is must Can do could Will able cannot did Has shall Would Are Have Should Did Had non etc. half Non

tableau 101 : résultats pour les différents indices avec l'algorithme ABIJ après suppression des mots outils

Indice	*Unités*	*P*	*R*	*F*
COb	formes	46,4 %	39,5 %	42,7 %
COb	lexies	42,1 %	42,8 %	42,4 %
COb	lemmes	44,2 %	43,8 %	44,0 %
IM	formes	70,8 %	60,3 %	65,1 %
IM	lexies	69,2 %	70,3 %	69,7 %
IM	lemmes	72,5 %	72,1 %	72,3 %
TS	formes	73,2 %	62,3 %	67,3 %
TS	lexies	73,1 %	74,1 %	73,6 %
TS	lemmes	76,2 %	75,6 %	75,9 %
RV	formes	75,9 %	64,6 %	69,8 %
RV	lexies	75,1 %	76,2 %	75,6 %
RV	lemmes	77,5 %	76,9 %	77,2 %
P0	formes	75,8 %	64,5 %	69,7 %
P0	lexies	75,0 %	76,0 %	75,5 %
P0	lemmes	77,5 %	76,9 %	77,2 %
PC	formes	76,8 %	65,3 %	70,6 %
PC	lexies	76,4 %	77,5 %	77,0 %
PC	lemmes	78,8 %	78,2 %	78,5 %

tableau 102 : proportion des couples erronés dus aux unités résiduelles (algorithme ABIJ)

tâche FS	***P***	***R***	***F***	***Proportion de couples avec au moins une unité résiduelle***		***Proportion de couples avec deux unités résiduelles***	
				p_{res1} empirique	*p_{res1} théorique*	p_{res2} empirique	*p_{res2} théorique*
CO	39,9 %	28,9 %	33,5 %	35,9 %	*45,3 %*	5,8 %	*8,6 %*
IM	67,9 %	49,2 %	57,0 %	54,1 %	*58,3 %*	20,0 %	*15,7 %*
TS	69,6 %	50,4 %	58,5 %	59,1 %	*59,4 %*	21,6 %	*16,5 %*
RV	71,9 %	52,1 %	60,4 %	61,2 %	*60,7 %*	24,1 %	*17,5 %*
P0	71,6 %	51,9 %	60,2 %	60,8 %	*60,6 %*	23,9 %	*17,4 %*
PC	72,6 %	52,6 %	61,0 %	62,7 %	*61,4 %*	24,9 %	*18,0 %*

tâche LEX	***P***	***R***	***F***	***Proportion de couples avec au moins une unité résiduelle***		***Proportion de couples avec deux unités résiduelles***	
				p_{res1} empirique	*p_{res1} théorique*	p_{res2} empirique	*p_{res2} théorique*
CO	36,1 %	29,4 %	32,4 %	38,8 %	*51,7 %*	7,5 %	*11,5 %*
IM	66,8 %	54,4 %	60,0 %	54,6 %	*66,7 %*	24,9 %	*21,8 %*
TS	68,6 %	55,9 %	61,6 %	63,4 %	*68,3 %*	27,2 %	*23,1 %*
RV	71,1 %	57,9 %	63,8 %	64,2 %	*69,7 %*	30,2 %	*24,7 %*
P0	71,0 %	57,8 %	63,7 %	64,7 %	*69,7 %*	30,1 %	*24,7 %*
PC	72,2 %	58,8 %	64,8 %	67,3 %	*71,0 %*	31,6 %	*25,6 %*

tâche LEM	***P***	***R***	***F***	***Proportion de couples avec au moins une unité résiduelle***		***Proportion de couples avec deux unités résiduelles***	
				p_{res1} empirique	*p_{res1} théorique*	p_{res2} empirique	*p_{res2} théorique*
CO	39,0 %	30,1 %	34,0 %	37,1 %	*51,8 %*	7,5 %	*11,7 %*
IM	69,6 %	53,7 %	60,6 %	56,5 %	*67,3 %*	25,4 %	*22,1 %*
TS	72,4 %	55,8 %	63,1 %	65,7 %	*69,4 %*	29,4 %	*23,8 %*
RV	74,1 %	57,1 %	64,5 %	66,1 %	*70,4 %*	32,4 %	*25,0 %*
P0	74,0 %	57,0 %	64,4 %	66,1 %	*70,4 %*	32,4 %	*24,9 %*
PC	75,0 %	57,8 %	65,3 %	68,7 %	*71,4 %*	33,8 %	*25,7 %*

tableau 103 : résultats pour les différents indices avec l'algorithme ABIJ, après suppression des unités résiduelles (tâche LEM)

Indice	*P*	*R*	*F*
CO	46,4 %	31,4 %	37,4 %
IM	81,8 %	55,3 %	66,0 %
TS	86,2 %	58,2 %	69,5 %
RV	87,3 %	58,9 %	70,4 %
P0	87,3 %	58,9 %	70,3 %
PC	88,6 %	59,8 %	71,4 %

A-VIII Filtrages des résultats

tableau 104 : filtrage relatif des résultats de la tâche LEX pour les différents indices et algorithmes (NB : Proportion = 1 équivaut à ne pas filtrer)

Indice	*Proportion*	*AMAX*			*ABIJ*		
		P	***R***	***F***	***P***	***R***	***F***
CO	1	33,6 %	28,1 %	30,6 %	36,1 %	29,4 %	32,4 %
	0,8	37,6 %	27,7 %	31,9 %	40,3 %	28,9 %	33,7 %
	0,6	46,8 %	26,5 %	33,9 %	48,9 %	27,2 %	34,9 %
	0,4	61,2 %	24,9 %	35,4 %	63,5 %	25,2 %	36,1 %
	0,2	80,1 %	18,9 %	30,6 %	81,5 %	18,9 %	30,7 %
IM	1	30,8 %	25,8 %	28,1 %	66,8 %	54,4 %	60,0 %
	0,8	33,3 %	24,6 %	28,3 %	72,7 %	52,2 %	60,8 %
	0,6	36,2 %	20,6 %	26,2 %	75,7 %	42,1 %	54,1 %
	0,4	39,8 %	16,2 %	23,0 %	73,8 %	29,4 %	42,0 %
	0,2	45,6 %	10,8 %	17,4 %	69,3 %	16,1 %	26,1 %
TS	1	58,7 %	49,1 %	53,4 %	68,6 %	55,9 %	61,6 %
	0,8	62,6 %	46,1 %	53,1 %	73,5 %	52,7 %	61,4 %
	0,6	67,8 %	38,5 %	49,1 %	78,1 %	43,4 %	55,8 %
	0,4	74,1 %	30,2 %	42,9 %	82,0 %	32,6 %	46,7 %
	0,2	80,0 %	18,9 %	30,5 %	84,3 %	19,6 %	31,8 %
RV	1	64,5 %	54,0 %	58,8 %	71,1 %	57,9 %	63,8 %
	0,8	69,7 %	51,3 %	59,1 %	77,2 %	55,4 %	64,5 %
	0,6	76,5 %	43,4 %	55,4 %	83,3 %	46,3 %	59,5 %
	0,4	83,0 %	33,8 %	48,0 %	87,8 %	34,9 %	50,0 %
	0,2	88,0 %	20,8 %	33,6 %	90,2 %	21,0 %	34,0 %
P0	1	64,2 %	53,7 %	58,5 %	71,0 %	57,8 %	63,7 %
	0,8	69,2 %	51,0 %	58,7 %	77,1 %	55,3 %	64,4 %
	0,6	76,2 %	43,3 %	55,2 %	83,2 %	46,2 %	59,4 %
	0,4	82,8 %	33,7 %	47,9 %	87,8 %	34,9 %	49,9 %
	0,2	88,0 %	20,8 %	33,6 %	90,1 %	20,9 %	34,0 %
PC	1	65,1 %	54,4 %	59,3 %	72,2 %	58,8 %	64,8 %
	0,8	70,1 %	51,6 %	59,5 %	78,1 %	56,1 %	65,3 %
	0,6	76,7 %	43,5 %	55,5 %	83,8 %	46,5 %	59,8 %
	0,4	82,9 %	33,8 %	48,0 %	87,9 %	35,0 %	50,0 %
	0,2	88,0 %	20,8 %	33,6 %	90,2 %	21,0 %	34,0 %

tableau 105 : filtrage différentiel des résultats de la tâche LEX pour les différents indices et algorithmes (NB : rapport = 1 équivaut à ne pas filtrer).

Indice	*rapport seuil*	*AMAX*			*ABIJ*		
		P	***R***	***F***	***P***	***R***	***F***
CO	1	33,6 %	28,1 %	30,6 %	36,1 %	29,4 %	32,4 %
	1,05	83,8 %	23,4 %	36,6 %	86,9 %	23,3 %	36,8 %
	1,2	84,0 %	23,2 %	36,3 %	86,9 %	23,1 %	36,5 %
	1,5	85,2 %	22,5 %	35,6 %	87,7 %	22,4 %	35,7 %
	2	86,1 %	22,0 %	35,0 %	88,1 %	22,0 %	35,2 %
	2,5	86,5 %	21,4 %	34,3 %	88,1 %	21,4 %	34,4 %
	3	86,7 %	20,9 %	33,7 %	87,8 %	20,9 %	33,8 %
	4	86,9 %	20,4 %	33,0 %	87,7 %	20,4 %	33,1 %
IM	1	30,8 %	25,8 %	28,1 %	66,8 %	54,4 %	60,0 %
	1,05	45,1 %	17,0 %	24,6 %	70,2 %	25,6 %	37,5 %
	1,2	71,4 %	7,9 %	14,3 %	75,0 %	9,3 %	16,6 %
	1,5	92,9 %	3,5 %	6,7 %	88,0 %	3,6 %	7,0 %
	2	98,0 %	1,5 %	2,9 %	94,8 %	1,5 %	3,0 %
	2,5	0,0 %	0,8 %	0,0 %	0,0 %	0,8 %	0,0 %
	3	0,0 %	0,6 %	0,0 %	0,0 %	0,6 %	0,0 %
	4	0,0 %	0,4 %	0,0 %	0,0 %	0,4 %	0,0 %
TS	1	58.7 %	49.1 %	53.4 %	68.6 %	55.9 %	61.6 %
	1,05	66,0 %	40,2 %	50,0 %	77,3 %	43,3 %	55,5 %
	1,2	76,6 %	34,8 %	47,9 %	86,3 %	35,8 %	50,6 %
	1,5	88,2 %	27,0 %	41,4 %	93,7 %	27,2 %	42,1 %
	2	93,8 %	20,0 %	33,0 %	96,1 %	20,0 %	33,2 %
	2,5	95,4 %	14,4 %	25,1 %	97,0 %	14,4 %	25,1 %
	3	96,0 %	10,2 %	18,5 %	97,5 %	10,2 %	18,5 %
	4	96,6 %	5,2 %	9,8 %	97,5 %	5,2 %	9,8 %
RV	1	64.5 %	54.0 %	58.8 %	71.1 %	57.9 %	63.8 %
	1,05	70,7 %	49,2 %	58,0 %	77,8 %	52,1 %	62,4 %
	1,2	77,0 %	46,2 %	57,8 %	83,5 %	47,5 %	60,5 %
	1,5	83,4 %	42,6 %	56,4 %	89,2 %	43,0 %	58,0 %
	2	87,9 %	38,2 %	53,3 %	92,5 %	38,3 %	54,2 %
	2,5	90,6 %	34,9 %	50,4 %	94,2 %	34,9 %	50,9 %
	3	92,5 %	32,8 %	48,5 %	95,1 %	32,8 %	48,8 %
	4	94,3 %	29,3 %	44,8 %	96,0 %	29,3 %	44,9 %
P0	1	64,2 %	53,7 %	58,5 %	71,0 %	57,8 %	63,7 %
	1,05	70,9 %	48,9 %	57,8 %	77,8 %	51,5 %	61,9 %
	1,2	78,1 %	45,5 %	57,5 %	84,9 %	46,5 %	60,1 %
	1,5	84,9 %	41,7 %	56,0 %	90,4 %	42,1 %	57,4 %
	2	88,9 %	37,2 %	52,4 %	92,9 %	37,2 %	53,1 %
	2,5	91,6 %	33,8 %	49,4 %	94,5 %	33,8 %	49,8 %
	3	93,3 %	31,6 %	47,2 %	95,7 %	31,5 %	47,4 %
	4	94,8 %	28,0 %	43,2 %	96,3 %	27,9 %	43,3 %
PC	1	65,1 %	54,4 %	59,3 %	72,2 %	58,8 %	64,8 %
	1,05	71,3 %	49,8 %	58,7 %	78,5 %	52,6 %	63,0 %
	1,2	78,4 %	46,2 %	58,2 %	85,1 %	47,2 %	60,7 %
	1,5	85,0 %	42,1 %	56,3 %	90,4 %	42,4 %	57,7 %
	2	88,9 %	37,5 %	52,7 %	92,9 %	37,5 %	53,4 %
	2,5	91,7 %	34,0 %	49,6 %	94,5 %	34,0 %	50,0 %
	3	93,3 %	31,7 %	47,3 %	95,7 %	31,6 %	47,6 %
	4	94,9 %	28,1 %	43,3 %	96,3 %	28,1 %	43,5 %

tableau 106 : filtrage absolu des résultats de la tâche LEX pour les différents indices et algorithmes (NB : coefficient = 0 équivaut à ne pas filtrer).

Indice	*Coefficient du seuil*	*AMAX*			*ABIJ*		
		P	***R***	***F***	***P***	***R***	***F***
CO	0	33,6 %	28,1 %	30,6 %	36,1 %	29,4 %	32,4 %
	0,25	33,6 %	28,1 %	30,6 %	36,1 %	29,4 %	32,4 %
	0,5	33,6 %	28,1 %	30,6 %	36,1 %	29,4 %	32,4 %
	1	79,1 %	24,6 %	37,5 %	83,9 %	24,4 %	37,9 %
	2,5	79,1 %	24,6 %	37,5 %	83,9 %	24,4 %	37,9 %
	4	79,1 %	24,6 %	37,5 %	83,9 %	24,4 %	37,9 %
	6	79,1 %	24,6 %	37,5 %	83,9 %	24,4 %	37,9 %
	10	79,1 %	24,6 %	37,5 %	83,9 %	24,4 %	37,9 %
IM	0	30,8 %	25,8 %	28,1 %	66,8 %	54,4 %	60,0 %
	0,25	30,9 %	25,8 %	28,1 %	68,0 %	54,4 %	60,5 %
	0,5	31,0 %	25,8 %	28,2 %	71,7 %	53,9 %	61,6 %
	1	31,9 %	23,6 %	27,1 %	76,7 %	46,7 %	58,1 %
	2,5	32,4 %	5,4 %	9,2 %	55,6 %	6,9 %	12,2 %
	4	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %
	6	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %
	10	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %
TS	0	58,7 %	49,1 %	53,4 %	68,6 %	55,9 %	61,6 %
	0,25	58,7 %	49,1 %	53,4 %	68,7 %	55,9 %	61,6 %
	0,5	58,7 %	49,1 %	53,4 %	68,8 %	55,9 %	61,7 %
	1	62,5 %	46,1 %	53,1 %	76,6 %	50,7 %	61,0 %
	2,5	71,0 %	36,8 %	48,5 %	82,2 %	38,2 %	52,2 %
	4	75,9 %	30,4 %	43,4 %	84,6 %	31,0 %	45,3 %
	6	80,2 %	23,9 %	36,8 %	86,2 %	24,2 %	37,8 %
	10	85,2 %	15,7 %	26,5 %	88,6 %	15,9 %	26,9 %
RV	0	64,5 %	54,0 %	58,8 %	71,1 %	57,9 %	63,8 %
	0,25	64,8 %	53,9 %	58,9 %	74,4 %	57,4 %	64,8 %
	0,5	68,7 %	52,8 %	59,7 %	79,5 %	55,1 %	65,1 %
	1	74,4 %	48,5 %	58,7 %	84,4 %	50,1 %	62,9 %
	2,5	79,4 %	41,9 %	54,9 %	87,2 %	42,8 %	57,5 %
	4	81,7 %	38,1 %	52,0 %	88,5 %	38,8 %	54,0 %
	6	84,2 %	35,2 %	49,7 %	89,8 %	35,9 %	51,3 %
	10	86,4 %	31,3 %	45,9 %	90,8 %	31,8 %	47,1 %
P0	0	64,2 %	53,7 %	58,5 %	71,0 %	57,8 %	63,7 %
	0,25	64,3 %	53,7 %	58,5 %	72,1 %	57,7 %	64,1 %
	0,5	67,7 %	52,8 %	59,3 %	78,6 %	55,4 %	65,0 %
	1	74,5 %	47,6 %	58,1 %	84,5 %	49,2 %	62,2 %
	2,5	80,0 %	40,6 %	53,9 %	87,5 %	41,5 %	56,3 %
	4	82,7 %	37,0 %	51,1 %	89,0 %	37,7 %	53,0 %
	6	85,0 %	33,7 %	48,2 %	90,2 %	34,3 %	49,7 %
	10	86,9 %	29,7 %	44,3 %	91,0 %	30,2 %	45,3 %
PC	0	65,1 %	54,4 %	59,3 %	72,2 %	58,8 %	64,8 %
	0,25	65,1 %	54,4 %	59,3 %	73,2 %	58,7 %	65,2 %
	0,5	68,6 %	53,5 %	60,1 %	79,7 %	56,3 %	66,0 %
	1	74,7 %	48,2 %	58,6 %	84,6 %	49,8 %	62,7 %
	2,5	80,0 %	40,7 %	54,0 %	87,5 %	41,6 %	56,4 %
	4	82,8 %	37,1 %	51,2 %	89,0 %	37,8 %	53,1 %
	6	85,0 %	33,7 %	48,2 %	90,2 %	34,3 %	49,7 %
	10	86,9 %	29,7 %	44,3 %	91,0 %	30,2 %	45,3 %

figure 79 : résultats des indices avec le filtrage relatif

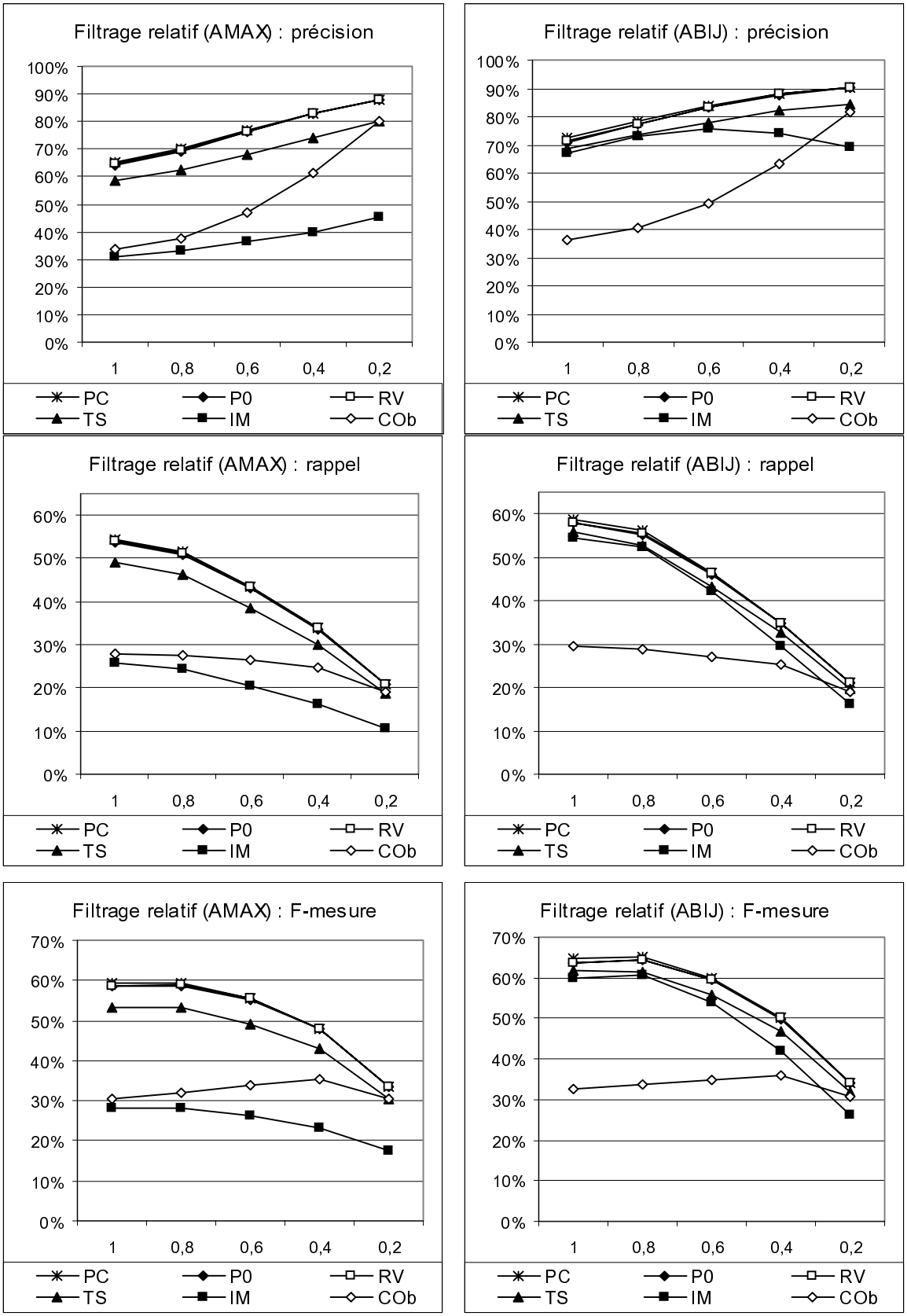


figure 80 : résultats des indices avec le filtrage absolu

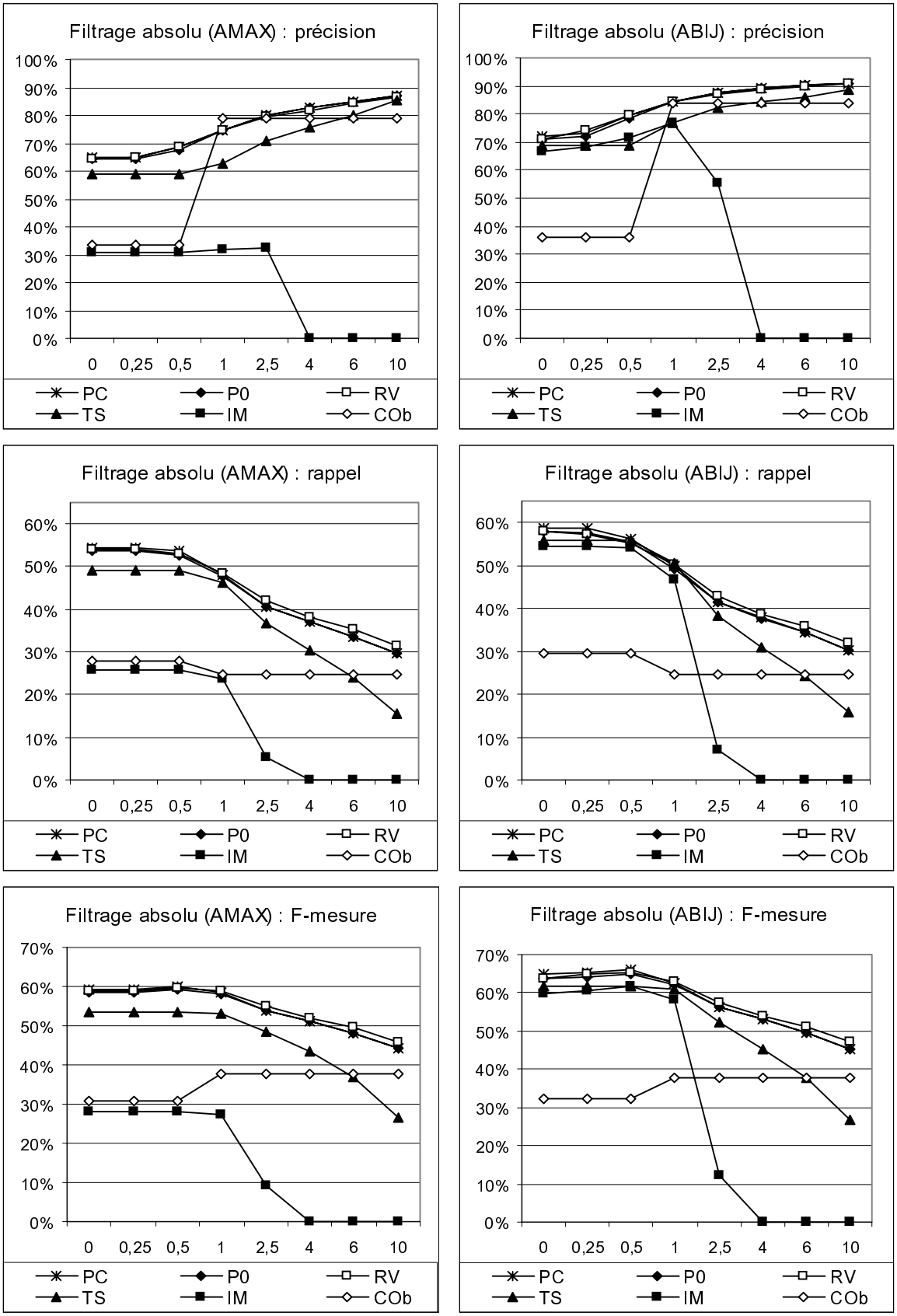
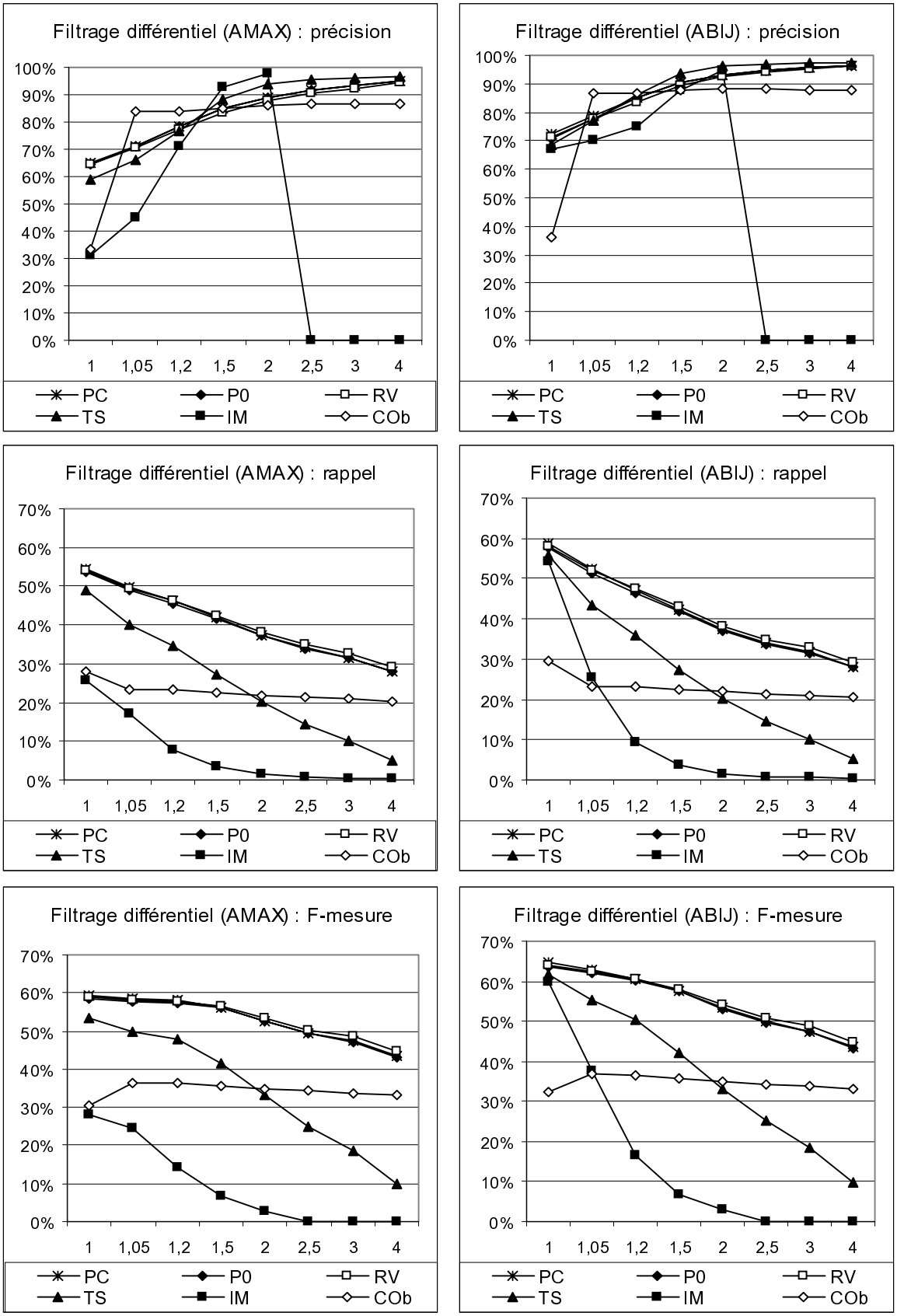


figure 81 : résultats des indices avec le filtrage différentiel



A-IX Evolution des résultats avec la taille du corpus d'apprentissage (algorithme ABIJ, extraction LEX)

– *Le corpus d'apprentissage incluant le corpus d'évaluation :*

tableau 107 : évolution de la précision avec la taille du corpus d'apprentissage

Taille	*IM*	*TS*	*RV*	*P0*	*PC*
767	37,8 %	42,9 %	43,5 %	43,5 %	54,3 %
1 990	44,4 %	49,9 %	50,4 %	50,3 %	58,2 %
4 502	51,2 %	56,0 %	57,1 %	57,1 %	62,6 %
8 181	55,6 %	59,6 %	61,0 %	60,9 %	65,2 %
15 528	59,0 %	63,2 %	64,7 %	64,8 %	68,0 %
30 238	63,1 %	65,6 %	67,7 %	67,6 %	69,7 %
49 780	65,6 %	68,2 %	70,1 %	70,1 %	71,7 %
69 160	66,8 %	68,6 %	71,1 %	71,0 %	72,2 %

tableau 108 : évolution du rappel avec la taille du corpus d'apprentissage

Taille	*IM*	*TS*	*RV*	*P0*	*PC*
767	30,7 %	34,8 %	35,4 %	35,3 %	44,2 %
1 990	36,1 %	40,6 %	41,0 %	40,9 %	47,4 %
4 502	41,7 %	45,5 %	46,4 %	46,5 %	51,0 %
8 181	45,3 %	48,5 %	49,7 %	49,6 %	53,0 %
15 528	48,0 %	51,4 %	52,7 %	52,8 %	55,4 %
30 238	51,4 %	53,4 %	55,1 %	55,1 %	56,7 %
49 780	53,5 %	55,5 %	57,1 %	57,0 %	58,4 %
69 160	54,4 %	55,9 %	57,9 %	57,8 %	58,8 %

tableau 109 : évolution de la F-mesure avec la taille du corpus d'apprentissage

Taille	*IM*	*TS*	*RV*	*P0*	*PC*
767	33,9 %	38,4 %	39,1 %	39,0 %	48,7 %
1 990	39,9 %	44,7 %	45,2 %	45,1 %	52,2 %
4 502	46,0 %	50,2 %	51,2 %	51,2 %	56,2 %
8 181	49,9 %	53,5 %	54,7 %	54,7 %	58,5 %
15 528	52,9 %	56,7 %	58,1 %	58,2 %	61,1 %
30 238	56,7 %	58,9 %	60,7 %	60,7 %	62,6 %
49 780	58,9 %	61,2 %	63,0 %	62,9 %	64,3 %
69 160	60,0 %	61,6 %	63,8 %	63,7 %	64,8 %

– *Le corpus d'apprentissage excluant le corpus d'évaluation :*

tableau 110 : évolution de la précision avec la taille du corpus d'apprentissage

Taille	*IM*	*TS*	*RV*	*P0*	*PC*
629	59,3 %	66,8 %	66,5 %	66,6 %	51,2 %
1833	61,6 %	68,7 %	68,6 %	68,9 %	56,7 %
3616	63,3 %	70,2 %	70,4 %	70,7 %	59,9 %
7334	65,1 %	71,2 %	71,6 %	72,0 %	63,2 %
14622	66,5 %	71,9 %	72,4 %	72,7 %	66,2 %
29414	68,5 %	72,7 %	73,7 %	73,8 %	68,3 %
48748	69,5 %	73,0 %	74,2 %	74,3 %	69,8 %
68393	70,1 %	72,9 %	74,5 %	74,5 %	70,5 %

tableau 111 : évolution du rappel avec la taille du corpus d'apprentissage

Taille	*IM*	*TS*	*RV*	*P0*	*PC*
629	24,7 %	26,5 %	26,7 %	26,6 %	41,8 %
1833	32,7 %	34,2 %	34,8 %	34,7 %	46,3 %
3616	37,4 %	38,9 %	39,6 %	39,5 %	48,9 %
7334	41,8 %	42,8 %	43,7 %	43,8 %	51,6 %
14622	45,9 %	46,8 %	47,7 %	47,7 %	54,1 %
29414	49,8 %	50,2 %	51,3 %	51,2 %	55,7 %
48748	51,8 %	52,1 %	53,3 %	53,2 %	56,9 %
68393	53,0 %	53,0 %	54,5 %	54,4 %	57,6 %

tableau 112 : évolution de la F-mesure avec la taille du corpus d'apprentissage

Taille	*IM*	*TS*	*RV*	*P0*	*PC*
629	34,9 %	37,9 %	38,1 %	38,0 %	46,0 %
1833	42,7 %	45,7 %	46,1 %	46,1 %	51,0 %
3616	47,0 %	50,0 %	50,7 %	50,7 %	53,8 %
7334	50,9 %	53,5 %	54,3 %	54,4 %	56,8 %
14622	54,3 %	56,7 %	57,5 %	57,6 %	59,5 %
29414	57,7 %	59,4 %	60,5 %	60,5 %	61,4 %
48748	59,3 %	60,8 %	62,1 %	62,0 %	62,7 %
68393	60,4 %	61,4 %	62,9 %	62,9 %	63,4 %

A-X Résultats par tranches de fréquence

tableau 113 : résultats de l'indice CO par tranches de fréquence (LEM avec ABIJ)

Tranche	*Anglais*			*Français*		
	Précision	***Rappel***	***F-mesure***	***Précision***	***Rappel***	***F-mesure***
1	39,5 %	40,6 %	40,1 %	33,4 %	33,4 %	33,4 %
2	40,1 %	41,5 %	40,8 %	37,8 %	38,3 %	38,1 %
3	33,8 %	34,2 %	34,0 %	39,8 %	40,6 %	40,2 %
4	31,8 %	33,1 %	32,4 %	29,6 %	30,9 %	30,3 %
5	36,7 %	37,5 %	37,1 %	32,4 %	33,9 %	33,2 %
6	36,2 %	37,5 %	36,8 %	34,4 %	36,8 %	35,6 %
7	38,7 %	40,5 %	39,5 %	38,0 %	38,9 %	38,5 %
8	37,9 %	40,8 %	39,3 %	36,5 %	38,0 %	37,3 %
9	36,1 %	37,5 %	36,8 %	35,3 %	38,4 %	36,8 %
10	35,5 %	41,1 %	38,1 %	37,5 %	42,4 %	39,8 %

tableau 114 : résultats de l'indice IM par tranches de fréquence (LEM avec ABIJ)

Tranche	*Anglais*			*Français*		
	Précision	***Rappel***	***F-mesure***	***Précision***	***Rappel***	***F-mesure***
1	40,0 %	42,2 %	41,1 %	40,3 %	43,8 %	42,0 %
2	49,1 %	52,1 %	50,6 %	45,1 %	49,1 %	47,0 %
3	55,9 %	59,5 %	57,7 %	51,3 %	56,5 %	53,8 %
4	61,9 %	66,9 %	64,3 %	57,3 %	65,8 %	61,3 %
5	63,8 %	66,7 %	65,2 %	61,7 %	69,5 %	65,3 %
6	70,9 %	75,0 %	72,9 %	66,5 %	73,8 %	69,9 %
7	71,9 %	77,9 %	74,8 %	75,6 %	80,5 %	78,0 %
8	73,3 %	80,9 %	76,9 %	75,4 %	81,9 %	78,5 %
9	78,8 %	82,1 %	80,4 %	78,6 %	84,0 %	81,2 %
10	73,5 %	83,1 %	78,0 %	77,2 %	82,2 %	79,6 %

tableau 115 : résultats de l'indice TS par tranches de fréquence (LEM avec ABIJ)

Tranche	*Anglais*			*Français*		
	Précision	***Rappel***	***F-mesure***	***Précision***	***Rappel***	***F-mesure***
1	43,0 %	44,2 %	43,6 %	43,1 %	44,8 %	43,9 %
2	59,4 %	60,6 %	60,0 %	56,1 %	56,2 %	56,1 %
3	68,5 %	69,3 %	68,9 %	70,9 %	71,2 %	71,0 %
4	76,2 %	78,5 %	77,3 %	70,2 %	73,0 %	71,6 %
5	76,5 %	76,2 %	76,3 %	74,6 %	74,7 %	74,6 %
6	76,6 %	78,8 %	77,7 %	75,5 %	75,5 %	75,5 %
7	75,5 %	78,3 %	76,9 %	81,6 %	81,6 %	81,6 %
8	77,8 %	83,3 %	80,5 %	79,6 %	83,1 %	81,3 %
9	80,9 %	83,8 %	82,3 %	79,2 %	84,2 %	81,6 %
10	71,7 %	71,7 %	83,9 %	77,3 %	84,3 %	77,9 %

tableau 116 : résultats de l'indice RV par tranches de fréquence (LEM avec ABIJ)

Tranche	*Anglais*			*Français*		
	Précision	***Rappel***	***F-mesure***	***Précision***	***Rappel***	***F-mesure***
1	44,0 %	45,9 %	44,9 %	44,7 %	47,5 %	46,1 %
2	61,4 %	64,3 %	62,8 %	54,5 %	58,0 %	56,2 %
3	69,2 %	71,6 %	70,3 %	72,3 %	75,0 %	73,6 %
4	77,2 %	81,0 %	79,0 %	69,6 %	74,3 %	71,9 %
5	78,5 %	80,3 %	79,4 %	77,1 %	78,4 %	77,8 %
6	79,2 %	82,5 %	80,8 %	75,9 %	77,9 %	76,9 %
7	77,8 %	81,4 %	79,5 %	84,4 %	84,5 %	84,5 %
8	78,7 %	84,2 %	81,3 %	81,5 %	85,3 %	83,4 %
9	82,4 %	85,4 %	83,9 %	82,0 %	86,3 %	84,1 %
10	73,6 %	85,0 %	78,9 %	74,6 %	85,1 %	79,5 %

tableau 117 : résultats de l'indice P0 par tranches de fréquence (LEM avec ABIJ)

Tranche	*Anglais*			*Français*		
	Précision	***Rappel***	***F-mesure***	***Précision***	***Rappel***	***F-mesure***
1	43,8 %	44,6 %	44,2 %	44,4 %	46,4 %	45,4 %
2	61,6 %	64,3 %	62,9 %	55,0 %	58,3 %	56,6 %
3	69,2 %	71,6 %	70,3 %	72,3 %	74,5 %	73,4 %
4	76,9 %	80,7 %	78,7 %	69,4 %	74,3 %	71,8 %
5	77,8 %	79,1 %	78,4 %	78,6 %	78,6 %	78,6 %
6	78,9 %	81,7 %	80,3 %	76,6 %	77,9 %	77,3 %
7	77,6 %	81,1 %	79,3 %	83,8 %	84,2 %	84,0 %
8	79,1 %	84,5 %	81,7 %	81,0 %	85,1 %	83,0 %
9	82,3 %	85,0 %	83,7 %	81,5 %	85,7 %	83,6 %
10	73,5 %	84,9 %	78,8 %	74,4 %	84,9 %	79,3 %

tableau 118 : résultats de l'indice PC par tranches de fréquence (LEM avec ABIJ)

Tranche	*Anglais*			*Français*		
	Précision	***Rappel***	***F-mesure***	***Précision***	***Rappel***	***F-mesure***
1	58,5 %	60,1 %	59,3 %	55,8 %	58,5 %	57,1 %
2	66,7 %	69,3 %	68,0 %	60,0 %	63,6 %	61,7 %
3	70,8 %	73,0 %	71,9 %	74,2 %	76,3 %	75,2 %
4	77,8 %	81,6 %	79,6 %	69,6 %	74,3 %	71,9 %
5	78,3 %	79,7 %	79,0 %	79,3 %	79,3 %	79,3 %
6	78,9 %	81,7 %	80,3 %	77,4 %	78,4 %	77,8 %
7	77,6 %	81,2 %	79,3 %	84,1 %	84,5 %	84,3 %
8	79,2 %	84,5 %	81,8 %	81,5 %	85,5 %	83,5 %
9	82,8 %	85,7 %	84,2 %	81,4 %	86,0 %	83,7 %
10	73,4 %	85,1 %	78,8 %	74,4 %	85,1 %	79,4 %

A-XI Résultats par catégories morphosyntaxiques (LEM avec ABIJ)

tableau 119 : résultats de l'indice IM par catégorie morphosyntaxique pour l'anglais et le français

Précision				***Rappel***			
Anglais		*Français*		*Anglais*		*Français*	
Catégorie	*P*	*Catégorie*	*P*	*Catégorie*	*R*	*Catégorie*	*R*
mot outil	46,7 %	mot outil	49,8 %	adverbe	62,0 %	verbe	69,3 %
adverbe	53,8 %	verbe	61,5 %	mot outil	68,3 %	adverbe	75,0 %
verbe	64,2 %	adverbe	68,9 %	verbe	71,2 %	adjectif	76,6 %
verbe/sub.	73,8 %	adjectif	73,1 %	verbe/sub.	78,0 %	mot outil	77,7 %
adjectif	74,8 %	verbe/sub.	74,1 %	adjectif	80,3 %	verbe/sub.	78,4 %
substantif	79,8 %	substantif	78,1 %	substantif	82,2 %	substantif	79,7 %
nom propre	84,5 %	adjectif/sub.	86,3 %	nom propre	85,0 %	adjectif/sub.	82,7 %
adjectif/sub.	86,8 %	nom propre	86,6 %	adjectif/sub.	86,8 %	nom propre	86,9 %

tableau 120 : résultats de l'indice TS par catégorie morphosyntaxique pour l'anglais et le français

Précision				***Rappel***			
Anglais		*Français*		*Anglais*		*Français*	
Catégorie	*P*	*Catégorie*	*P*	*Catégorie*	*R*	*Catégorie*	*R*
mot outil	44,5 %	mot outil	44,4 %	adverbe	68,4 %	verbe	69,6 %
adverbe	61,4 %	verbe	66,9 %	mot outil	68,9 %	mot outil	78,5 %
verbe	68,0 %	adverbe	72,6 %	verbe	74,3 %	adjectif	79,6 %
verbe/sub.	78,5 %	verbe/sub.	74,2 %	verbe/sub.	82,2 %	adverbe	80,4 %
adjectif	79,0 %	adjectif	76,3 %	adjectif	83,7 %	verbe/sub.	80,9 %
substantif	83,5 %	substantif	82,6 %	substantif	85,1 %	adjectif/sub.	85,1 %
nom propre	88,8 %	adjectif/sub.	88,1 %	adjectif/sub.	89,1 %	substantif	87,0 %
adjectif/sub.	89,5 %	nom propre	90,6 %	nom propre	89,1 %	nom propre	90,1 %

tableau 121 : résultats de l'indice P0 par catégorie morphosyntaxique pour l'anglais et le français

Précision				***Rappel***			
Anglais		*Français*		*Anglais*		*Français*	
Catégorie	*P*	*Catégorie*	*P*	*Catégorie*	*R*	*Catégorie*	*R*
mot outil	46,1 %	mot outil	45,9 %	mot outil	70,3 %	verbe	73,0 %
adverbe	62,2 %	verbe	69,1 %	adverbe	70,9 %	mot outil	79,3 %
verbe	70,9 %	verbe/sub.	75,4 %	verbe	77,1 %	adjectif	81,1 %
adjectif	80,1 %	adjectif	77,8 %	verbe/sub.	84,3 %	verbe/sub.	81,9 %
verbe/sub.	80,2 %	adverbe	79,0 %	adjectif	85,1 %	adjectif/sub.	86,0 %
substantif	84,8 %	substantif	83,8 %	substantif	86,5 %	adverbe	87,5 %
nom propre	89,7 %	adjectif/sub.	88,2 %	adjectif/sub.	89,5 %	substantif	88,1 %
adjectif/sub.	90,0 %	nom propre	91,7 %	nom propre	90,5 %	nom propre	91,3 %

tableau 122 : effectif des tranches et des catégories morphosyntaxiques en anglais

Tranches \ *Catégories*	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***Total***
substantif	33	46	69	66	78	78	81	86	73	73	***683***
nom propre	21	26	29	27	16	22	14	14	15	9	***193***
verbe	11	11	22	24	24	27	26	33	26	17	***221***
verbe / sub.	4	6	14	19	24	32	48	34	44	44	***269***
adjectif	4	8	19	22	37	31	21	29	32	20	***223***
adjecti f /sub.	1	4	3	10	5	6	7	12	10	14	***72***
adverbe	0	4	6	7	19	11	7	5	1	2	***62***
mot outil	1	0	1	6	3	8	14	23	31	70	***157***
Total	***75***	***105***	***163***	***181***	***206***	***215***	***218***	***236***	***232***	***249***	***1880***

tableau 123 : effectif des tranches et des catégories morphosyntaxiques en français

Tranches \ *Catégories*	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***Total***
substantif	36	52	63	94	94	110	102	104	121	98	***874***
nom propre	20	21	38	17	15	18	15	10	9	7	***170***
verbe	22	29	41	34	42	38	35	39	35	20	***335***
verbe / sub.	3	2	7	8	7	12	10	9	13	14	***85***
adjectif	9	20	24	45	35	25	21	27	16	13	***235***
adjectif / sub.	4	5	8	15	10	13	11	11	21	14	***112***
adverbe	2	1	5	5	7	3	8	4	3	1	***39***
mot outil	0	0	3	1	6	17	17	23	30	58	***155***
Total	***96***	***130***	***189***	***219***	***216***	***236***	***219***	***227***	***248***	***225***	***2005***

A-XII Constitution de dictionnaires à partir d'extractions filtrées

tableau 124 : filtrages absolus de l'extraction complète du corpus JOC et résultats de l'indice* d *correspondant à chaque dictionnaire

extractions FS avec ABIJ	*paramètre*	*avec l'indice P0*			*avec l'indice d lié à l'extraction correspondante*		
		P	***R***	***F***	***Pd***	***Rd***	***Fd***
extraction complète	0	71,7 %	52,0 %	60,3 %	73,9 %	52,7 %	61,5 %
extraction complète après filtrage absolu	0,25	72,6 %	51,8 %	60,5 %	74,6 %	52,5 %	61,6 %
	0,5	78,3 %	50,2 %	61,2 %	79,5 %	50,8 %	62,0 %
	1	82,9 %	45,9 %	59,1 %	83,6 %	46,4 %	59,7 %
	2,5	86,8 %	39,4 %	54,2 %	87,0 %	39,6 %	54,4 %
	4	88,8 %	35,9 %	51,2 %	89,0 %	36,0 %	51,3 %
	6	90,5 %	32,5 %	47,8 %	90,6 %	32,5 %	47,8 %
	10	91,7 %	28,8 %	43,8 %	91,8 %	28,7 %	43,8 %
	20	93,2 %	22,7 %	36,5 %	93,2 %	22,6 %	36,3 %

tableau 125 : filtrages différentiels de l'extraction complète du corpus JOC et résultats de l'indice* d *correspondant à chaque dictionnaire

extractions FS avec ABIJ	*paramètre*	*avec l'indice P0*			*avec l'indice d lié à l'extraction correspondante*		
		P	***R***	***F***	***Pd***	***Rd***	***Fd***
extraction complète	1	71,7 %	52,0 %	60,3 %	73,9 %	52,7 %	61,5 %
extraction complète après filtrage différentiel	1,05	77,0 %	45,4 %	57,2 %	76,7 %	50,5 %	60,9 %
	1,2	84,0 %	40,2 %	54,4 %	80,1 %	48,3 %	60,3 %
	1,5	89,5 %	35,6 %	50,9 %	82,5 %	45,8 %	58,9 %
	2	92,8 %	30,7 %	46,1 %	85,0 %	43,3 %	57,4 %
	2,5	94,8 %	27,8 %	42,9 %	87,1 %	41,2 %	56,0 %
	3	95,6 %	25,6 %	40,4 %	88,7 %	39,7 %	54,9 %
	4	96,2 %	22,4 %	36,4 %	90,4 %	37,8 %	53,3 %
	5	96,7 %	20,1 %	33,2 %	91,4 %	36,4 %	52,1 %
	6	97,2 %	18,3 %	30,8 %	91,8 %	34,8 %	50,5 %
	7	97,4 %	16,8 %	28,7 %	92,0 %	33,9 %	49,5 %
	10	98,1 %	13,5 %	23,7 %	93,2 %	31,8 %	47,4 %

A-XIII Alignement du corpus *Verne* à partir des correspondances lexicales

Paramétrages spécifiques (par rapport aux précédents alignements) :

Elimination de la transition T_6, et Seuil_{diag} *= 0,8 (agrandissement de l'espace de recherche). Dans l'identification des cognats, on ne conserve que les chaînes véfiant* r_{seuil} *= 0,5.*

tableau 126 : modèles 1 et 2 – résultats sans filtrage

CoeffSimil /CoeffDist	***P_1***	***R_1***	***F_1***	***P_2***	***R_2***	***F_2***
0	64,9 %	64,6 %	64,8 %	64,9 %	64,6 %	64,8 %
0,5	53,3 %	67,0 %	59,4 %	41,7 %	55,8 %	47,7 %
1,5	43,0 %	60,7 %	50,3 %	33,9 %	49,1 %	40,1 %
2	41,5 %	59,2 %	48,8 %	33,2 %	48,3 %	39,4 %
2,5	41,6 %	59,6 %	49,0 %	33,0 %	48,1 %	39,1 %
3	41,3 %	59,4 %	48,7 %	32,8 %	47,9 %	39,0 %
5	40,5 %	58,5 %	47,9 %	32,6 %	47,8 %	38,8 %
+inf	39,5 %	57,0 %	46,7 %	31,3 %	45,7 %	37,1 %

tableau 127 : modèles 1 et 2 – résultats avec filtrage (r=2, s=3,5)

CoeffSimil /CoeffDist	***P_1***	***R_1***	***F_1***	***P_2***	***R_2***	***F_2***
0	64,9 %	64,6 %	64,8 %	64,9 %	64,6 %	64,8 %
0,5	77,6 %	77,0 %	77,3 %	83,1 %	82,8 %	83,0 %
1,5	80,6 %	80,9 %	80,8 %	85,2 %	87,5 %	86,4 %
2	81,4 %	82,0 %	81,7 %	85,6 %	88,2 %	86,9 %
2,5	83,1 %	83,8 %	83,5 %	83,7 %	87,8 %	85,7 %
3	82,4 %	84,0 %	83,2 %	81,8 %	87,2 %	84,4 %
5	80,6 %	83,5 %	82,0 %	78,9 %	86,5 %	82,5 %
+inf	74,7 %	76,2 %	75,5 %	75,1 %	83,2 %	78,9 %

tableau 128 : modèles 1 et 2 – résultats en fonction du filtrage des unités de haute fréquence (+ filtrage de paramètre r = 2, s = 3,5)

$s_{fréqmax}$	P_1	R_1	F_1	P_2	R_2	F_2
10	69,7 %	69,5 %	69,6 %	72,7 %	72,9 %	72,8 %
40	78,7 %	77,7 %	78,2 %	84,5 %	83,9 %	84,2 %
50	80,3 %	79,0 %	79,7 %	86,3 %	85,6 %	85,9 %
60	81,1 %	79,7 %	80,4 %	86,7 %	86,5 %	86,6 %
70	81,2 %	80,0 %	80,6 %	86,5 %	86,8 %	86,7 %
80	81,5 %	80,3 %	80,9 %	86,6 %	87,4 %	87,0 %
90	81,6 %	80,6 %	81,1 %	86,7 %	87,5 %	87,1 %
100	81,8 %	80,8 %	81,3 %	86,7 %	87,5 %	87,1 %
110	81,8 %	81,1 %	81,4 %	86,8 %	87,6 %	87,2 %
120	81,8 %	81,1 %	81,4 %	87,0 %	87,6 %	87,3 %
150	82,3 %	81,9 %	82,1 %	86,7 %	87,7 %	87,2 %
200	83,7 %	83,6 %	83,6 %	86,9 %	87,7 %	87,3 %
250	83,3 %	83,2 %	83,2 %	86,6 %	87,7 %	87,1 %
400	83,0 %	83,1 %	83,1 %	85,6 %	88,2 %	86,9 %
500	81,4 %	82,0 %	81,7 %	85,6 %	88,2 %	86,9 %
800	81,4 %	82,0 %	81,7 %	85,6 %	88,2 %	86,9 %
5000	81,4 %	82,0 %	81,7 %	85,6 %	88,2 %	86,9 %

tableau 129 : étape 2, modèles 1 et 2 - résultats en fonction du filtrage des unités de haute fréquence (+ filtrage de paramètre r = 2, s = 3,5)

$s_{fréqmax}$	P_1	R_1	F_1	P_2	R_2	F_2
10	80,4 %	80,6 %	80,5 %	76,9 %	76,8 %	76,8 %
40	87,6 %	87,8 %	87,7 %	86,6 %	86,4 %	86,5 %
50	88,5 %	88,3 %	88,4 %	86,9 %	86,8 %	86,9 %
60	88,5 %	88,4 %	88,5 %	87,3 %	87,5 %	87,4 %
70	88,3 %	88,7 %	88,5 %	87,2 %	87,7 %	87,5 %
80	88,2 %	88,7 %	88,5 %	88,7 %	88,5 %	88,6 %
90	88,1 %	88,8 %	88,5 %	88,6 %	88,4 %	88,5 %
100	88,1 %	88,8 %	88,4 %	88,5 %	88,4 %	88,5 %
110	88,5 %	88,9 %	88,7 %	88,5 %	88,4 %	88,5 %
120	88,6 %	88,9 %	88,7 %	88,3 %	88,2 %	88,3 %
150	88,3 %	89,2 %	88,7 %	88,0 %	88,4 %	88,2 %
200	87,4 %	89,0 %	88,2 %	87,1 %	88,4 %	87,7 %
250	87,2 %	89,4 %	88,3 %	86,9 %	88,6 %	87,7 %
400	84,6 %	88,5 %	86,5 %	86,1 %	88,6 %	87,4 %
500	84,9 %	88,8 %	86,8 %	86,1 %	88,6 %	87,4 %
800	84,9 %	88,8 %	86,8 %	86,1 %	88,6 %	87,4 %
5000	84,9 %	88,8 %	86,8 %	86,1 %	88,6 %	87,4 %

tableau 130 : modèle 2 - résultats en fonction des paramètres de filtrage r et s
Indice $Score_{combiné}$ avec $k = 2$

RapportDiff	*IndiceMin*	*P*	*R*	*F*
1	0	33,2 %	48,3 %	39,4 %
1	1	52,5 %	71,9 %	60,7 %
1	2	73,1 %	86,1 %	79,1 %
1	3	80,5 %	88,0 %	84,1 %
1	3,5	82,1 %	88,2 %	85,0 %
1	4	83,4 %	88,3 %	85,8 %
1	5	84,2 %	88,0 %	86,1 %
1	6	84,1 %	87,2 %	85,6 %
1	7	83,8 %	86,3 %	85,1 %
1	8	83,7 %	85,8 %	84,7 %
1	9	83,8 %	85,7 %	84,7 %
1	10	84,0 %	84,8 %	84,4 %
1,25	0	52,6 %	68,3 %	59,4 %
1,25	1	69,1 %	83,7 %	75,7 %
1,25	2	77,6 %	87,3 %	82,2 %
1,25	3	82,2 %	87,9 %	84,9 %
1,25	3,5	82,9 %	87,9 %	85,3 %
1,25	4	83,9 %	88,0 %	85,9 %
1,25	5	84,2 %	87,5 %	85,8 %
1,25	6	83,6 %	86,6 %	85,1 %
1,25	7	83,2 %	85,5 %	84,3 %
1,25	8	82,8 %	84,7 %	83,7 %
1,25	9	81,2 %	82,7 %	81,9 %
1,25	10	82,8 %	83,7 %	83,3 %
1,5	0	67,9 %	79,8 %	73,3 %
1,5	1	76,5 %	85,9 %	80,9 %
1,5	2	80,8 %	87,7 %	84,1 %
1,5	3	83,7 %	88,0 %	85,8 %
1,5	3,5	84,6 %	88,3 %	86,4 %
1,5	4	84,6 %	88,2 %	86,4 %
1,5	5	84,3 %	87,7 %	86,0 %
1,5	6	83,6 %	86,6 %	85,0 %
1,5	7	83,0 %	85,2 %	84,1 %
1,5	8	82,8 %	84,6 %	83,7 %
1,5	9	81,1 %	82,7 %	81,9 %
1,5	10	82,8 %	83,7 %	83,2 %

tableau 130 (suite)

RapportDiff	*IndiceMin*	*P*	*R*	*F*
1,75	0	74,3 %	83,0 %	78,4 %
1,75	1	80,7 %	87,5 %	84,0 %
1,75	2	82,5 %	88,0 %	85,2 %
1,75	3	84,4 %	88,4 %	86,4 %
1,75	3,5	85,2 %	88,4 %	86,8 %
1,75	4	85,1 %	88,4 %	86,7 %
1,75	5	84,4 %	87,6 %	85,9 %
1,75	6	83,7 %	86,5 %	85,1 %
1,75	7	82,1 %	84,3 %	83,2 %
1,75	8	81,3 %	83,1 %	82,2 %
1,75	9	80,5 %	82,4 %	81,4 %
1,75	10	81,7 %	82,8 %	82,2 %
2	0	78,2 %	84,7 %	81,3 %
2	1	83,2 %	87,6 %	85,3 %
2	2	83,9 %	88,0 %	85,9 %
2	3	84,7 %	88,0 %	86,3 %
2	3,5	85,6 %	88,2 %	86,9 %
2	4	85,3 %	88,1 %	86,7 %
2	5	85,1 %	87,6 %	86,3 %
2	6	83,9 %	86,4 %	85,1 %
2	7	82,0 %	84,5 %	83,2 %
2	8	80,9 %	83,3 %	82,1 %
2	9	80,4 %	82,4 %	81,4 %
2	10	81,7 %	82,8 %	82,3 %
2,25	0	79,7 %	84,7 %	82,1 %
2,25	1	84,2 %	87,4 %	85,8 %
2,25	2	84,7 %	87,7 %	86,2 %
2,25	3	85,2 %	87,8 %	86,5 %
2,25	3,5	85,9 %	88,1 %	87,0 %
2,25	4	85,5 %	87,9 %	86,7 %
2,25	5	84,9 %	87,0 %	85,9 %
2,25	6	82,6 %	85,1 %	83,9 %
2,25	7	81,4 %	83,5 %	82,5 %
2,25	8	81,0 %	83,0 %	82,0 %
2,25	9	80,7 %	82,1 %	81,4 %
2,25	10	81,4 %	82,6 %	82,0 %

tableau 130 (suite)

RapportDiff	*IndiceMin*	*P*	*R*	*F*
2,5	0	80,5 %	84,6 %	82,5 %
2,5	1	85,2 %	87,2 %	86,2 %
2,5	2	85,4 %	87,4 %	86,4 %
2,5	3	85,8 %	87,5 %	86,7 %
2,5	3,5	86,2 %	87,7 %	87,0 %
2,5	4	86,0 %	87,6 %	86,8 %
2,5	5	84,6 %	86,7 %	85,6 %
2,5	6	82,9 %	85,1 %	84,0 %
2,5	7	81,5 %	83,5 %	82,5 %
2,5	8	81,0 %	83,0 %	82,0 %
2,5	9	80,7 %	82,1 %	81,4 %
2,5	10	81,4 %	82,6 %	82,0 %
2,75	0	80,6 %	83,9 %	82,2 %
2,75	1	84,9 %	87,4 %	86,1 %
2,75	2	85,0 %	87,5 %	86,2 %
2,75	3	85,3 %	87,4 %	86,4 %
2,75	3,5	85,5 %	87,6 %	86,6 %
2,75	4	84,6 %	87,3 %	85,9 %
2,75	5	84,1 %	86,3 %	85,2 %
2,75	6	82,1 %	84,7 %	83,4 %
2,75	7	81,6 %	83,5 %	82,5 %
2,75	8	80,9 %	82,5 %	81,7 %
2,75	9	80,8 %	82,1 %	81,5 %
2,75	10	81,5 %	82,6 %	82,1 %
3	0	81,1 %	83,8 %	82,4 %
3	1	85,0 %	86,8 %	85,9 %
3	2	85,1 %	86,9 %	86,0 %
3	3	84,3 %	86,4 %	85,3 %
3	3,5	84,5 %	86,6 %	85,5 %
3	4	84,6 %	86,6 %	85,6 %
3	5	85,0 %	86,5 %	85,7 %
3	6	82,9 %	84,8 %	83,8 %
3	7	82,6 %	84,3 %	83,4 %
3	8	81,6 %	83,1 %	82,4 %
3	9	81,6 %	82,7 %	82,2 %
3	10	81,5 %	82,6 %	82,1 %

tableau 130 (suite)

RapportDiff	*IndiceMin*	*P*	*R*	*F*
3,25	0	81,4 %	82,9 %	82,1 %
3,25	1	85,1 %	86,2 %	85,6 %
3,25	2	85,2 %	86,3 %	85,7 %
3,25	3	85,4 %	86,4 %	85,9 %
3,25	3,5	85,5 %	86,5 %	86,0 %
3,25	4	84,6 %	85,8 %	85,2 %
3,25	5	84,4 %	85,3 %	84,9 %
3,25	6	84,2 %	84,6 %	84,4 %
3,25	7	83,5 %	84,2 %	83,9 %
3,25	8	82,7 %	83,5 %	83,1 %
3,25	9	82,1 %	82,8 %	82,4 %
3,25	10	81,8 %	82,5 %	82,1 %
3,5	0	81,5 %	82,5 %	82,0 %
3,5	1	86,1 %	86,4 %	86,2 %
3,5	2	86,3 %	86,5 %	86,4 %
3,5	3	86,2 %	86,6 %	86,4 %
3,5	3,5	85,3 %	85,8 %	85,5 %
3,5	4	85,1 %	85,8 %	85,4 %
3,5	5	84,6 %	85,2 %	84,9 %
3,5	6	83,8 %	84,5 %	84,2 %
3,5	7	83,1 %	84,2 %	83,6 %
3,5	8	82,4 %	83,3 %	82,9 %
3,5	9	82,1 %	82,7 %	82,4 %
3,5	10	81,6 %	82,3 %	81,9 %
3,75	0	82,0 %	82,8 %	82,4 %
3,75	1	85,7 %	85,6 %	85,7 %
3,75	2	85,7 %	85,6 %	85,6 %
3,75	3	85,6 %	85,6 %	85,6 %
3,75	3,5	85,6 %	85,6 %	85,6 %
3,75	4	84,6 %	84,9 %	84,7 %
3,75	5	84,2 %	84,4 %	84,3 %
3,75	6	83,9 %	84,2 %	84,0 %
3,75	7	83,7 %	83,8 %	83,8 %
3,75	8	81,9 %	82,4 %	82,1 %
3,75	9	82,1 %	82,6 %	82,3 %
3,75	10	81,7 %	82,3 %	82,0 %

tableau 130 (suite)

RapportDiff	*IndiceMin*	*P*	*R*	*F*
4	0	81,4 %	82,3 %	81,8 %
4	1	85,1 %	85,0 %	85,1 %
4	2	85,1 %	85,1 %	85,1 %
4	3	84,2 %	84,4 %	84,3 %
4	3,5	84,2 %	84,4 %	84,3 %
4	4	84,2 %	84,5 %	84,4 %
4	5	83,8 %	84,0 %	83,9 %
4	6	83,4 %	83,7 %	83,5 %
4	7	83,6 %	83,8 %	83,7 %
4	8	81,3 %	81,9 %	81,6 %
4	9	81,3 %	82,0 %	81,6 %
4	10	81,0 %	81,8 %	81,4 %
4,25	0	80,3 %	81,5 %	80,9 %
4,25	1	83,5 %	83,7 %	83,6 %
4,25	2	83,6 %	83,8 %	83,7 %
4,25	3	83,6 %	83,7 %	83,6 %
4,25	3,5	83,5 %	83,6 %	83,6 %
4,25	4	83,1 %	83,3 %	83,2 %
4,25	5	83,0 %	83,1 %	83,0 %
4,25	6	82,7 %	83,0 %	82,8 %
4,25	7	82,5 %	83,2 %	82,8 %
4,25	8	80,6 %	81,5 %	81,0 %
4,25	9	80,7 %	81,7 %	81,2 %
4,25	10	80,8 %	81,4 %	81,1 %
4,5	0	80,9 %	81,7 %	81,3 %
4,5	1	84,0 %	84,1 %	84,1 %
4,5	2	84,1 %	84,2 %	84,1 %
4,5	3	84,0 %	84,1 %	84,1 %
4,5	3,5	84,0 %	84,2 %	84,1 %
4,5	4	83,6 %	83,9 %	83,7 %
4,5	5	83,1 %	83,3 %	83,2 %
4,5	6	82,3 %	82,6 %	82,4 %
4,5	7	82,1 %	82,7 %	82,4 %
4,5	8	80,3 %	80,8 %	80,5 %
4,5	9	80,4 %	81,2 %	80,8 %
4,5	10	80,4 %	80,6 %	80,5 %

figure 82 : évolution de P en fonction des paramètres de filtrage (r,s)

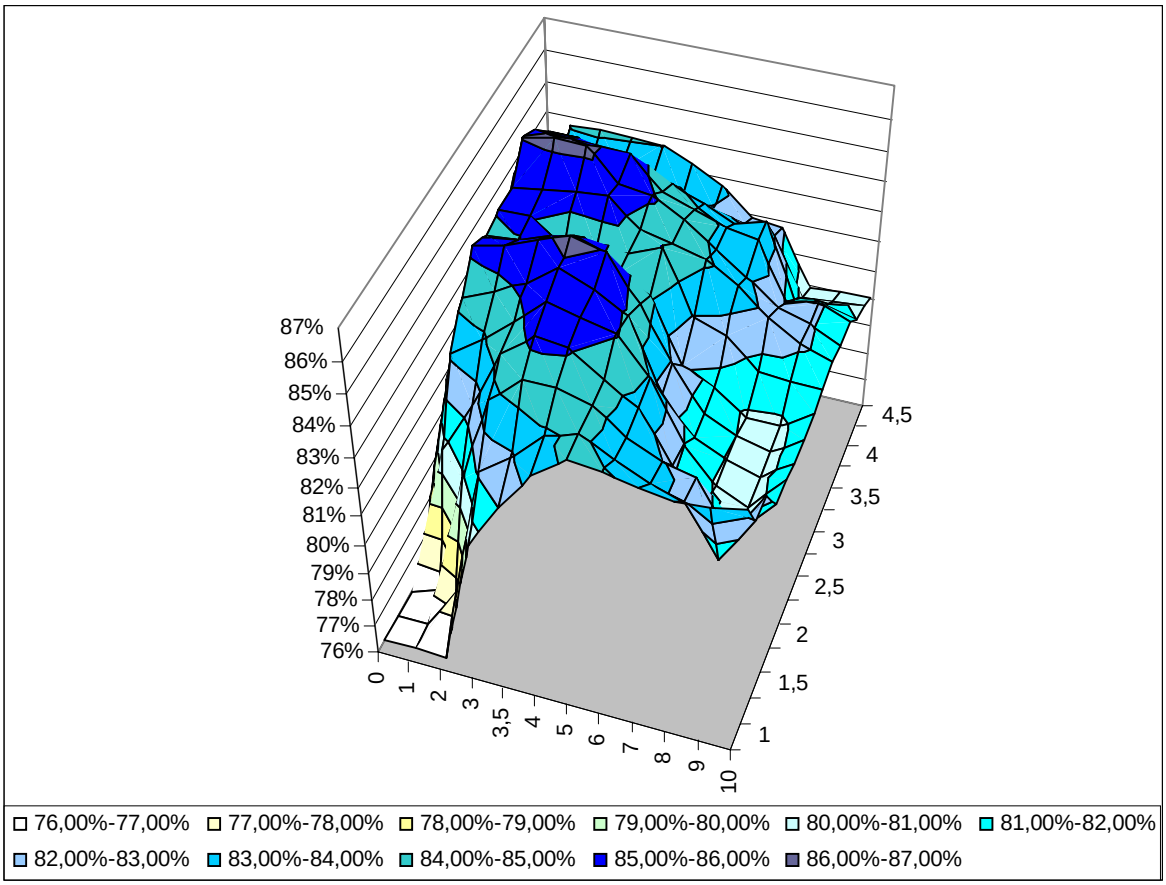


figure 83 : évolution de R en fonction des paramètres de filtrage (r,s)

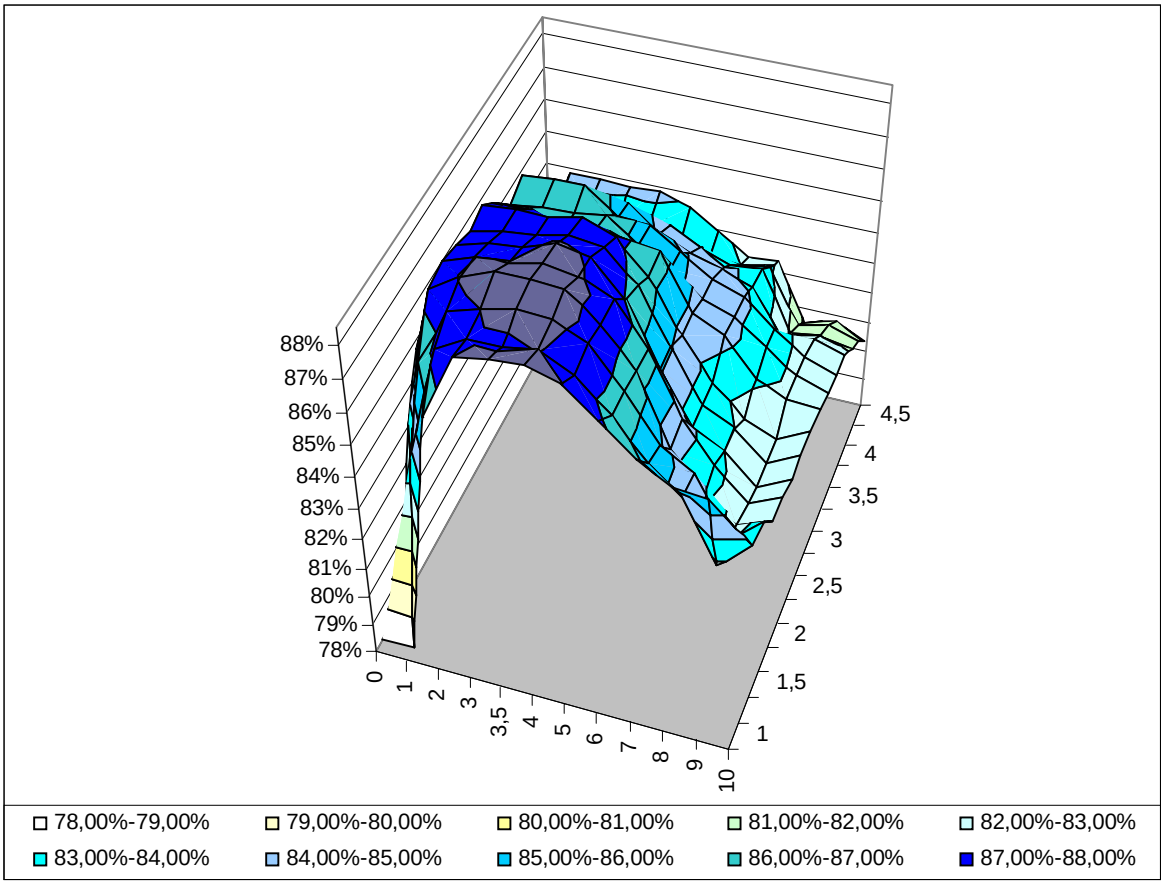


tableau 131 : étape 2, modèle 1 - résultats en fonction des paramètres de filtrage r et s
Indice $Score_{combiné}$ avec $k = 2$

RapportDiff	*IndiceMin*	*P*	*R*	*F*
1,2	1,5	76,4 %	87,3 %	81,5 %
1,2	1,75	78,4 %	88,2 %	83,0 %
1,2	2	80,0 %	88,7 %	84,1 %
1,2	2,25	80,3 %	88,6 %	84,3 %
1,2	2,5	82,0 %	89,2 %	85,4 %
1,2	2,75	82,4 %	89,0 %	85,6 %
1,2	3	82,1 %	88,9 %	85,4 %
1,2	3,5	83,6 %	88,7 %	86,1 %
1,2	3,75	84,2 %	89,0 %	86,5 %
1,2	4	84,3 %	89,0 %	86,6 %
1,2	4,25	84,1 %	88,9 %	86,4 %
1,2	4,5	84,0 %	88,9 %	86,4 %
1,2	4,75	83,9 %	88,6 %	86,2 %
1,2	5	84,2 %	88,6 %	86,4 %
1,5	1,5	78,6 %	87,8 %	83,0 %
1,5	1,75	80,1 %	88,4 %	84,0 %
1,5	2	81,4 %	88,7 %	84,9 %
1,5	2,25	82,4 %	89,0 %	85,6 %
1,5	2,5	82,8 %	88,8 %	85,7 %
1,5	2,75	83,1 %	89,2 %	86,0 %
1,5	3	83,2 %	89,2 %	86,1 %
1,5	3,5	84,0 %	89,0 %	86,4 %
1,5	3,75	84,6 %	89,1 %	86,8 %
1,5	4	84,6 %	89,1 %	86,8 %
1,5	4,25	84,6 %	89,0 %	86,7 %
1,5	4,5	84,7 %	89,0 %	86,8 %
1,5	4,75	84,6 %	88,8 %	86,6 %
1,5	5	84,3 %	88,6 %	86,4 %
1,75	1,5	81,8 %	88,7 %	85,1 %
1,75	1,75	81,7 %	88,5 %	85,0 %
1,75	2	82,0 %	88,4 %	85,1 %
1,75	2,25	83,4 %	89,0 %	86,1 %
1,75	2,5	83,6 %	88,9 %	86,2 %
1,75	2,75	83,8 %	89,0 %	86,3 %
1,75	3	84,1 %	88,9 %	86,4 %
1,75	3,5	84,4 %	88,9 %	86,6 %
1,75	3,75	85,0 %	88,9 %	86,9 %
1,75	4	85,0 %	88,9 %	86,9 %
1,75	4,25	84,9 %	88,8 %	86,8 %
1,75	4,5	84,9 %	88,8 %	86,8 %
1,75	4,75	84,8 %	88,6 %	86,6 %
1,75	5	84,5 %	88,5 %	86,5 %

tableau 131 (suite)

RapportDiff	*IndiceMin*	*P*	*R*	*F*
2	1,5	83,0 %	88,7 %	85,8 %
2	1,75	83,4 %	88,9 %	86,1 %
2	2	83,6 %	88,8 %	86,1 %
2	2,25	83,7 %	88,8 %	86,2 %
2	2,5	83,9 %	88,8 %	86,2 %
2	2,75	84,1 %	88,8 %	86,4 %
2	3	84,2 %	88,7 %	86,4 %
2	3,5	84,9 %	88,8 %	86,8 %
2	3,75	84,9 %	88,7 %	86,8 %
2	4	85,0 %	88,8 %	86,8 %
2	4,25	84,9 %	88,7 %	86,8 %
2	4,5	84,8 %	88,7 %	86,7 %
2	4,75	84,8 %	88,5 %	86,6 %
2	5	84,9 %	88,5 %	86,7 %
2,25	1,5	83,7 %	88,5 %	86,0 %
2,25	1,75	83,7 %	88,5 %	86,0 %
2,25	2	84,1 %	88,5 %	86,2 %
2,25	2,25	84,0 %	88,5 %	86,2 %
2,25	2,5	84,1 %	88,5 %	86,2 %
2,25	2,75	84,0 %	88,4 %	86,1 %
2,25	3	84,0 %	88,3 %	86,1 %
2,25	3,5	84,6 %	88,4 %	86,4 %
2,25	3,75	84,8 %	88,5 %	86,6 %
2,25	4	84,8 %	88,5 %	86,6 %
2,25	4,25	84,7 %	88,3 %	86,5 %
2,25	4,5	84,4 %	88,2 %	86,2 %
2,25	4,75	84,3 %	88,0 %	86,1 %
2,25	5	84,4 %	88,0 %	86,2 %
2,5	1,5	83,4 %	88,2 %	85,7 %
2,5	1,75	83,8 %	88,4 %	86,0 %
2,5	2	84,2 %	88,4 %	86,2 %
2,5	2,25	84,3 %	88,4 %	86,3 %
2,5	2,5	84,3 %	88,3 %	86,3 %
2,5	2,75	84,5 %	88,4 %	86,4 %
2,5	3	84,5 %	88,4 %	86,4 %
2,5	3,5	84,6 %	88,3 %	86,4 %
2,5	3,75	84,9 %	88,4 %	86,6 %
2,5	4	84,9 %	88,3 %	86,6 %
2,5	4,25	84,8 %	88,2 %	86,5 %
2,5	4,5	84,8 %	88,2 %	86,5 %
2,5	4,75	84,7 %	88,1 %	86,3 %
2,5	5	84,5 %	87,9 %	86,2 %

tableau 131 (suite)

RapportDiff	*IndiceMin*	*P*	*R*	*F*
2,75	1,5	84,7 %	88,5 %	86,6 %
2,75	1,75	84,5 %	88,5 %	86,4 %
2,75	2	84,6 %	88,4 %	86,5 %
2,75	2,25	84,6 %	88,4 %	86,5 %
2,75	2,5	84,6 %	88,3 %	86,4 %
2,75	2,75	84,6 %	88,4 %	86,5 %
2,75	3	84,7 %	88,3 %	86,5 %
2,75	3,5	84,8 %	88,3 %	86,5 %
2,75	3,75	85,0 %	88,3 %	86,6 %
2,75	4	84,9 %	88,3 %	86,6 %
2,75	4,25	84,9 %	88,2 %	86,5 %
2,75	4,5	84,9 %	88,2 %	86,5 %
2,75	4,75	84,7 %	88,0 %	86,3 %
2,75	5	84,5 %	87,9 %	86,2 %
3	1,5	84,4 %	88,1 %	86,2 %
3	1,75	84,3 %	88,0 %	86,2 %
3	2	84,6 %	88,0 %	86,2 %
3	2,25	84,5 %	88,0 %	86,2 %
3	2,5	84,5 %	87,9 %	86,2 %
3	2,75	84,6 %	88,0 %	86,2 %
3	3	84,6 %	88,0 %	86,3 %
3	3,5	84,5 %	88,0 %	86,3 %
3	3,75	84,8 %	88,1 %	86,4 %
3	4	84,7 %	88,1 %	86,4 %
3	4,25	84,7 %	88,1 %	86,4 %
3	4,5	84,7 %	88,1 %	86,4 %
3	4,75	84,6 %	87,9 %	86,2 %
3	5	84,6 %	87,9 %	86,2 %
3,5	1,5	84,5 %	87,5 %	86,0 %
3,5	1,75	84,5 %	87,5 %	86,0 %
3,5	2	84,7 %	87,5 %	86,1 %
3,5	2,25	84,8 %	87,6 %	86,2 %
3,5	2,5	85,2 %	87,8 %	86,5 %
3,5	2,75	85,2 %	87,8 %	86,5 %
3,5	3	85,3 %	87,9 %	86,6 %
3,5	3,5	85,2 %	87,8 %	86,5 %
3,5	3,75	85,3 %	87,8 %	86,5 %
3,5	4	85,1 %	87,8 %	86,5 %
3,5	4,25	85,2 %	87,9 %	86,5 %
3,5	4,5	85,2 %	87,9 %	86,5 %
3,5	4,75	85,0 %	87,7 %	86,3 %
3,5	5	85,0 %	87,7 %	86,3 %

tableau 131 (suite)

RapportDiff	*IndiceMin*	*P*	*R*	*F*
4	1,5	84,7 %	87,1 %	85,9 %
4	1,75	84,7 %	87,1 %	85,9 %
4	2	84,7 %	87,0 %	85,9 %
4	2,25	84,9 %	87,0 %	86,0 %
4	2,5	84,9 %	87,0 %	86,0 %
4	2,75	84,9 %	87,0 %	85,9 %
4	3	85,0 %	87,1 %	86,0 %
4	3,5	84,7 %	87,1 %	85,9 %
4	3,75	84,7 %	87,1 %	85,9 %
4	4	84,6 %	87,1 %	85,8 %
4	4,25	84,6 %	87,1 %	85,8 %
4	4,5	84,6 %	87,1 %	85,8 %
4	4,75	84,5 %	87,0 %	85,8 %
4	5	84,5 %	87,0 %	85,7 %

figure 84 : étape 2 - évolution de P en fonction des paramètres de filtrage (r,s)

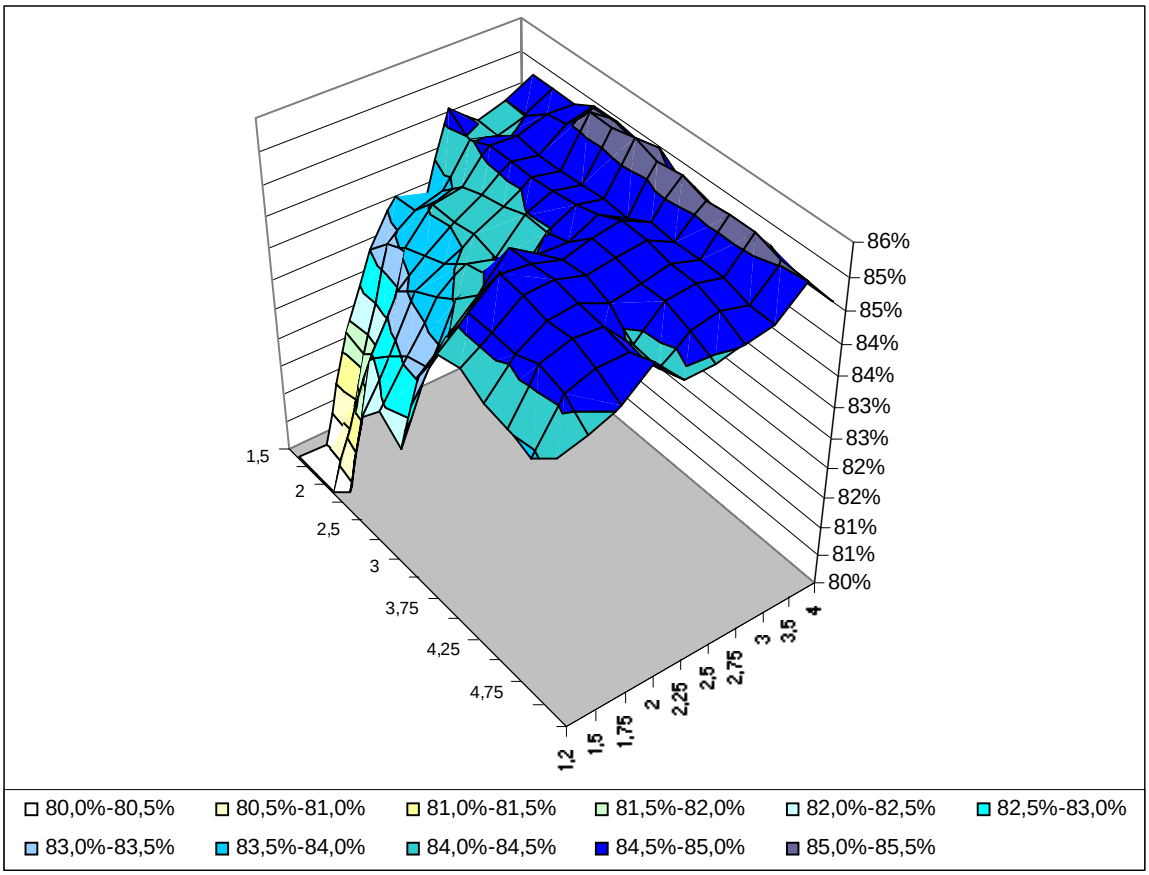


figure 85 : étape 2 - évolution de R en fonction des paramètres de filtrage (r,s)

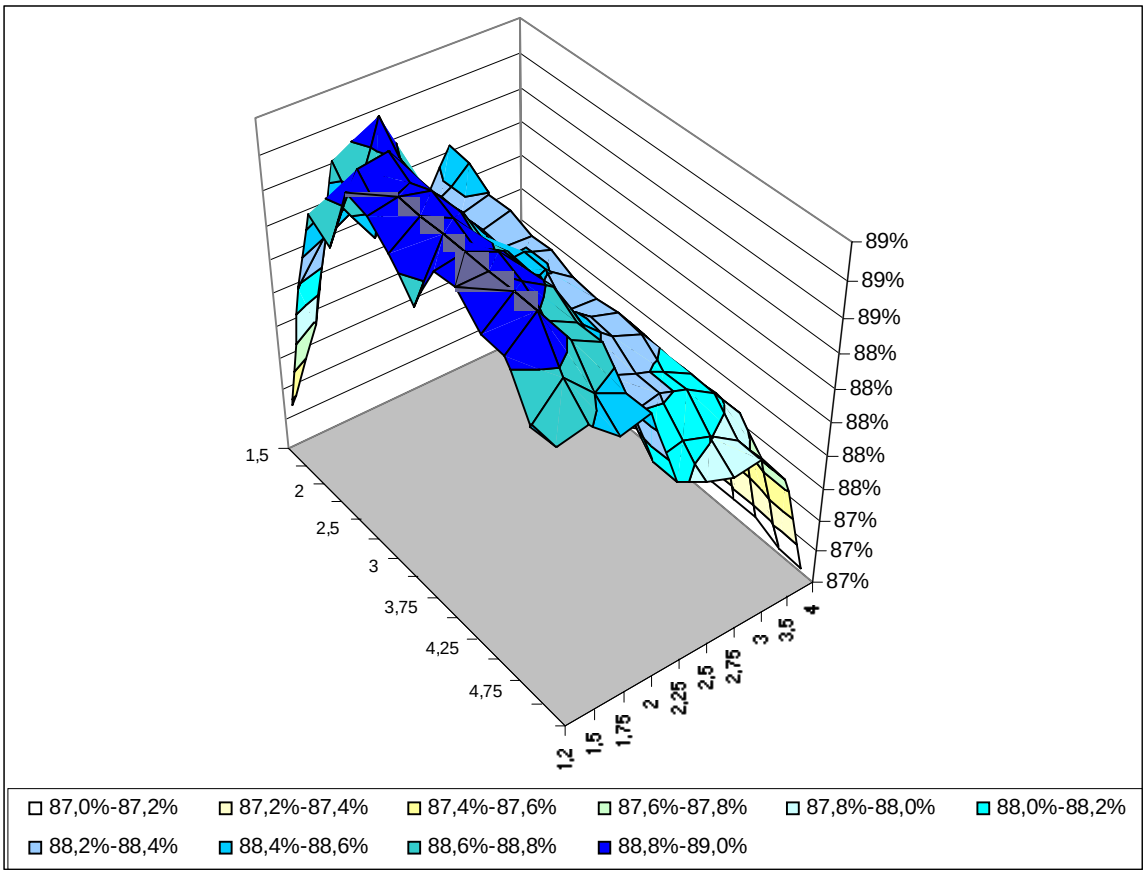


tableau 132 : nombre de couples de formes extraits avant et après filtrage (r = 2, s = 3,5)

Ces couples sont comptés à l'intérieur de l'espace de recherche ou sur le chemin optimal. Dans ce dernier cas, on fait figurer le nombre moyen de couples filtrés par binômes.

	Espace de recherche			***Chemin***						
	Couples	***Filtrés***		***Couples***	***Filtrés***		***par binômes***	***P***	***R***	***F***
Etape 1 modèle 1	477639	7636	1,6 %	31229	5741	18 %	2,28	81,41 %	81,96 %	81,69 %
Etape 1 modèle 2	477639	22775	4,8 %	31731	6663	21 %	2,64	85,57 %	88,24 %	86,88 %
Etape 2 modèle 1	477639	29784	6,2 %	32000	6924	22 %	2,75	84,90 %	88,76 %	86,79 %
Etape 2 modèle 2	477639	24323	5,1 %	31625	6976	22 %	2,77	86,13 %	88,62 %	87,36 %

tableau 133 : exemples de couples de formes extraits avant et après filtrage (r = 2, s = 3,5)

Etape 1, modèle 2

Numéro des segments	***Correspondances extraites***	***Correspondances filtrées***
1 <-> 1	(MOON, A) (EARTH, LUNE) (TO, TERRE) (FROM, DE) (THE, LA)	
9 <-> 12	(singularly, dans) (which, quoi) (distanced, surpassèrent) (point, singulièrement) (Europeans, Européens) (gunnery, balistique) (science, science) (was, fut) (Americans, Américains) (But, Mais)	(science, science) (was, fut) (Americans, Américains) (But, Mais)
13 <-> 16	(fact, doit) (one, personne) (surprise, Ceci) (need, étonner) (no, ne)	
56 <-> 82	(By, perdu) (gunnery, artillerie) (future, avenir) (America, Amérique) (is, est)	(America, Amérique) (is, est)
93 <-> 119	(few, trois) (one, au) (sink, droit) (defiance, mépris) (Shall, ne) (countrymen, pendront) (hang, couleront) (rights, seul) (nations, nationaux) (steamers, steamers) (French, Français) (our, nos) (English, Anglais) (or, ou) (not, pas)	(French, Français) (English, Anglais) (or, ou)
99 <-> 131	(out, loin) (did, t) (once, Et) (way, appartenu) (s, aux) (going, autrefois) (one, a) (stop, n) (belong, tenez) (But, si) (find, chercher) (for, elle) (cause, motif) (North, Nord) (America, Amérique) (English, Anglais) (without, sans) (war, guerre) (not, pas)	(America, Amérique) (English, Anglais) (without, sans) (war, guerre) (not, pas)
107 <-> 138	(By, prochaines) (Jove, élections)	
117 <-> 148	(BARBICANE, _)	
117 <-> 149	(Very, G.) (P.G.C, leur_) (cordially, cordialement) (IMPEY, IMPEY) (BARBICANE, BARBICANE)	
124 <-> 157	(when, «) (so, [) (reach, élevées) (against,]) (perfect, simple) (ran, connaître) (front, poussant) (There, cette) (they, premiers) (up, s) (learn, portes) (struggling, avides) (eager,	(ideas, idées) (action, action) (with, avec) (President, président)

Numéro des segments	*Correspondances extraites*	*Correspondances filtrées*
	populaire) (educated, cherchant) (vulgar, rangs) (which, aux) (pressed, pressait) (ranks, Gouvernement) (squeezing, écrasant) (herd, personnel.) (crushing, bousculant) (pushing, rencontrait) (self, self) (doors, gagner) (each, chacun) (government., government) (peculiar, particulière) (freedom, liberté) (is, dans) (that, là) (masses, masses) (important, importante) (nature, se) (ideas, idées) (all, tous) (action, action) (with, avec) (who, qui) (President, président) (communication, communication) (Barbicane, Barbicane)	(communication, communication) (Barbicane, Barbicane)
134 <-> 169	(by, salle) (large, large) (occupied, occupait) (saloon, esplanade) (platform, extrémité) (assisted, assisté) (secretaries, secrétaires) (four, quatre) (president, président) (At, A)	(secretaries, secrétaires) (four, quatre) (president, président) (At, A)
157 <-> 197	(seek, chercher) (pine, ordre) (We, aliment) (make, dévore) (all, prendre) (up, Il) (for, ») (train, parti) (our, dans) (some, son) (activity, activité) (then, donc) (another, autre) (ideas, idées) (must, faut) (which, qui) (we, nous)	(then, donc) (another, autre) (ideas, idées) (must, faut) (which, qui) (we, nous)
167 <-> 214	(for, peut) (It, Il) (Columbuses, Colombs) (this, ce) (perhaps, être) (reserved, réservé) (us, nous) (world, monde) (unknown, inconnu) (is, est)	(reserved, réservé) (us, nous) (world, monde) (unknown, inconnu) (is, est)
208 <-> 267	(chests, ») (from, «) (amazement, oh) (words, paroles) (these, ces) (At, A)	(words, paroles) (these, ces) (At, A)
277 <-> 367	(institution, établissement) (by, donc) (reposed, titres) (justified, justifiait) (confidence, confiance) (This, Cet) (all, tous) (celebrated, célèbre) (Club, Club) (Gun, Gun)	(all, tous) (celebrated, célèbre)
347 <-> 419	(S, 0) (N, °) (or, ou)	(or, ou)
381 <-> 450	(endued, furent) (own, aussitôt) (immediately, rotation) (masses, animés) (around, autour) (rotary, amas) (These, Ces) (their, leur) (central, central) (point, point) (motion, mouvement)	(These, Ces) (their, leur) (point, point)
397 <-> 487	(has, Cléomène) (attention, réfléchie) (moon, brillait) (by, qu) (phases, enseigna) (produced, lumière) (her, elle)	
419 <-> 518	(dried, savoir) (unable, manière) (thoroughly, complète) (Whether, non) (up, purent) (ancient, anciennes) (ascertain, rainures) (beds, lits) (rivers, rivières) (chasms, desséchés) (were, étaient) (these, ces) (not, ne) (or, ou) (they, ils)	(these, ces) (not, ne) (or, ou) (they, ils)
472 <-> 582	(but, dernier) (third, lieu) (class, ignorants) (remains, superstitieuse) (superstitious, classe) (There, Restait)	
491 <-> 610	(that, répondit) (appeared, En) (Nevertheless, général) (depend, Morgan) (projectile, effet)	
51 <-> 655	(ever, huit) (believe, seraient) (I, cents) (This,, atteinte) (is, ces) (attained, jusqu) (maximum, maximum) (replied, reprit) (velocity, vitesse) (Barbicane, Barbicane)	(velocity, vitesse) (Barbicane, Barbicane)
542 <-> 687	(then, Parfaitement)	
552 <-> 702	(diameter., Précisément)	
556 <-> 711	(For, qui) (employed, II) (instance, taille) (were, belle) (1,900, lança) (II., pesaient) (weight, neuf) (shot, boulets) (Mahomet, Mahomet) (during, Ainsi) (siege, siège) (1453, 1453) (Constantinople, Constantinople) (stone, pierre) (by, par) (pounds, livres)	(Constantinople, Constantinople) (stone, pierre) (by, par) (pounds, livres)
569 <-> 737	(it, absolument) (must, faut)	(must, faut)

Numéro des segments	*Correspondances extraites*	*Correspondances filtrées*
591 <-> 759	(iron., fonte) (Employ, Employer) (another, autre) (metal, métal)	(another, autre)
598 <-> 765	(asked, dit) (major, major)	(major, major)
605 <-> 772	(very, projectile) (purpose, forme) (wrought, extrêmement) (It, tout) (than, se) (created, léger) (lighter, créé) (easily, facilement) (forming, fournir) (with, nature) (furnishing, alumine) (widely, répandu) (distributed, travaille) (express, exprès) (rocks, roches) (projectile., puisque) (have, dans) (base, base) (seems, semble) (most, plus) (material, matière) (iron, fer) (us, nous) (for, pour) (been, été) (times, fois) (our, notre) (three, trois) (is, est)	(seems, semble) (been, été) (times, fois) (our, notre) (three, trois)
634 <-> 808	(following, son) (evening, aussitôt) (renewed, préambule) (was, reprit) (The, La) (discussion, discussion)	(The, La) (discussion, discussion)
681 <-> 877	(whatever, déperdition) (So, employée) (loss, ainsi) (propulsion., impulsion) (employed, aucune) (all, toute) (be, aura) (force, force) (expansive, expansive) (no, n) (there, y) (will, sera) (gas, gaz) (powder, poudre)	(there, y) (gas, gaz) (powder, poudre)
686 <-> 903	(committee, Et) (adjourned, pourquoi)	
735 <-> 960	(during, peuvent) (point, doute) (question, même) (raised, procès) (taken, dans) (before, ai) (artillery., artillerie) (for, mis) (depositions, verbaux) (called, relevés) (myself, moi) (cannot, ne) (be, être) (facts, faits) (committee, Comité) (These, Ces) (I, je)	(be, être) (facts, faits) (committee, Comité) (These, Ces) (I, je)
743 <-> 969	(going, jusque) (am, se) (mind, propre) (presently, Mon) (easy, choses) (always, amour) (propose, proposerai) (cried, dans) (Our, répliqua) (jokes, quantités) (at, qui) (make, bientôt) (propensities, rassure) (matters,, folâtre) (satisfy, satisferont) (serious, sérieuses) (him, qu) (artillerist, artilleur) (gunpowder, poudre) (his, son) (friend, ami) (but, mais) (major, major) (I, je) (is, est) (Maston, Maston)	(gunpowder, poudre) (friend, ami) (but, mais) (major, major) (I, je) (is, est) (Maston, Maston)
767 <-> 1000	(this, par) (followed, suivit) (triple, triple) (proposal, proposition) (A, Un) (silence, silence) (moment, moment)	(proposal, proposition) (A, Un) (silence, silence) (moment, moment)
796 <-> 1050	(property, faudra) (take, inaltérable) (charge, puisqu) (valuable, qualité) (unaltered, précieuse) (as, yeux) (would, charger) (us, nos) (inasmuch, De) (it, plus) (moisture, humidité) (by, pour) (several, plusieurs) (pyroxyle, pyroxyle) (cannon, canon) (days, jours) (is, est)	(pyroxyle, pyroxyle) (cannon, canon) (days, jours) (is, est)
824 <-> 1086	(was, dans) (who, Un) (attempt, tentative) (protested, protesta) (individual, homme) (against, contre) (alone, seul) (all, tous) (Union, Union) (States, États) (Club, Club) (Gun, Gun)	(against, contre) (Union, Union) (States, États)
854 <-> 1125	(chance, voulut) (offering, laissant) (tempt, surexcité) (disgusted, inqualifiable) (every, toutes) (tried, chances) (this, cet) (obstinacy, entêtement) (him, lui) (by, par) (Barbicane, Barbicane) (Nicholl, Nicholl)	(him, lui) (Barbicane, Barbicane) (Nicholl, Nicholl)
871 <-> 1140	(his, suprême) (with, Il) (mingled, y) (absolute, absolu) (intense, impuissance) (impotence, mêlait) (feeling, sentiment) (was, s) (jealousy, jalousie)	(jealousy, jalousie)
877 <-> 1151	(hot, comme) (quitting, cette) (it, pour) (still, En) (red, inflammation) (melt, tenant) (heads, sortie) (that, cents) (be, sa) (granting, regardant) (support, supporterait) (developed,	(spectators, spectateurs) (by, par) (temperature,

Numéro des segments	*Correspondances extraites*	*Correspondances filtrées*
	développés) (acquired, acquise) (shower, pluie) (Further, fondrait) (ignition, résistât) (1,600,000, seize) (imprudent, imprudents) (on, mille) (resist, résisterait) (regarding, crâne) (would, bouillante) (fall, pareille) (back, retomberait) (could, ne) (sufficient, suffisante) (shell, obus) (spectators, spectateurs) (pressure, pression) (upon, sur) (by, par) (temperature, température) (gas, gaz) (less, moins) (velocity, vitesse) (pounds, livres) (Columbiad, Columbiad) (not, pas) (powder, poudre)	température) (gas, gaz) (less, moins) (velocity, vitesse) (pounds, livres) (Columbiad, Columbiad) (powder, poudre)
924 <-> 1225	(without, coulée) (result, décidé) (however, serait) (States, sol) (different, soit) (two, Il) (entirely, Columbiad) (The, donc) (these, celui) (this, dans) (decision, Floride) (towns, Texas) (was, fut)	(was, fut)
947 <-> 1252	(pure, Floride) (metal, fonte)	(metal, fonte)
984 <-> 1290	(do, faire) (nothing, dans) (with, elles) (they, n) (prepossessions, personnalités) (for, aux) (political, politiques) (had, avaient) (As, Quant) (question, question)	(As, Quant) (question, question)
1016 <-> 1326	(this, C) (nothing, là) (But, belle) (mere, bonne) (English, anglaise) (jealousy, jalousie) (was, était)	(jealousy, jalousie) (was, était)
1023 <-> 1336	(The, au) (Mobilier, mobilier) (Credit, Crédit) (Paris, Paris) (At, A)	(Paris, Paris) (At, A)
1084 <-> 1410	(things, vînt) (established, provoquer) (feared, craignaient) (disturb, déranger) (order, globe) (would, ne) (it, dans) (that, qu) (They, ils)	(They, ils)
1122 <-> 1452	(passage, ne) (long, longue) (The, La) (was, fut) (not, pas)	(was, fut)
1175 <-> 1537	(smaller, s) (became, dans) (among, bois) (thickets, gros) (thinly, éparpillèrent) (dense, épais) (trees, arbres) (less, moins)	
1182 <-> 1544	(no, ces) (however, inquiéter) (upon, sans) (effect, bornèrent) (had, ils) (hostile, hostiles) (These, se) (demonstrations, démonstrations) (companions, compagnons) (his, ses) (Barbicane, Barbicane)	(his, ses) (Barbicane, Barbicane)
1243 <-> 1638	(removed, régulièrement) (advanced, avançaient) (actively, enlèvement) (regularly, activaient) (works, travaux) (rubbish, matériaux) (Nevertheless, Cependant) (cranes, grues) (steam, vapeur)	(Nevertheless, Cependant) (cranes, grues)
1263 <-> 1661	(XV, XV)	
1264 <-> 1662		
1341 <-> 1747	(patience, difficile) (The, Il) (sorely, calculer) (was, était)	(was, était)
1365 <-> 1781	(pretty, au) (sure, plus) (for, avait) (now, n) (There, y) (only, sûr) (fail, manquerait) (would, ne) (they, Il) (that, qu) (were, vous) (rendezvous, rendez) (wait, attendre) (was, était) (she, elle) (not, pas) (moon, Lune)	(rendezvous, rendez) (wait, attendre) (she, elle) (moon, Lune)
1441 <-> 1894	(be, écria) (person, t) (passage, s)	
1498 <-> 1992	(passion, risqua) (have, davantage) (impossible, souvent) (In, pas) (I, n) (his, Ardan) (was, avait)	
1578 <-> 2105	(gentlemen,, aujourd) (certainty, aller) (which, on) (opinions, On) (Yes, ira) (minded, océans) (magic, atmosphérique) (now, hui) (travel, traversé) (must, sera) (some, va) (make, bientôt) (facility, facilement) (outstep, sûrement) (narrow, aux) (would, océan) (rapidity, rapidement) (Liverpool, Liverpool) (as, comme) (stars, étoiles) (York, York) (New,	(stars, étoiles) (planets, planètes)

Numéro des segments	*Correspondances extraites*	*Correspondances filtrées*
	New) (planets, planètes) (moon, Lune)	
1613 <-> 2164	(from, bien) (far, fallait) (he, sans) (desirable, empêcher) (was, vers) (now, fût) (conversant, tiré) (divert, dévier) (with, donc) (It, Il) (which, dont) (practical, pratiques) (doubtless, doute) (nature, se) (questions, questions) (less, moins)	(practical, pratiques) (doubtless, doute) (questions, questions) (less, moins)
1657 <-> 2216	(A, ses) (followed, emporté) (which, Il) (course, instincts) (J, ingénieur) (no, fougueux) (T, secrétaire) (thunder, probable) (than, avait) (other, été) (applause, hardie) (author, hasarder) (Maston, par) (this, cette) (proposal, proposition) (was, est)	(this, cette) (proposal, proposition)
1712 <-> 2280	(state, logique) (however, été) (time, sans) (at, pendant) (one, moi) (These, certaine) (were, ces) (activity, activité) (volcanoes, volcans)	
1835 <-> 2406	(Captain, capitaine) (Nicholl, Nicholl)	(Nicholl, Nicholl)
1873 <-> 2444	(are, au) (fighting, battent) (this, ce) (Skersnaw, Skersnaw) (wood, bois) (morning, matin) (They, Ils)	(wood, bois) (morning, matin) (They, Ils)
1923 <-> 2494	(vain, compagnons) (an, s) (intensified, arrêtèrent) (spent, vaines) (pursuit, recherches) (After, Après) (hour, heure) (two, deux)	(After, Après) (two, deux)
2044 <-> 2611	(me, présente) (at, Michel) (introduce, te) (permit, permets) (Himself,, Lui) (you, je) (time, temps) (same, même) (worthy, digne) (Captain, capitaine) (replied, répondit) (Ardan, Ardan) (Nicholl, Nicholl)	(worthy, digne) (replied, répondit) (Nicholl, Nicholl)
2062 <-> 2627	(Nothing, chose) (more, Pas)	
2067 <-> 2632	(certainly,, certes) (replied, répliqua) (Yes, Oui) (president, président)	(replied, répliqua) (Yes, Oui) (president, président)
2125 <-> 2712	(sold, correspondait) (his, univers) (for, public) (As, mettait) (all, disposition) (taken, entier) (were, se) (size, avec) (they, Il)	
2126 <-> 2713	(million, faisait) (More, répétait) (incredibly, propageait) (disposed, ceux) (copies, ses) (space, surtout) (time, on) (an, ne) (short, bons) (half, On) (than, qu) (were, pas)	
2127 <-> 2714	(as, riche) (But, prêtait) (who, habitude) (only, côté) (well, suivant) (but, car) (it, On) (not, ce) (him, lui) (was, était)	(him, lui) (was, était)
2162 <-> 2760	(parabola, fit) (that, On) (fired, feu)	
2224 <-> 2833	(opened, toute) (On, ailleurs) (P.M, suffirait) (November, devait) (exactly, pesanteur) (at, sa) (plate, poids) (th, Il) (20, quatre) (six, cents) (was, est)	(six, cents) (was, est)
2279 <-> 2998	(indeed, Et) (were, Hill) (precautions, amener) (Americans, abord) (before, s) (had, bien) (success, enceinte) (dangers, Stone) (accruing, chargement) (all, tout) (from, se) (carelessness, garda) (his, son)	(his, son)
236 <-> 3091	(may, perdiez) (that, C) (bets, paris) (your, vos) (other, autres) (two, deux) (is, est) (you, vous)	(other, autres) (two, deux (is, est) (you, vous)
2411 <-> 3145	(all, au) (her, vous) (gracefully, exacte) (rendezvous, rendez) (She, Elle) (was, était)	(rendezvous, rendez)
2487 <-> 3221	(as, Hurrah) (at, pour) (had, «) (him, Ardan)	(him, Ardan)
2537 <-> 3300	(astronomy, jamais) (immortalized, leur) (events, Les) (were, cercle) (all, temps) (annals, sont) (At, se) (certain, dans) (names, noms) (be, être) (Michel, Michel) (Ardan, Ardan) (Barbicane, Barbicane) (Nicholl, Nicholl)	(be, être) (Barbicane, Barbicane) (Nicholl, Nicholl)

tableau 134 : couples de formes extraits et filtrés (r = 2, s = 3,5)

Etape 1, modèle 2. Nb *désigne le nombre de fois que le couple a été extrait.*

Anglais	*Français*	*Nb*	*Anglais*	*Français*	*Nb*
1	1	5	machines	machines	5
2	2	4	magistrates	magistrats	1
3	3	7	major	major	30
4	4	8	man	homme	16
5	5	2	masonry	maçonnerie	4
6	6	3	mass	masse	10
7	7	2	Maston	Maston	18
15	15	2	material	matière	1
18	18	2	mathematical	mathématiques	2
22	22	4	me	me	4
28	28	1	me	moi	4
$	dollars	8	me	vous	1
:	:	30	means	moyen	9
@	°	1	mechanical	mécaniques	2
_	_	22	meeting	meeting	6
_	Lune	6	meeting	séance	8
_	Sur	1	members	membres	8
A	Un	18	men	hommes	8
A	Une	7	menaces	menaces	2
absolute	absolue	3	metal	fonte	1
absolutely	absolument	5	metal	métal	18
according	suivant	2	Mexicans	Mexicains	1
action	action	7	Michel	Michel	3
adversary	adversaire	4	midnight	minuit	2
after	après	38	midst	milieu	12
against	contre	7	miles	milles	27
air	air	17	millions	millions	7
all	tous	20	minutes	minutes	4
all	toutes	10	molecules	molécules	1
alone	seul	4	moment	instant	2
already	déjà	2	moment	moment	26
also	aussi	5	money	argent	3
altitude	hauteur	1	month	mois	8
always	toujours	12	months	mois	10
America	Amérique	15	moon	astre	3
American	Américain	3	moon	Lune	113
American	américaine	9	moon	projectile	1
Americans	Américains	15	more	plus	17
And	Et	22	Morgan	Morgan	14
angle	angle	3	morning	matin	8
animals	animaux	3	mortar	mortier	5
another	autre	11	mortars	mortiers	3
answer	répondre	1	most	plus	14
antipathy	antipathie	2	motion	mouvement	1
Ardan	Ardan	17	motions	mouvements	1
Ardan	Nicholl	3	mould	moule	6
are	sont	24	mouth	bouche	6
arguments	arguments	8	movement	mouvement	3

Anglais	*Français*	*Nb*	*Anglais*	*Français*	*Nb*
armor	cuirasse	1	multitude	foule	1
arms	bras	6	Murchison	Murchison	4
around	autour	1	must	faut	18
artillery	artillerie	5	my	ma	3
as	comme	14	my	mes	1
As	Quant	9	my	mon	8
asked	demanda	16	my	vous	2
aspect	aspect	1	mysterious	mystérieux	3
astronomers	astronomes	7	nature	nature	5
astronomical	astronomique	2	nature	se	1
astronomical	astronomiques	6	necessary	nécessaires	1
At	A	32	never	ne	12
At	Ce	1	never	pas	2
atmosphere	atmosphère	16	Nevertheless	Cependant	5
atoms	atomes	1	New	New	1
attention	attention	6	newspapers	journaux	5
attentively	attentivement	1	Nicholl	capitaine	1
attraction	attraction	5	Nicholl	Nicholl	50
axis	axe	5	night	nuit	3
ball	boulet	4	nine	neuf	10
balls	balles	3	No	ci.	3
Baltimore	Baltimore	14	nor	ni	1
Barbicane	Barbicane	177	not	n	12
bars	barres	1	not	ne	42
base	base	1	not	non	6
bay	baie	4	Not	Pas	58
be	être	12	nothing	rien	16
be	sera	6	November	novembre	5
beasts	bêtes	1	Now	Or	11
became	se	2	number	nombre	9
bed	couche	1	observation	observation	2
been	été	19	observations	observations	10
bellicose	belliqueux	2	October	octobre	14
belonging	appartenant	3	old	vieilles	2
Berlin	Berlin	1	on	sur	35
best	meilleur	4	operation	opération	15
better	mieux	3	opinion	opinion	3
between	entre	25	opportunity	occasion	5
Bilsby	Bilsby	1	or	ou	34
bird	oiseau	2	orator	orateur	3
board	bord	3	orbit	orbite	2
bottom	fond	5	order	afin	1
branches	branches	1	orifice	orifice	1
bronze	bronze	2	other	autre	25
brought	Tampa	1	other	autres	6
bushman	bushman	5	others	autres	6
But	Mais	125	ought	«	1
by	ne	12	our	nos	8
by	par	108	our	notre	25
calculations	calculs	7	Paris	Paris	1
Cambridge	Cambridge	1	Parrott	Parrott	1
cannon	canon	28	peace	paix	4
captain	capitaine	21	perceived	aperçu	2
care	soin	3	perfection	perfection	2

Anglais	*Français*	*Nb*	*Anglais*	*Français*	*Nb*
cast	fonte	2	period	temps	2
casting	fonte	13	persevering	persévérant	1
cat	chat	3	personage	personnage	4
cat	écureuil	1	persons	personnes	3
catastrophes	catastrophes	2	phases	phases	3
celebrated	célèbre	4	Philadelphia	Philadelphie	5
center	centre	8	piece	pièce	5
century	siècle	3	pistols	pistolets	1
certain	certain	9	place	lieu	12
change	changement	3	plain	plaine	1
CHAPTER	ET	1	plains	plaines	3
charged	chargé	3	plan	moyen	2
chemist	chimiste	1	plane	plan	2
children	enfants	3	planets	planètes	11
choose	choisir	3	plates	plaques	1
chronometer	chronomètre	2	platform	estrade	5
circle	cercle	4	point	point	19
citizens	citoyens	2	points	points	8
city	ville	1	position	position	1
classes	classes	1	possible	possible	10
clock	heures	1	post	poste	2
clouds	nuages	1	pounds	livres	26
club	club	4	powder	poudre	25
coal	charbon	1	practical	pratiques	1
Coldspring	Goldspring	1	precedes	précède	1
colleagues	collègues	15	presence	présence	4
Columbiad	Columbiad	43	president	président	78
committee	Comité	8	pressure	pression	5
communication	communication	9	principle	principe	5
condition	condition	2	principles	principes	1
conditions	conditions	16	problem	problème	4
conical	coniques	2	produced	produit	2
Constantinople	Constantinople	1	professor	professeur	1
contains	renferme	1	profound	profond	1
continents	continents	2	progress	progrès	3
contradictor	contradicteur	1	project	projet	3
convinced	convaincus	1	projectile	projectile	80
cotton	coton	5	projectiles	projectiles	10
could	ne	2	proposal	proposition	10
could	pouvait	5	proposed	proposa	4
country	pays	15	purely	purement	2
cranes	grues	4	pyroxyle	pyroxyle	6
craters	cratères	1	quantity	quantité	3
cried	écria	15	question	question	26
cried	Maston	1	questions	questions	11
crowd	foule	15	radius	rayon	1
curious	curieuse	3	rapidity	rapidité	5
cylinder	cylindre	5	rays	rayons	2
danger	danger	5	reason	raison	5
dangers	dangers	3	rendezvous	rendez	2
day	jour	36	replied	dans	1
days	jours	25	replied	répliqua	2
de	A	1	replied	répondit	55
dear	cher	8	reply	répondre	5

Anglais	*Français*	*Nb*	*Anglais*	*Français*	*Nb*
December	décembre	5	reserved	réservé	4
decision	décision	1	resistance	résistance	2
degrees	degrés	5	rifle	rifle	5
departure	départ	7	rifles	rifles	1
depth	profondeur	1	rival	rival	5
deputies	députés	5	rivalry	rivalité	5
diameter	diamètre	5	rivals	rivaux	4
diameter	neuf	1	Rodman	Rodman	7
different	divers	5	round	autour	17
difficult	difficile	7	said	dit	58
difficulties	difficultés	12	same	même	11
difficulty	difficulté	3	satellite	satellite	15
dimensions	dimensions	6	say	dire	13
direct	directe	4	scene	scène	5
director	directeur	6	science	science	13
disappeared	disparurent	4	scientific	scientifique	3
disc	disque	4	scientific	scientifiques	2
discussion	discussion	11	sea	mer	2
dispatch	dépêche	1	second	second	1
distance	distance	25	second	seconde	3
diversity	diversité	1	seconds	secondes	4
dollars	dollars	11	secretaries	secrétaires	2
doubt	doute	7	secretary	secrétaire	9
doubtless	doute	3	secrets	secrets	3
duel	duel	2	see	ne	1
During	Pendant	26	seems	semble	2
ears	oreilles	3	send	envoyer	4
earth	elle	1	separates	sépare	1
earth	Terre	40	September	septembre	6
earth	terrestre	1	series	série	1
easy	facile	5	serpents	serpents	1
eclipse	éclipse	2	seven	sept	11
effect	effet	10	she	elle	16
eight	huit	11	shell	bombe	5
eleven	onze	1	shell	obus	6
Elphinstone	Elphiston	5	shock	contrecoup	1
enclosure	enceinte	4	shot	boulet	21
enemy	ennemi	6	shot	poudre	1
energy	énergie	2	side	côté	4
engineer	ingénieur	2	sides	parois	1
English	Anglais	5	silence	silence	12
enormous	énorme	10	simple	simple	8
enterprise	entreprise	14	sir	monsieur	6
escort	escorte	2	Sir,	Monsieur	4
establish	établir	1	six	cents	3
Europe	Europe	4	six	six	20
evening	soir	5	sixty	soixante	8
every	chaque	7	sky	ciel	2
evidently	évidemment	5	some	quelques	10
example	exemple	2	space	espace	7
excavation	travaux	1	spectacle	spectacle	6
excellent	excellente	1	spectators	spectateurs	7
existence	existence	4	square	carré	1
experiment	expérience	18	stars	étoiles	6

Anglais	*Français*	*Nb*	*Anglais*	*Français*	*Nb*
exquisite	exquis	1	State	État	2
extreme	extrême	2	States	États	8
face	face	8	steam	vapeur	2
face	faisait	1	steamer	steamer	3
fact	effet	4	steel	acier	1
facts	faits	3	Stockholm	Stockholm	2
fall	chute	10	stone	pierre	9
favorable	favorable	1	Stones	s	3
fears	craintes	3	straight	droit	1
Federal	fédérale	3	streets	rues	2
feet	cents	3	studies	études	2
feet	pieds	33	subscription	souscription	4
feet	profondeur	1	success	succès	10
feet	toises	2	successful	réussir	1
few	quelques	13	succession	succession	1
field	champ	4	successively	successivement	4
fifty	cinquante	10	suddenly	subitement	1
figure	figure	2	sufficient	suffisante	1
fire	feu	7	sum	somme	8
first	premiers	4	summit	sommet	2
five	cinq	20	sun	Soleil	12
Florida	Floride	17	surface	surface	13
Florida	Texas	1	surprise	surprise	3
flotilla	flottille	1	table	table	1
following	lendemain	10	tables	tables	1
for	pour	71	Tampa	Tampa	2
forehead	front	1	Tampico_	Tampico_	2
formation	formation	1	telegram	télégramme	4
formidable	formidables	3	telegraph	télégraphe	3
formulae	formules	1	telescope	télescope	8
fortune	fortune	2	temperature	température	4
four	quatre	17	terrestrial	terrestre	1
France	France	4	Texas	Texas	6
francs	francs	3	thanks	grâce	6
French	Français	6	That	°	2
Frenchman	Français	9	That	ci.	1
friend	ami	20	The	L	12
friends	amis	14	The	La	31
furnaces	fours	4	The	Le	50
gained	aux	1	The	Les	21
gas	gaz	12	The	pas	7
generally	généralement	3	their	leur	33
genius	génie	1	their	leurs	18
Germans	Allemands	2	them	leur	11
gigantic	gigantesque	6	then	donc	25
globe	globe	11	theory	théorie	3
government	gouvernement	3	There	Là	11
great	grande	14	there	y	22
ground	sol	4	these	ces	55
gun	canon	11	they	ils	61
gunpowder	poudre	4	thickness	épaisseur	3
guns	canons	3	thickness	pieds	1
had	avait	63	third	troisième	4
had	est	3	thirty	trente	11

Anglais	*Français*	*Nb*	*Anglais*	*Français*	*Nb*
half	demi	12	this	cette	64
hands	mains	6	thousand	mille	23
hardness	dureté	3	three	trois	26
has	a	24	through	travers	3
have	au	5	thunder	tonnerre	2
have	eût	4	time	temps	13
having	avoir	6	times	fois	18
He	Il	20	too	trop	15
heat	chaleur	7	toward	vers	9
heavens	ciel	2	towns	villes	4
heavy	lourd	4	trace	trace	2
her	elle	1	travelers	voyageurs	10
her	Lune	1	trees	arbres	2
her	son	7	Tribune_	Tribune_	2
hero	héros	3	tube	tube	5
Herschel	Herschell	5	twelve	douze	5
hesitation	hésitation	2	twenty	vingt	15
Hillisborough	Hillisboro	1	two	deux	63
him	Ardan	1	under	sous	14
him	lui	17	Union	Union	13
himself	se	14	unknown	inconnu	12
his	bras	1	Up	Jusqu	3
his	ses	42	upon	sur	34
His	Son	41	us	nous	13
hitherto	jusqu	1	velocity	vitesse	20
homogeneous	homogène	2	veritable	véritable	4
honor	honneur	7	visit	visite	2
horizon	horizon	12	wait	attendre	3
hot	chaud	4	war	guerre	18
hour	heure	4	was	avait	10
hours	heures	13	was	est	14
However	Cependant	5	was	était	86
hundred	cent	9	was	fut	74
hundred	cents	7	was	je	1
Hurrah	Hurrah	3	was	nous	1
hurrahs	hurrahs	4	was	vous	2
I	au	2	Washington	Washington	4
I	j	13	water	eau	8
I	Je	61	we	nous	45
I	mondes	1	we	se	1
I	se	1	weight	livres	1
I	son	1	weight	poids	16
I	vous	2	Well	bien	1
idea	idée	6	were	étaient	9
ideas	idées	5	were	furent	8
if	si	23	What	Quelle	5
ignorance	ignorance	4	wheel	maçonnerie	1
immense	immense	15	wheel	rouet	3
Impey	Impey	2	when	quand	13
impossible	impossible	9	whether	si	1
In	En	17	which	qui	104
influences	influences	2	while	tandis	10
inhabitants	habitants	12	who	qui	43
insulted	insulté	1	will	fera	2

Anglais	*Français*	*Nb*	*Anglais*	*Français*	*Nb*
intensity	intensité	4	will	vous	2
invariably	invariablement	2	wings	ailes	2
invented	inventait	1	with	avec	91
inventors	inventeurs	1	without	sans	29
iron	fer	6	wood	bois	6
iron	fonte	21	words	paroles	12
is	est	123	workmen	ouvriers	7
it	se	9	world	monde	18
its	sa	23	worlds	mondes	1
its	son	27	worthy	digne	10
J	J.	1	would	eût	4
jealousy	jalousie	2	would	Si	1
Jupiter	Jupiter	3	Yankee	Yankee	5
large	poudre	1	Yankees	Yankees	5
later	tard	9	yards	yards	14
learn	apprendre	2	year	année	1
length	longueur	6	years	ans	9
less	moins	23	Yes	Oui	11
life	vie	5	yet	encore	6
light	lumière	4	you	au	1
like	comme	28	you	dans	2
limits	limites	6	you	se	1
Lisbon	Lisbonne	1	you	votre	1
little	peu	12	you	vous	98
Louisiana	Louisiane	3	your	votre	5
lunar	lunaires	1	your	vous	1
M	M	1	zenith	zénith	13

A-XIV Définition des termes employés dans un sens spécifique

Alignement : opération consistant à faire correspondre, au sein de deux textes parallèles, les *segments* qui sont en relation d'équivalence, dans les limites de la *compositionnalité traductionnelle*. Par extension, on désigne par alignement le produit de cette opération, c'est-à-dire un ensemble de segments appariés (ou *binômes*).

Analyse contrastive : étude des transformations linguistiques impliquées dans le *transcodage* d'un énoncé de langue source en langue cible.

Assignement *:* ensemble d'appariements de mots entre deux phrases.

Bijection : propriété des *segments* en relation de traduction, impliquant que chaque segment source a un équivalent et un seul dans la cible, et réciproquement. On parle de *quasi-bijection* lorsque cette propriété est vérifiée pour au moins 80 % des segments.

Binôme : couple de *segments* appariés lors de l'*alignement* (un segment regroupe en général zéro, une ou plusieurs phrases).

Bi-texte (Multi-texte) : base de données textuelle incluant deux (ou plus de deux) textes en relation d'*équivalence traductionnelle*, segmentés en unité plus petites (généralement des phrases), et dont les unités équivalentes sont appariées, i.e. reliées entre elles.

Chemin d'alignement : suite de transitions (regroupements de zéro, une ou plus phrases contiguës de chaque côté du bi-texte) permettant de représenter, de proche en proche, l'ensemble de l'*alignement*.

Choix de traduction : la notion de choix découle de l'impossibilité de traduire sans sacrifier certaines virtualités interprétatives du texte original. Le traducteur se trouve

donc aux prises avec des couples antagonistes (tel que « esprit » / « lettre », signifiant / signifié, local / global, connotation / dénotation) dont il doit privilégier un terme au détriment de l'autre. Les choix de traduction se manifestent aussi de manière positive au travers de la fonction pragmatique assumée par l'acte de traduire, avec des différences de stratégie telle que sourcier *vs* cibliste.

Cognats : mots apparentés susceptibles d'être en relation d'*équivalence traductionnelle*, et présentant des similitudes dans leur graphie.

Compositionnalité traductionnelle : propriété indiquant la possibilité de décomposer la relation d'*équivalence traductionnelle*, définie d'abord globalement au niveau des textes sources et cibles, en sous-relations définies entre des unités d'un rang inférieur : chapitre, paragraphe, phrases, unités lexicales.

Coordonnées pragmatiques : ensemble des paramètres liant le message à une situation donnée. Nous distinguons les *coordonnées situationnelles*, relatives aux situations singulières de production et de réception du message (temps, lieu, participants, référents particuliers, …), et les *coordonnées contextuelles*, relatives aux codifications sociales (codes linguistiques, références culturelles et encyclopédiques, autres systèmes sémiotiques…) pertinentes dans le contexte de cette situation.

Correspondances lexicales : ensemble de couples d'unités lexicales en relation d'*équivalence traductionnelle* à l'intérieur des segments appariés. Les correspondances dépendent bien sûr de la définition que l'on donne du concept d'unité lexicale (qui peut inclure, par extension, les expressions terminologiques et idiomatiques, ou non) et de celui d'équivalence traductionnelle (équivalence dynamique, dénotative, connotative, réutilisabilité dans d'autres contextes, etc.). Avec l'alignement, la segmentation des textes découlait des appariements, en imposant des regroupements au niveau des phrases ou des paragraphes. A l'inverse, l'extraction des correspondances lexicales présuppose une segmentation préalable des unités, indépendante des choix de traduction locaux, et les contraintes de parallélisme sont abandonnées.

Désignation : rapport extrinsèque entre le signe et la réalité extra-linguistique qu'il désigne (le *designatum*, qui peut être un *concept* ou un *référent* concret).

Designatum : extrémité extra-linguistique de la relation de désignation. Qu'il s'agisse d'une classe d'objets du monde, d'un concept explicitement défini ou issu d'une pratique sociale, ou d'une représentation psychologique implicite, le rapport de désignation est le même d'un point de vue linguistique.

Diagonale d'un bi-texte : ligne théorique correspondant à l'appariement d'unités dont la position relative est identique dans les deux textes.

Equivalence traductionnelle : relation d'équivalence établie entre un texte source et le texte cible produit au cours de l'acte de traduire. Cette relation peut se situer sur différents niveaux en fonction des *choix de traduction* : niveau de l'effet sur le récepteur (*équivalence dynamique*), niveau de la désignation (*équivalence dénotationnelle*), niveaux connotatif, rhétorique ou expressif (valeur poétique), niveau des contenus linguistiques (*équivalence sémantique* définie au niveau des unités linguistiques). Notons que ce dernier niveau n'est pas explicitement recherché dans l'acte de traduire mais peut en découler (il s'agit d'une équivalence de moyen et non de fin).

Exégèse : étude des facteurs déterminant la production et l'interprétation du texte traduit, dans la perspective globale de la traduction comme acte de communication.

Idiome : ensemble des productions du code conformes à l'usage. L'*idiome* constitue un sous-ensemble des réalisations autorisées par la grammaire et le lexique. Les réalisations de l'*idiome* sont considérées par le locuteur natif comme plus « naturelle » ou habituelle (p. ex. « faire des courses » est plus naturel que « acheter ses commissions »).

Lexie, lexème, phrasème : par *lexie* nous désignons l'unité lexicale en général. Le *lexème* est une lexie simple (un mot, une forme simple), le *phrasème* est une lexie composée (une unité polylexicale au sens de Gross).

Littéralité vs mot-à-mot : nous définissons la traduction *littérale* comme le produit d'un *transcodage*, dans le respect des transformations nécessitées par la traduction des unités polylexicales et par le respect de l'idiome d'arrivée. Le *mot à mot* est en deçà du transcodage, puisqu'en se situant au seul niveau des équivalences lexicales, il ne tient pas compte des contraintes de l'idiome d'arrivée (p. ex. traduire *it makes cold* pour *il fait froid*).

Métataxe : transformations syntaxiques intervenant dans le passage d'un *idiome* vers un autre.

Monotonie : propriété des segments équivalents lorsqu'ils qui apparaissent dans le même ordre séquentiel dans la source et dans la cible. On parle de quasi-monotonie quand cette propriété est vérifiée pour au moins 80 % des segments.

Point d'ancrage : couple d'unités dont l'appariement est considéré comme fiable, entre lesquels les zones sont présumées alignées.

Segment : portion de texte regroupant une ou plusieurs unités issues de la segmentation primaire (p. ex. paragraphes, phrases) des textes.

Sens : objet de la situation de communication, lié au vouloir-dire de l'émetteur et à l'interprétation du récepteur. Par opposition à la *signification*, le *sens* est une valeur instantanée, singulière, liés à des facteurs situationnels particuliers.

Signification : valeur sémantique conventionnelle attachée à une unité ou une construction linguistique. A la différence du *sens*, la signification est inscrite de manière stable dans le code linguistique.

Textes parallèles : textes en relation d'équivalence traductionnelle satisfaisant aux critères de parallélisme, i.e. segmentables en unités d'un rang inférieur (généralement en paragraphes ou en phrases) et dont les segments respectent les conditions de *monotonie* et de *bijection*. Le parallélisme est une propriété scalaire, suivant la taille des segments concernés, et le respect plus ou moins total de ces deux conditions.

Transcodage : transformation d'un énoncé en langue source en un énoncé équivalent en langue cible, en ne tenant compte que des contraintes linguistiques. Le *transcodage* opère sur les *significations* tandis que la traduction porte sur le *sens*.

Transfuge : chaîne de caractère inchangée dans le passage à la traduction.

Version source, version cible : les deux textes composant un bi-texte (dénomination purement conventionnelle puisque nous définissons le bi-texte de manière symétrique, la relation d'équivalence n'étant pas orientée).

Bibliographie

Abbou, A., sous la dir. de (1989) *Traduction Assistée par Ordinateur, Perspectives technologiques, industrielles et économiques envisageables à l'horizon 1990*, Paris, DAICADIF.

Adam, J.-M. (1992) *Les textes : types et prototypes*, Paris, Nathan, Fac, 223 p.

Ahrenberg, L., Andersson, M., Merkel, M. (1998) A simple hybrid aligner for generating lexical correspondences in parallel texts. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montréal, Canada, 10-14 August 1998, pp. 29-35.

Association pour le traitement automatique des langues (1996) *Traitements probabilistes et corpus*, Paris.

Austin, J. L. (1962, 1970) *Quand dire c'est faire*, Paris, Editions du Seuil.

Bailly, G., Perrin, A., Yves, L. (1988) Common Approaches in Speech Synthesis and Automatic Translation of Text. In *Bulletin du Laboratoire de la Communication Parlée*, Grenoble, 2B, pp. 295-311.

Bar-Hillel, Y. (1951) The State of Machine Translation in 1951. *American Documentation*, vol. 2, pp. 229-237.

Bar-Hillel, Y. (1960) The present status of Automatic Translation of Languages. *Advances in Computers*, New-York, Academic Press, 1, pp. 91-163.

Bar-Hillel, Y. (1964) The future of Machine Translation. *Language and Information : Selected Essays on their Theory and Application*, London, Addison-Wesley, pp. 180-184.

Bathgate, R. H. (1980) A communicative model of translation. *Incorporated linguist*, 19(4) 20(1), pp. 113-144.

Bellman, R. (1957) *Dynamic programming*, Princeton, Princeton University Press.

Benjamin, W. (1923, 1993) Il compito del traduttore. In Nergaard, S. (a cura di) *La teoria della traduzione nella storia*, Milano, Strumenti Bompiani, pp. 221-236.

Benveniste, E. (1966) *Problèmes de linguistique générale*, Paris, Gallimard, tel.

Berger, A. L., Della Pietra, S. A., Della Pietra, V. J. (1996) A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, vol. 22, n. 1, pp. 41-71.

Berman, A. (1984) *L'épreuve de l'étranger. Culture et traduction dans l'Allemagne romantique*, Paris, Gallimard.

Biber, D. (1988) *Variation across speech and writing*, Cambridge, Cambridge University Press.

Biber, D. (1993) Representativeness in Corpus Design. *Literary and Linguistic Computing*, vol. 8, n. 4, pp. 243-257.

Biber, D., Conrad, S., Reppen, R. (1998) *Corpus Linguistics : Investigating language structure and use*, Cambrigdge University Press, 300 p.

Bindi, R., Calzolari, N., Monachini, M., Pirelli, V., Zampolli, A. (1994) Corpora and Computational Lexica : Integration of Different Methodologies of Lexical Knowledge Acquisition. *Literary and Linguistic Computing*, vol. 9, n. 1, pp. 29-46.

Blanc, E. (2000) From the UNL hypergraph to GETA's multilevel tree. In *Proceedings of MT 2000*, University of Exeter, Great Britain, 20-22 novembre 2000, 14-1 :14-9.

Blanchon, H. (1991) Problèmes de désambiguïsation interactive en TAO Personnelle. In Clas, A., Safar, H. (sous la dir. de), *Actes du colloque de Mons, L'environnement traductionnel ; La station de travail du traducteur de l'an 2001*, Mons, Belgique, 25-27 avril 1991, pp. 31-47.

Blanchon, H. (1994) Perspectives en TAFD pour auteur monolingue après une première expérience : la maquette LIDIA-1. In *Actes de TALN-94*, Marseille, 7-8 avril 1994, pp. 12-23.

Blank, I. (1995) Sentence alignement : methods and implementations. *T.A.L.*, vol. 36, n. 1-2, pp. 81-99.

Blank, I. (1998) Computer-aided analysis of multilingual patent documentation. In *Proceedings of First International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain, 28-30 May 1998, pp. 765-774.

Blank, I. (1998) Lexical knowledge extraction from technical texts. *Third European Robotics, Intelligent Systems and Control Conference,* Athens, Greece, 22-25 June 1998.

Boitet, C. (1991) Quelle automatisation de la traduction peut-on souhaiter et réaliser sur les stations de travail individuelles ? In Clas, A., Safar, H. (sous la dir. de), *Actes du colloque de Mons, L'environnement traductionnel ; La station de travail du traducteur de l'an 2001*, Mons, Belgique, 25-27 avril 1991, pp. 3-19.

Boitet, C. (1993) La TAO comme technologie scientifique : la cas de la traduction automatique fondée sur le dialogue. In Bouillon, P., Clas, A. (sous la dir. de), *La traductique*, Montréal, Les presses de l'université de Montréal, pp. 109-148.

Boitet, C. (1997) GETA's methodology and its current developments. In *Proceedings of PACLING'97*, Meisei University, Ohme, Japon, septembre 1997, pp. 23-57.

Bonhomme, P., Romary, L. (1995) The Lingua Parallel Concordancing Project, Managing Multilingual Texts for Educational Purposes. In *Proceedings of Language Engineering 95*, Montpellier, June 26-30, 1995 (disp. à l'adresse : http://www.loria.fr).

Bouillon, P., Boesefeldt, K. (1992) La traduction automatique des bulletins d'avalanches. In Clas, A., Safar, H. (sous la dir. de), *Actes du colloque de Mons, L'environnement traductionnel ; La station de travail du traducteur de l'an 2001*, Mons, Belgique, 25-27 avril 1991, pp.69-78.

Bouillon, P., Boesefeldt, K. (1992) Problèmes de traduction automatique dans le sous-langage des bulletins d'avalaches. *META*, Outremont, PQ, déc. 37 :4, pp. 625-646.

Bouillon, P., Clas, A., sous la dir. de (1993) *La traductique*, Montréal, Les presses de l'université de Montréal.

Bourigault, D. (1992) Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, Nantes, pp. 977-981.

Bourquin, G. (1991) Que peut-on automatiser en traduction ? *La traduction littéraire scientifique et technique, Actes du Colloque International organisé par l'AEPLP, 21-22 mars 1991*, Paris, La Tilu éditeur.

Bourquin, G. (1993) Préalables linguistiques à la construction d'un système de traduction automatique. In Bouillon, P., Clas, A. (sous la dir. de), *La traductique*, Montréal, Les presses de l'université de Montréal, pp. 27-34.

Boutsis, S., Piperidis, S. (1996) Automatic extraction of lexical equivalences from parallel corpora. *Workshop on Multilinguality in the Software Industry: the AI Contribution (MULSAIC'96), 12th European Conference on Artificial Intelligence (ECAI'96)*, Budapest, Hungary, 11-16 August 1996, pp. 27-31.

Boutsis, S., Piperidis, S. (1998) Aligning clauses in parallel texts. *Third Conference on Empirical Methods in Natural Language Processing*, Granada, Spain, 2 June 1998, pp. 17-26.

Boutsis, S., Piperidis, S. (1998) OK with alignment of sentences. What about clauses?. In *Proceedings of the Panhellenic Conference on New Information Technology-(NIT'98*), Athens, Greece, 8-10 October 1998, pp. 288-297.

Brown, P., Chen, S., Della Pietra, S., Della Pietra, V., Kehler, S., Mercer (1992c) Automatic speech recognition in machine aided translation. *Computer Speech and Language*, vol. 8, pp. 177-187.

Brown, P., Cocke, J., Della Pietra, S., Jelinek, F., Lafferty, J., Mercer, R., Roossin P. (1990) A Statistical Approach to Machine Translation. *Computational Linguistics*, vol. 16, n. 2, pp. 79-85.

Brown, P., Della Pietra, S., Della Pietra, V., Lai, J., Mercer, R. (1992a) An Estimate of an Upper Bound for the Entropy of English. *Computational Linguistics*, vol. 18, n. 1, pp. 31-40.

Brown, P., Della Pietra, S., Della Pietra, V., Mercer, R. (1992b) Word-Sense Disambiguation Using Statistical Methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, ACL-91*, pp. 264-270.

Brown, P., Della Pietra, S., Della Pietra, V., Mercer, R. (1993a) The Mathematics of Statistical Machine Translation : Parameter Estimation. *Computational Linguistics*, vol. 19, n. 2, pp. 263-311.

Brown, P., Della Pietra, V., De Souza, P., Lai, J., Mercer, R. (1993b) Class-Based n-gram Models of Natural Language. *Computational Linguistics*, vol. 18, n. 4, pp. 467-479.

Brown, P., Lai, J., Mercer, R. (1991) Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, ACL-91*, Morristown, NJ, pp. 169-176.

Brown, R. D. (1996) Example-Based Machine Translation in the Pangloss System. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, Copenhagen, 5-9 August 1996, pp. 125-130 (disp. à l'adresse : http:// www.cs.cmu.edu).

Brown, R. D., J. G. Carbonell, Yang Yiming (2000) Automatic dictionary extraction for cross-language information retrieval. In Véronis, J. (Ed.), *Parallel Text Processing*, Dordrecht, Netherlands, Kluwer Academic Publishers, § 14, 24 p.

Catford, J. C. (1965) *A Linguistic Theory of Translation*, London, Oxford University Press.

Catizone, R. G., Russell, G., Warwick, S. (1993) Deriving Translation Data from Bilingual Texts. In *Proceedings of 1st Lexical Acquisition Workshop*, Detroit, Michigan, 7 p.

Chang, J. J. S., Ker, S. J. (1996) Aligning More Words with High Precision for Small Bilingual Corpora. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, Copenhagen, 5-9 August 1996.

Charniak, E. (1993) *Statistical Language Learning*, Cambridge, MU, The MIT Press.

Charteris Black, J. (1999) A Comparative Approach of English and Malay Idioms and Phraseological Units. *Actes du colloque Linguistique contrastive et Traduction Approches Empiriques*, Louvain-la-Neuve, 5-6 février 1999, pp. 25-26.

Chen, S. (1993) Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, ACL-93*, Columbus, Ohio, pp. 9-16.

Choueka, Y., Conley, E. S., Dagan, I. (2000) A comprehensive bilingual alignment system – Application to disparate Languages : Hebrew and English. In Véronis, J. (Ed.), *Parallel Text Processing*, Dordrecht, Netherlands, Kluwer Academic Publishers, § 4, 28 p.

Church, K. W. (1993) Char align : A program for Aligning Parallel Texts at the Character Level. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, ACL-93*, Columbus Ohio, pp. 1-8.

Church, K. W., Hanks, P. (1990) Word Association Norms, Mutual Information, and Lexicography. *Machine Translation*, vol. 16, n. 1, pp. 22-29.

Church, K. W., Hovy, E. H. (1993) Good Applications for Crummy Machine Translation. *Machine Translation*, Dordrecht, Netherlands, vol. 8, n. 4, pp. 239-258.

Church, K. W., Mercer, R. L. (1993) Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics*, vol. 19, n. 1, pp. 1-24.

Clas, A. (1994) Collocations et Langues de spécialité. *META*, Outremont, PQ, XXXIX, 4, pp. 576-580.

Clas, A., Safar, H., sous la dir. de (1992) *Actes du colloque de Mons, L'environnement traductionnel ; La station de travail du traducteur de l'an 2001*, Mons, Belgique, 25-27 avril 1991.

Collins, B., Cunningham, P. (1995) A methodology for EBMT. *4th International Conference on the Cognitive Science of Natural Processing*, Dublin.

Cowie J., Dunning, T., Guthrie, L., Wilks, Y. (1994) Text Processing Using Multilingual Resources at the Computing Research Laboratory. *Literary and Linguistics Computing*, vol. 9, n. 1, pp. 65-78.

Cranias, L., Papageorgiou, H., Piperidis, S. (1997) Example retrieval from a translation memory. *Journal of Natural Language Engineering, 3*, February 1997, pp. 255-277.

Daelemans, W, van den Bosch, A., Buchholz, S., Veenstra, J., Zavrel, J. (1998) Memory-Based Word Sense Disambiguation for Senseval, 11 p.

Dagan, I., Church, K. W., Gale, W. A. (1993) Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora : Academic and Industrial Perspective*, Columbus, Ohio, pp. 1-8.

Dagan, I., Itai, A. (1994) Word Sense Disambiguation Using a Second Language Monolingual Corpus. *Computational Linguistics*, vol. 20, n. 4, pp. 563-596.

Dagan, I., Itai, A., Shwall, U. (1991) Two Languages Are More Informative Than One. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, ACL-91*, Morristown, NJ, pp. 130-137.

Dagan, I., Pereira, F., Lee, L. (1994) Similarity-Based Estimation of Word Cooccurrence Probabilities. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, ACL-94*, New Mexico State University, June 1994, pp. 272-278.

Daille, B.,Gaussier, E., Langé, J.-M. (1994) Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, Kyoto, Japon, pp. 712-716.

Daille, B. (1994) *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*, *Thèse de doctorat*, Paris, Université de Paris VII.

Danlos, L., Laurens, O. (1991) Présentation du projet EUROTRA et des grammaires d'Eurotra-France. *Rapports techniques d'Eurotra-France*, Paris, TALANA.

Davis, M. (1998) On the effective use of large parallel corpora in cross-language text retrieval. In Grefenstette, G. (Ed.) *Cross-Language Information Retrieval.* Boston, Kluwer Academic Publishers, pp. 11-23.

Davis, M. W., Dunning T. E., Ogden W. C. (1995) Text Alignment in the Real World : Improving Alignments of Noisy Translations Using Common Lexical Features. In *Proceedings of EACL 95*, 8 p. (disp. à l'adresse : http://www.crl.nmsu.edu).

Davis, M., Dunning, T. (1995) Query translation using evolutionary programming for multilingual information retrieval. In *Proceedings of the Fourth Annual Conference on Evolutionary Programming*, March 1995.

Davis, M., Dunning, T. (1996) Query translation using evolutionary programming for multilingual information retrieval II. In *Proceedings of the Fifth Annual Conference on Evolutionary Programming.*

Davis, M., Ogden, W.C. (1997) Free resources, advanced alignment for cross-language text retrieval. In Harman, D.K. (Ed.), *NIST Special Publication: The Sixth Text Retrieval Conference (TREC-6)*, Computer Systems Laboratory, NIST.

Davis, M., Ogden, W.C. (1997) QUILT: implementing cross-language text retrieval systems for large-scale text collections. *SIGIR97*, Philadelphia, PA, August 1997.

Davis, M., Ren, F. (1998) Automatic Japanese-Chinese parallel text alignment. In *Proceedings of the International Conference on Chinese Information Processing (ICCIP 98)*.

Débili, F. (1997) L'appariement : quels problèmes ? *1ères JST 1997 FRANCIL de l'AUPELF-UREF*, Avignon, 15-16 avril 1997, pp. 199-206.

Débili, F., Sammouda, E. (1992) Appariement des phrases de textes bilingues Français - Anglais et Français - Arabe. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING-92*, Nantes, 23-28 août 1992, pp. 518-524.

Delavenay, E. (1963) La machine à traduire., PUF, Que sais-je ?, 128 p.

Dempster, A., Laird, N., Rubin, D. (1977) Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society*, 39 (B), pp. 1-38.

Doi, S., Muraki, K. (1992) Translation Ambiguity Resolution based on Text Corpora of Source and Target Languages. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING-92*, Nantes, 23-28 août 1992, pp. 525-531.

Dolet, E. (1540, 1963) La manière de bien traduire d'une langue en aultre. In Edmond Cary *Les grands traducteurs français*, Genève, Georg & Cie, pp. 5-14.

Donovan, C. (1990) *La fidélité en interprétation, thèse de doctorat*, Paris, Université Paris III.

Dunning, T. (1993) Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, Morristown, NJ, vol. 19, n. 1, pp. 61-74.

Eco, U. (1995) Riflessioni teorico-pratiche sulla traduzione. In Nergaard, S. (a cura di), *teorie contemporanee della traduzione*, Milano, Strumenti Bompiani, pp.121-146.

El-Bèze, M. (1994) Exemples de domaines d'application des modèles de langage probabiliste. *Publications scientifiques et techniques d'IBM France*, juin 1994, pp. 33-57.

Fairon, C., Senellart, J. (1999) Réflexions sur la localisation, l'étiquetage, la reconnaissance et la traduction d'expressions linguistiques complexes. In *Actes de TALN'99*, Cargèse, Corse, 12-17 juillet 1999, pp. 135-143.

Federici S., Montemagni, S., Pirelli, V., Calzolari, N. (1997) Analogy-based Extraction of Lexical Knowledge from Corpora : the SPARKLE Experience, pp. 75-81.

Fluhr, C., Bisson, F., Elkateb, F. (2000) Parallel text alignment using crosslingual information retrieval. In Véronis, J. (Ed.), *Parallel Text Processing*, Dordrecht, Netherlands, Kluwer Academic Publishers, § 9, 14 p.

Fluhr, C., Schmit, D., Ortet, Ph., Elkateb, F., Gurtner, K., Radwan, R. (1998) Distributed cross-lingual information retrieval. In Grefenstette G. (Ed.), *Cross-Language Information Retrieval*, Dordrecht, Netherlands, Kluwer Academic Publishers.

Fuchs, C. (1982) *La Paraphrase*, Paris, PUF.

Fuchs, C., avec la collaboration de Lacheret-Dujour, A., Victorri, B., Danlos, L., Luzzati, D. (1993) *Linguistique et Traitements Automatiques des Langues*, Paris, Hachette.

Fung, P. (1995a) A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. In *Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics*, Boston, Massachussetts, March 1995, pp. 226-233 (disp. à l'adresse : http://www.cs.columbia.edu).

Fung, P. (1995b) Compiling Bilingual Lexicon Entries From a Non-Parallel English-Chinese Corpus. In *Proceedings of Third Workshop on Very Large Corpora (WVLC3) at ACL-95*, *Boston*, *Massachusetts*, June 1995, pp. 173-183.

Fung, P. (2000) A statistical view on bilingual lexicon extraction - From parallel corpora to non-parallel corpora. In Véronis, J. (Ed.), *Parallel Text Processing*, Dordrecht, Netherlands, Kluwer Academic Publishers, § 11, p. 18.

Fung, P., Church, K. W. (1994) K-vec : A New approach for Aligning Parallel Texts. In *Proceedings of the 15th International Conference on Computational Linguistics*, *COLING-94*, Kyoto, pp. 1096-1102.

Fung, P., Mc Keown, K. (1994) Aligning noisy parallel corpora across language group : Word pair feature matching by dynamic time warping. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, pp. 81-88.

Fung, P., McKeown, K. (1997) Finding Terminology Translations from Non-parallel Corpora. *The 5th Annual Workshop on Very Large Corpora*, Hong Kong, August 1997, pp. 192-202.

Fung, P., Wu, D. (1994) Statistical augmentation of a Chinese machine-readable dictionary. In *Proceedings of the 2nd Annual Workshop on Very Large Corpora*, Kyoto, pp. 69-85.

Furuse, O., Iida, H. (1992) An example-based Method for Transfer-driven Machine Translation. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, *TMI-92*, Montréal, CCRIT, pp. 67-81.

Gale, W., Church, K. W. (1993) A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, vol. 19, n. 1, pp. 75-91.

Gaussier, E. (1995) *Modèles statistiques et patrons morphosyntaxiques pour l'extraction de lexiques bilingues*, Thèse, Université de Paris VII.

Gaussier, E., Hull, D., Aït-Mokhtar, S. (2000) Term alignment in use. In Véronis, J. (Ed.), *Parallel text Processing*, Dordrecht, Netherlands, Kluwer Academic Publisher, § 13, p. 22.

Gaussier, E., Langé, J.-M. (1995) Modèles statistiques pour l'extraction de lexiques bilingues. *T.A.L.*, vol. 36, n. 1-2, pp. 133-155.

Gouadec, D., sous la resp. de (1993) *Terminologie et terminotique, outils, modèles et méthodes : actes de la première Université d'automne en terminologie*, Rennes 2, 21 au 26 septembre 1992.

Granger, G.-G. (1967) *Pensée Formelle et Sciences de l'Homme*, Paris, Aubier – Montagne.

Grefenstette, G., Tapanainen, P. (1994) What is a word, What is a sentence ? Problems of Tokenization. In *Proceedings of the 3rd International Conference on Computational Lexicography*, Budapest, pp. 79-87.

Greimas, A. J. (1970) *Du Sens, Essais sémiotiques*, Paris, Editions du Seuil.

Greimas, A. J., Courtès, J. (1993) *Sémiotique*, Paris, Hachette, Coll. HU linguistique.

Grundy, V. (1996) L'utilisation d'un corpus dans la rédaction du dictionnaire bilingue. In Béjoint, H., Thoiron, P. (sous la dir. de) Les dictionnaires bilingues, Louvain-la-Neuve, Duculot, Coll. Universités francophones / Champs linguistiques, pp. 127-149.

Habert, B., Nazarenko, A., Salem, A. (1997) *Les linguistiques de corpus*, Paris, Armand Colin.

Harris, B. (1988a) Bi-text, a new concept in Translation Theory. *Language Monthly*, n° 54, pp. 8-10.

Harris, B. (1988b) Are you Bi-Textual ? *Language Technology*, n° 7, pp. 41-41.

Harris, Z., Gottfried, M., Ryckman, T., Mattick, Jr P., Daladier, A., Harris, T., Harris, S. (1989) *The Form of Information in Science, Analysis of Immunology Sublanguage*, Dordrecht, Netherlands, Kluwer Academic Publisher.

Haruno, M., Yamazaki, T. (1996) High-performance bilingual text alignment using statistical and dictionary information. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, University of California, Santa Cruz, California, 24-27 June 1996, pp. 131-138.

Heid, U. (1993) Le lexique : quelques problèmes de description de représentation lexicale pour la traduction automatique. In Bouillon, P., Clas, A. (sous la dir. de), *La traductique*, Montréal, Les presses de l'université de Montréal, pp. 167-196.

Hofland, K. (1995) A program for Aligning English and Norwegian Sentences. In *Proceedings of ACH/ALLC 95*, Santa Barbara, 11-15 July 1995, 13 p. (disp. à l'adresse : http://www.hd.uib.no/enpc.html).

Hughes, 0. J., Souter, C., Atwell, E. (1995) Automatic Extraction of Tagset Mappings from Parallel-Annotated Corpora. In *Proceedings of EACL-95 SIGDAT Workshop*.

Hull, D. (1998) A practical approach to terminology alignment. In *Proceedings of the First Workshop on Computational Terminology, COLING-ACL '98*, Montreal, Canada, 1998, pp. 1-7.

Hutchins, W. J. (1986) *Machine translation : past, present, future*, Chichester, Ellis Hoorwood Ldt.

Hutchins, W. J. (1996) Computer-based translation systems and tools. *ELRA Newsletter*, vol. 1, n° 4, dec. 1996 (disp. à l'adresse http://www.bcs.org.uk/siggroup\nalatran\nala_018.htm).

Hutchins, W. J., Somers, H. L. (1992) An Introduction to Machine Translation., Londres, Academic Press.

Ikehara, S., Shirai, S., Uchino, H. (1996) A Statistical Method for Extracting Uninterrupted and Interrupted Collocations from Very Large Corpora. In *Proceedings of the 16th International Conference on Computational Linguistics*, *COLING-96*, Copenhagen, 5-9 August 1996, pp. 574-579.

Iordanskaja, L., Kim, M., Lavoie, B., Polguere, A. (1992) Generation of Extended Bilingual Statistical Reports. In *Proceedings of the 14th International Conference on Computational Linguistics*, *COLING-92*, Nantes, 23-28 août 1992, pp. 1019-1023.

Isabelle, P. (1992a) La bi-textualité : vers une nouvelle génération d'aides à la traduction et la terminologie. *META*, Outremont, PQ, XXXVII, 4, pp. 721-731.

Isabelle, P. (1992b) Bi-Textual Aids for Translators. In *Proceedings of the Eight Annual Conference of the UW Centre for the New OED and Text Research*, University of Waterloo, Waterloo, pp. 76-89.

Isabelle, P. (1993) Current Research in Machine Translation : A reply to Harry Somers. *Machine Translation*, *7*, 265-272

Isabelle, P. (1996) The state of machine translation in 1996. *Invited report prepared for the National Research Council of the U.S.A.*, 2 p. (disp. à l'adresse : http://www-rali.iro.umontreal.ca).

Isabelle, P., Dymetman, M. Foster, G., Jutras, J.-M., Macklovitch, E., Perrault, F., Ren, X., Simard, M. (1993) Translation Analysis and Translation Automation. In *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in MT*, Kyoto, pp. 201-217.

Isabelle, P., Macklovitch, E. (1990) Où en est la traduction automatique ? *Actes du colloque annuel CIPS/CATA*, Ottawa, 11 p. (disp. à l'adresse : http://www-rali.iro.umontreal.ca).

Isabelle, P., Simard, M. (1996) Propositions pour la représentation et l'évaluation des alignements de textes parallèles dans l'ARC A2. *Rapport technique*, Laval, Canada, CITI, 6 p. (disp. à l'adresse : http://www-rali.iro.umontreal.ca/arc-a2/PropEval).

Isabelle, P., Warwick-Armstrong, S. (1993) Les corpus bilingues : une nouvelle ressource pour le traducteur. In Bouillon, P., Clas, A. (sous la dir. de), *La traductique*, Montréal, Les presses de l'université de Montréal, pp. 288-306.

Isahara, H. (1998) JEIDA's English-Japanese bilingual corpus project. In *Proceedings of First International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain, 28-30 May 1998, pp. 471-474.

Jacquemin, C. (1991) *Transformation des noms composés*, *Thèse de doctorat*, Paris, Université de Paris VII.

Jakobson, R. (1963) Aspects linguistiques de la traduction. *Essais de linguistique générale*, Paris, Les éditions de Minuit, pp. 78-86.

Jones, D. (1992) Non-hybrid Example-based Machine Translation Architectures. *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, TMI-92*, Montréal, CCRIT.

Jordan, P., Dorr, B., Benoit, J. (1993) A First-Pass Approach for Evaluation of Machine Translation Systems. *Machine Translation*, Dordrecht, Netherlands, 8:1-2, pp. 49-58.

Kaji H., Kida Y., Morimoto Y. (1992) Learning translation templates from bilingual text. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING-92*, Nantes, 23-28 août 1992, pp. 672-678.

Kaji, H., Ono, T. (1996) Extracting Word Correspondences from Bilingual Corpora Based on Word Co-occurrence Information. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96,* Copenhagen, 5-9 August 1996.

Kay M., Röscheisen, M. (1993) Text-Translation Alignement. *Computational Linguistics*, Morristown, NJ, vol. 19, n. 1, pp. 121-142.

Kay, M. (1997) The Proper Place of Men and Machines in Language Translation. *Machine Translation, special issue*, Dordrecht, Netherlands, Kluwer Academic Publisher, vol. 12, nº 1-2, p. 3-23.

Kerbrat-Orecchioni, C. (1977) *La Connotation*, Lyon, P.U.L., 256 p.

Kilgariff, A. (1996) Comparing word frequencies across corpora : Why chi-square doesn't work, and an improved LOB-Brown comparison. In *Proceedings of ALLC-ACH Conference*, pp. 169-172 (disp. à l'adresse : http://www.itri.brighto.ac.uk/~Adam.Kilgarriff/euralex.asc).

King, M. (1993) Sur l'évaluation des systèmes de traduction assistée par ordinateur. In Bouillon, P., Clas, A. (sous la dir. de), *La traductique*, Montréal, Les presses de l'université de Montréal, pp. 261-269.

Kitamura, M., Matsumoto, Y. (1996) Automatic Extraction of Word Sequence Correspondences in Parallel Corpora. In *Proceedings of Fourth Workshop on Very Large Corpora* (*WVLC4*) *at ACL-96.*

Kittrege, R. (1987) The Significance of Sublanguage for Automatic Translation. In Nirenburg, S. (Ed.), *Machine Translation : Theoretical and Methodological Issues*, Cambridge.

Klavans, J., Tzoukermann, E. (1995) Combining Corpus and Machine-Readable Dictionary Data for Building Bilingual Lexicons. *Machine Translation*, Dordrecht, Netherlands, Kluwer Academic Publisher, vol. 10 (3), pp. 185-218.

Kleiber, G. (1990) *La sémantique du prototype*, Paris, PUF.

Kleiber, G. (1999) *Problèmes de sémantique, La polysémie en questions*, Paris, Presses Universitaires du Septentrion.

Knowles, F. (1996) L'informatisation de la fabrication des dictionnaires bilingues. In Béjoint, H., Thoiron, P. (sous la dir. de) Les dictionnaires bilingues, Louvain-la-Neuve, Duculot, Coll. Universités francophones / Champs linguistiques, pp. 150-.

Kraif, O. (1995) *Contribution a l'Elaboration d'un système de Traduction Automatique Basée sur l'Exemple*. *Mémoire de DEA*, Nice, Université de Nice Sophia Antipolis.

Krauwer, S. (1993) Evaluation of MT Systems : A Programmatic View. *Machine Translation*, Dordrecht, Netherlands, 8:1-2, pp. 59-66.

Krovetz, R. (1998) More than One Sense Per Discourse, in *Pilot Senseval, Hot-off-the-Press Papers*, Herstmonceux Castle, 10 p.

Kumano, A., Hideki, H. (1994) Building a MT dictionnary from parallel texts based on linguistic and statistical information. In *Proceedings of the 15th International Conference on Computational Linguistics, COLING-94*, Kyoto, pp. 76-81.

Kupiec, J. (1993) An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics, ACL-93*, Colombus, Ohio, pp. 17-22.

Ladmiral, J.-R. (1986) Sourciers et ciblistes. In Revue d'Esthétique, n°12, pp. 33-41.

Ladouceur, J. (1997) L'alignement de termes complexes plurilingues dans des textes spécialisés : une approche interactive. *1ères JST 1997 FRANCIL de l'AUPELF-UREF*, Avignon, 15-16 avril 1997, pp. 493-500.

Langé, J.-M., Gaussier, É. (1995) Alignement de corpus multilingues au niveau des phrases. *T.A.L.*, vol. 36, n. 1-2, pp. 67-79.

Langlais, P. (1997) Alignement de corpus multilingues : intérêts, algorithmes et évaluations. In *Actes de FRACTAL 1997*, Besançon, pp. 245-254.

Langlais, P., El-Bèze, M. (1997) Alignement de corpus bilingues : algorithmes et évaluation. *1ères JST 1997 FRANCIL de l'AUPELF-UREF*, Avignon, 15-16 avril 1997, pp. 191-197.

Langlais, P., Simard, M., Véronis, J. (1998) Methods and Practical Issues in Evaluating Alignment Techniques. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistic*, Montréal, Canada, August 10-14, 7 p.

Langlais, P., Simard, M., Véronis, J., Armstrong, S., Bonhomme P., Débili, F., Isabelle, P., Souissi, E., Théron, P. (1998) ARCADE: A co-operative research project on bilingual text alignment. In *Proceedings of First International Conference on Language Resources and Evaluation* (*LREC*), *Granada, Spain, 28-30 May 1998*, pp. 289-292.

Laplace, C. (1994) *Théorie du langage et théorie de la traduction*, Paris, Didier érudition.

Larose, R. (1994) Qualité et efficacité en traduction : réponse à F.W. Sixel. *META*, Outremont, PQ, XXXIX vol. 2, pp. 362-373.

Laurière, J. L. (1979) *Eléments de programmation dynamique*, Paris, Gauthiers-Villars.

Lederer, M., Israël, F., réunis par (1991) *La liberté en traduction, Actes du colloque international tenu à l'E.S.I.T. les 7,8 et 9 juin 90*, Paris, Didier érudition, Traductologie.

Lehrberger, J. (1982) Automatic Translation and the Concept of Sublanguage. In Kittrege, R., Lehrberger, J. (Ed.), *Sublanguage : Studies of Language in Restricted Semantic Domains*, Berlin, de Gruyter, pp. 81-106.

Lemaréchal, A. (1989) *Les parties du discours, sémantique et syntaxe*, Paris, PUF, Linguistique nouvelle.

Lerat, P. (1995) *Les langues spécialisées*, Paris, PUF, Linguistique Nouvelle.

Lévi-Strauss, C. (1962) *La pensée sauvage*, Paris, Plon, Press Pocket.

Levý, J. (1967, 1995) La traduzione come processo decisionale. In Nergaard, S. (a cura di) *Teorie contemporanee della traduzione*, Milano, Strumenti Bompiani, pp. 63-83.

Loffler, L. (1983) Pour une typologie des erreurs dans la traduction automatique. *Multilingua : Journal of Cross Cultural and Interlanguage Communication*, Berlin, pp. 65-78.

Lonsdale, D., Mitamura T., Nyberg, E. (1994) Acquisition of Large Lexicons for Practical Knowledge-Based MT. *Machine Translation*, Dordrecht, Netherlands, 9:3-4.

Macklovitch, E. (1992) Corpus-Based Tools for Tranlators. In *Proceedings of the 33rd Annual Conference of the American Tranlators Association*, San Diego, Californy, novembre 1992, CITI (disp. à l'adresse : http://www-rali.iro.umontreal.ca).

Macklovitch, E. (1993a) Des outils à base de corpus à l'intention des traducteurs., Laval, Canada, octobre 1993, CITI, 13 p. (disp. à l'adresse : http://www-rali.iro.umontreal.ca).

Macklovitch, E. (1993b) Le poste de travail du traducteur (PTT), ou les aides à la traduction. In Bouillon, P., Clas, A. (sous la dir. de), *La traductique : Études et Recherches de traduction par ordinateur*, Montréal, Les Presses de l'Université de Montréal, p. 281-287.

Macklovitch, E. (1994) Using bi-textual alignment for translation validation: the TransCheck system. *Actes du First Conference of the Association for Machine Translation in the Americas (AMTA-94)*, Columbia, 5-8 octobre 1994 (disp. à l'adresse : http://www-rali.iro.umontreal.ca).

Macklovitch, E. (1995a) Can Terminological Consistency be Validated Automatically? *In Proceedings of the IVes Journées scientifiques, Lexicommatique et Dictionnairiques, org. by AUPELF-UREF*, Lyon, 28-30 septembre, 17 p. (disp. à l'adresse : http://www-rali.iro.umontreal.ca).

Macklovitch, E. (1995b) The Future of MT is Now and Bar-Hillel was (almost entirely) Right. In *Proceedings of the Fourth Bar-Ilan Symposium on the Foundations of Artificial Intelligence*, Ramat Gan, Israel, 12 p. (disp. à l'adresse : http://www-rali.iro.umontreal.ca).

Macklovitch, E., Russel, G. (2000) What's been forgotten with translation memory. In *Proceedings of AMTA-2000*, 10 p. (disp. à l'adresse : http://www-rali.iro.umontreal.ca).

Mahimon, M.-D. (1999) *Identification des équivalences traductionnelles sur un corpus Français / Anglais, Mémoire de DEA*, Université de Provence Aix-Marseille 1, Aix-en-Provence.

Manning, C. D., Schütze, H. (1999) *Foundations of Statistical Natural Language Processing, Cambridge*, MU, The MIT Press.

Matsumoto, Y. (1993) Structural Matching of Parallel Texts. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, ACL-94*, pp. 23-30.

Mc Lean, I. J. (1992) Example-based Machine Translation Using Connectionnist Matching. *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, TMI-92*, Montréal, CCRIT.

McEnery, A. M. Oakes, M. P. (1996) Sentence and word alignment in the CRATER project. In Thomas, J., Short, M. (Ed.), *Using Corpora for Language Research*, London, Longman, pp. 211-231.

McEnery, A. M., Nieto-Serrano, A., Smalley, J.P. (1996) *Cognate extraction using approximate string matching techniques*. CRATER project Internal Report.

McEnery, A. M., Oakes, M. P. (1995) Sentence and word alignment in the CRATER project : methods and assessment. In *Proceedings of the EACL-SIGDAT Workshop*, Dublin.

McEnery, A., Langé, J.-M., Oakes, M., Véronis, J. (1997) The exploitation of multilingual annotated corpora for term extraction. In Garsid, R., Leech, G.,Mc Enery, A. M. (Ed.), *Corpus annotation : Linguistics Information from Computer Text Corpora*, London, Addison-Wesley Longman, pp. 220-230.

Malavazos, C., Piperidis, S., Carayannis, G. (2000), Towards memory and template-based translation synthesis, In *Proceedings of MT 2000*, University of Exeter, United Kingdom, 20-22 November 2000

Mel'čuk, I. (1984) *Dictionnaire explicatif et combinatoire du français contemporain*, Montréal, Les Presses de l'Université de Montréal, vol. I.

Mel'čuk, I., Clas, A., Polguere, A. (1995) *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-neuve, Edition Duculot, Champs linguistiques, 256 p.

Melamed, I. D. (1996a) A geometric Approach to Mapping Bitext Correspondence. In *Proceedings of the First Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, pp. 1-12 (disp. à l'adresse : http://www.cis.upenn.edu/~melamed/home.html)

Melamed, I. D. (1996b) Automatic Construction of Clean Broad-Coverage Translation Lexicons. *2nd Conference of the Association for Machine Translation in the Americas*, Montréal, Canada, pp. 125-134 (disp. à l'adresse : http://www.cis.upenn.edu/~melamed/home.html)

Melamed, I. D. (1996c) A Geometric Approach to Mapping Bitext Correspondence. In *Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing*, May 17-18, 1996, pp. 1-10 (disp. à l'adresse : http://www.cis.upenn.edu/~melamed/home.html).

Melamed, I. D. (1996d) Automatic Detection of Omissions in Translations. In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, 5-9 August 1996, pp. 764-769 (disp. à l'adresse : http://www.cis.upenn.edu/~melamed/home.html).

Melamed, I. D. (1997a) A Word-to-Word Model of Translational Equivalence. In *Proceedings of the 35th Conference of the Association for Computational Linguistics*, Madrid, 7-12 July 1997, pp. 490-497.

Melamed, I. D. (1997b) Measuring Semantic Entropy. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics*, Washington, DC, 4-5 April 1997, pp. 41-46 (disp. à l'adresse : http://www.cis.upenn.edu/~melamed/home.html).

Melamed, I. D. (1997c) A Portable Algorithm for Mapping Bitext Correspondence. In *Proceedings of the 35th Conference of the Association for Computational Linguistics*, Madrid, 7-12 July 1997, pp. 305-312 (disp. à l'adresse : http://www.cis.upenn.edu/~melamed/home.html).

Melamed, I. D. (1997d) Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Providence, RI, 1-2 August 1997, pp. 97-108 (disp. à l'adresse : http://www.cis.upenn.edu/~melamed/home.html).

Melamed, I. D. (1997e) A Scalable Architecture for Bilingual Lexicography. *Technical Report # MS-CIS-91-01*, Philadelphia, PA, Dept. of Computer and Information Science, University of Pennsylvania, 6 p. (disp. à l'adresse : http://www.cis.upenn.edu/~melamed/home.html).

Melamed, I. D. (1998a) Word-to-Word Models of Translational Equivalence. *Technical Report # 98-06*, Philadelphia, PA, Institute for Research in Cognitive Science, University of Pennsylvania, 34 p.

Melamed, I. D. (1998b) Models of Co-occurrence. *Technical Report # 98-05*, Philadelphia, PA, Institute for Research in Cognitive Science, University of Pennsylvania, 8 p. (disp. à l'adresse : http://www.cis.upenn.edu/~melamed/home.html).

Melamed, I. D. (1998c) *Empirical Methods for Exploiting Parallel Texts, Ph.D. Dissertation*, Philadelphia, PA, University of Pennsylvania (disp. à l'adresse : http://www.cis. upenn.edu/~melamed/home.html).

Melamed, I. D. (1998d) Manual Annotation of Translational Equivalence: The Blinker Project. *Technical Report # 98-07*, Philadelphia, PA, Institute for Research in Cognitive Science, University of Pennsylvania, 13 p. (disp. à l'adresse : http://www.cis.upenn.edu/~melamed/ home.html).

Melamed, I. D. (1998e) Empirical Methods for MT Lexicon Development. In *Proceedings of AMTA-1998*, 13 p. (disp. à l'adresse : http://www.cis.upenn.edu/~melamed/home.html).

Melby, A. (1993) La typologie des textes: son importance pour la traduction automatique. In Bouillon, P., Clas, A. (sous la dir. de), *La traductique*, Montréal, Les presses de l'université de Montréal, pp. 35-40

Melby, A. (1995) Why Can't a Computer Translate More Like a Person?, Barker Lecture (disp. à l'adresse : http://www.ttt.org/theory/barker.html).

Merkel, M. (1998) Consistency and variation in technical translations - a study of translators' attitudes. In Bowker, L., Cronin, M., Kenny, D., Pearson, J. (Eds.) *Unity in Diversity?, Current Trends in Translation Studies*, St Jerome Publishing, pp. 137-149.

Meyers, A., Yangarber, R., Grishman, R. (1996) Alignment of Shared Forests for Bilingual Corpora. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, Copenhagen, 5-9 August 1996, pp. 460-465.

Minnaja, C.,Paccagnella, L. (2000) A Part-of-Speech Tagger for Esperanto oriented to MT. In *Proceedings of MT 2000*, University of Exeter, Great Britain, 20-22 november 2000, 13-1 :13-5.

Minnis, S. (1993) Constructive Machine Translation Evaluation. *Machine Translation*, Dordrecht, Netherlands, Kluwer Academic Publishers, vol. 8, n. 1-2, pp. 67-75.

Minnis, S. (1994) A Simple and Practical Method for Evaluating Machine Translation Quality. *Machine Translation*, Dordrecht, Netherlands, Kluwer Academic Publishers, vol. 9, pp. 133-149.

Mood, A. M.; Graybill, F. A., Boes, D. C. (1974) *Introduction to the Theory of Statistics*, McGraw Hill.

Mori, S., Nagao, M. (1996) Word Extraction from Corpora Using Distributional Analysis. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, Copenhagen, 5-9 August 1996.

Mounin, G. (1955) *Les Belles Infidèles*, Paris, Cahiers du Sud.

Mounin, G. (1963) *Les problèmes théoriques de la traduction*, Paris, Gallimard.

Mounin, G. (1972) *La sémantique*, Paris, Editions Seghers.

Muller, C. (1968) *Initiation à la statistique linguistique*, Paris, Larousse.

Nagao, M. (1984) A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In A. Elithorn et R. Banerji, Eds, *Artificial and Human Intelligence*, Amsterdam, Elsevier Science Publishers, pp. 173-180.

Nergaard, S. (1993) Introduzione. In Nergaard, S. (a cura di) *La teoria della traduzione nella storia*, Milano, Strumenti Bompiani, pp. 3-49.

Nida, E. (1959, 1995) Principi di traduzione esemplificati dalla traduzione della bibbia. In Nergaard, S. (a cura di) *Teorie contemporanee della traduzione*, Milano, Strumenti Bompiani, pp. 149-180.

Nida, E. (1969) *The theory and practice of translation*, Leiden, Brill.

Niremburg, S., Domashnev, C., Grannes, D. (1993) Two Approaches to Matching in Example-Based Translation. In *Proceedings of TMI-93*, Kyoto, Japan, pp. 47-57.

Nirenburg, S. (1993) L'interlangue et le traitement du sens dans les systèmes de T.A. In Bouillon, P., Clas, A. (sous la dir. de), *La traductique*, Montréal, Les presses de l'université de Montréal, p. 91-108.

Nirenburg, S. et al. (1989) *KBMT-89 project report*. Pittsburg, avril 1989, Center for Machine Translation, Carnegie Mellon University.

Palmer, D. (1995) SATZ, An Adaptative Sentence Segmentation System. *Report n. UCB/CSD-94-846*, Berkeley, Computer Science Division University of California, 29 p.

Pergnier, M. (1993) *Les fondements sociolinguistiques de la traduction*, Lille, Presses Universitaires de Lille.

Piperidis, S. (1995) Translearn : interactive corpus-based translation drafting tool. In *Aslib Proceedings*, Vol. 47, n°3, March 1995, pp. 83-92.

Piperidis, S., Papageorgiou, H., Demiros, I., Malavazos, C., Triantafyllou, I (1998) A Framework for Example-based Translation-Aid Tools. In *Proceedings of the Panhellenic Conference on New Information Technology*-(*NIT'98*), Athens, Greece, 8-10 October 1998, pp. 269-278.

Pottier, B. (1992a) *Sémantique générale*, Paris, PUF.

Pottier, B. (1992b) *Théorie et Analyse en Linguistique*, Paris, Hachette, Collection HU Linguistique.

Rapp, R. (1995) Identifying Word Translations in Non-Parallel Texts. In *Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics*, Boston, Massachussetts, March 1995, pp. 320-322.

Rastier, F. (1987) Microsémantique et textualité. In Charolles M., Petöfi J.S., Sözer E. (Ed.), *Research in Text Connexity and Text Coherence*, Hambourg, Helmut Buske Verlag, pp. 147-166.

Rastier, F. (1988) Problématiques Sémantiques. *Hommage à Bernard Pottier*, Paris, Klincksieck, Tome II, pp. 671-686.

Rastier, F. (1989) *Sens et textualité*, Paris, Hachette, Collection HU Linguistique.

Rastier, F. (1990a) La triade sémiotique, le trivium et la sémantique linguistique. *Nouveaux Actes sémiotiques*, Paris.

Rastier, F. (1990b) Mot, phrase, texte : pour une sémantique unifiée. *Support de Cours*, Paris, LIMSI – CNRS, vol. 1, p. 10.

Rastier, F., Cavazza, M., Abeille, A. (1994) *Sémantique pour l'analyse, de la linguistique à l'informatique*, Paris, Masson, Collection Sciences Cognitives.

Resnik, P., Melamed, D. (1997) Semi-Automatic Acquisition of Domain-Specific Translation Lexicons. In *Proceedings of the 35th Conference of the Association for Computational Linguistics*, Madrid, 31 March - 3 April 1997, 340-347.

Reynar, J. C., Ratnaparkhi, A. (1997) A Maximum Entropy Approach to Identifying Sentence Boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington D.C., pp. 16-19.

Ricoeur P. (1994) L'herméneutique et la méthode des sciences sociales. In Amselek, P. (Ed.), *Théorie du droit et science*, Paris, PUF.

Risset, J. (1985) Traduire Dante. In Dante, *L'enfer*, Paris, Garnier - Flammarion, pp. 15-22.

Rosch, E. et al. (1976) Basic Objects in Natural Categories. *Cognitive Psychology*, 8, pp. 382-436.

Roudaud, B., Puerta, M.-C., Gamrat, O. (1993) A Procedure for the Evaluation and Improvement of an MT System by the End-User. *Machine Translation*, Dordrecht, Netherlands, Kluwer Academic Publishers, 8:1-2, pp. 109-116.

Sager, J. C. (1994) *Language Engineering and Translation : Consequences of automation*, Amsterdam, John Benjamins.

Sammouda, E. (1994) *Algorithmes pour l'appariement des mots ete des phrases de textes bilingues Français-Anglais*, Thèse, Université de Paris X.

Santos, D. (2000) The translation network. In Véronis, J. (Ed.), *Parallel Text Processing*, Dordrecht, Netherlands, Kluwer Academic Publishers, § 8, p. 18.

Sato, S. (1992) CTM : An Example-based Translation Aid System. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING-92*, Nantes, pp. 1259-1263.

Sato, S. (1995) MBT2 : A Method for Combining Fragments of Examples in Example-Based Machine Translation. *Artificial Intelligence*, n° 75, pp. 31-49.

Sato, S., Nagao, M. (1990) Toward Memory-based Translation. In *Proceedings of the 13th International Conference on Computational Linguistics, COLING-90*, Helsinky, pp. 247-252.

Satoshi, S., Jun-Ich, T. (1995) Automatic Acquisition of Semantic Collocation from Corpora. *Machine Translation*, Dordrecht, Netherlands, Kluwer Academic Publishers, vol. 10, n. 3, pp. 219-258.

Seleskovitch, D. (1975a) *Langage, langues et mémoire, étude de la prise de notes en interprétation consécutive*, Paris, Minard, Lettres Modernes, 272 p.

Seleskovitch, D. (1975b) Pour une théorie de la traduction inspirée de sa pratique. *META*, Outremont, PQ, vol. 25, n° 4, dec. 1980, pp. 401-408.

Seleskovitch, D. (1984) En collaboration avec Lederer, M., *Interpréter pour Traduire*, Paris, Didier érudition, 312 p.

Senellart, J., Fairon, C. (1999) Classes d'expression bilingues gérées par des transducteurs finis (FST), dates et titres de personnalité (anglais-français) *Actes du colloque Linguistique contrastive et Traduction Approches Empiriques*, Louvain-la-Neuve, 5-6 février 1999, Université Catholique de Louvain, pp. 37-38.

Shannon, C. E. (1949) *The Mathematical Theory of Communication.*, Urbana, University of Illinois Press.

Shin, J. H., Choi, K.-S., Han, Y. S. (1996) Bilingual Knowledge Acquisition From Korean-English Parallel Corpus Using Alignment Method: Korean-English Alignment At

Word And Phrase Level. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, Copenhagen, 5-9 August 1996.

Shiwen, Y. (1993) Automatic Evaluation of Output Quality for Machine Translation Systems. *Machine Translation*, Dordrecht, Netherlands, Kluwer Academic Publishers, pp. 117-126.

Simard, M. (1998) The BAF : A Corpus of English-French Bitext. *First International Conference on Language Resources and Evaluation*, Granada, Espagne, pp. 489-494.

Simard, M. (2000) Multilingual text alignment – Aligning three or more versions of a text. In Véronis, J. (Ed.), *Parallel Text Processing*, Dordrecht, Netherlands, Kluwer Academic Publishers, § 3, 20 p.

Simard, M., Foster, G., Isabelle, P. (1992) Using Cognates to Align Sentences in Bilingual Corpora. *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, TMI-92*, Montréal, CCRIT, pp. 67-81.

Simard, M., Foster, G., Perrault, F. (1993) TransSearch : un concordancier bilingue. *CITI Technical Report*, Laval, Canada, 20 p. (disp. à l'adresse : http://www-rali.iro.umontreal.ca).

Simard, M., Plamondon, P. (1996) Bilingual Sentence Alignment : balancing robustness and accuracy. In *Proceedings of AMTA-96*, Montréal, Canada, pp. 135-144.

Sinclair J. *et al*., eds. (1987) *Collins Cobuild English Language Dictionary*, London, Collins.

Smajda, F. (1993) Retrieving Collocations from Text : Xtract. *Computational Linguistics*, vol. 19, n.1, pp. 143-177.

Smajda, F., Mc Keown, K., Hatzivassiloglou, V. (1996) Translating Collocations for Bilingual Lexicons : A Statistical Approach. *Computational Linguistics*, pp. 1-38.

Somers, H. L. (1993a) La traduction automatique basée sur l'exemple ou sur les corpus. In Bouillon, P., Clas, A. (sous la dir. de), *La traductique*, Montréal, Les presses de l'université de Montréal, pp. 149-166.

Somers, H. L. (1993b) Current Research in Machine Translation. *Machine Translation*, Dordrecht, Netherlands, Kluwer Academic Publishers, vol. 7, pp. 231-246.

Somers, H. L. (1998) Further Experiments in Bilingual Text Alignment. *International Journal of Corpus Linguistics*, Vol. 3, n°1, 36 p.

Sowa, J. F. (1984) *Conceptual Structures: Information Processing in Mind and Machine*, London, Addison-Wesley.

Sta, J.-D. (1995) Comportement statistique des termes et acquisition terminologique à partir de corpus. *T.A.L.*, vol. 36, n. 1-2, pp. 119-132.

Su, K.-Y., Wu, M.-W., Chang, J.-S. (1994) A corpus-based Approach to Automatic Compound Extraction. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, ACL-94*, New Mexico State University, June 1994, pp. 242-247.

Sumita, E., Iida, H. (1991) Experiments and Prospects of Example-based Machine Translation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, ACL-91*, Morristown, NJ, pp. 185-192.

Sumita, E., Tsutsumi, Y. (1988) A translation Aid System Using Flexible Text Retrieval Based on Syntax-matching. *Second International Conference on Theoretical and Methological Issues in Machine Translation of Natural Languages, TMI-88*, Pittsburg.

Tanaka, K., Iwasaki, H. (1996) Extraction of Lexical Translations from Non-Aligned Corpora. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, Copenhagen, 5-9 August 1996, pp. 580-585.

Tesnière, L. (1959) *Eléments de syntaxe structurale*, Paris, Klincksieck.

Toury, G. (1980, 1995) Comunicazione e traduzione. Un approccio semiotico. In Nergaard, S. (a cura di) *Teorie contemporanee della traduzione*, Milano, Strumenti Bompiani, pp. 103-119.

Utsuro, T., Matsumoto Y., Nagao, M. (1992) Lexical Knowledge Acquisition from Bilingual Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING-92*, Nantes, pp. 581-587 (disp. à l'adresse : http://www.kuee.kyoto-u.ac.jp).

Van Der Eijk, P. (1997) Comparative Discourse Analysis of Parallel Texts. In Armstrong et al. (Ed.), *NLP using Very Large Corpora*, Dordrecht, Netherlands, Kluwer Academic Publishers, pp. 219-233.

Van Slype, G. (1982) Conception d'une méthodologie générale d'évaluation de la traduction automatique. *Multilingua : Journal of Cross Cultural and Interlanguage Communication*, Berlin, &:4, pp. 221-237.

Vandooren, F. (1993) Divergences de traduction et architectures de transfert. In Bouillon, P., Clas, A. (sous la dir. de), *La traductique*, Montréal, Les presses de l'université de Montréal, pp. 77-90.

Vauquois, B. (1992) La Traduction Automatique. In Pottier, B. (Ed.), *Les Sciences du langage en France au XXème siècle*, Paris, Peeters.

Véronis, J. (1997) Une action d'évaluation des systèmes d'alignement de textes multilingues. *1ères JST 1997 FRANCIL de l'AUPELF-UREF*, Avignon, 15-16 avril 1997, pp. 191-197.

Véronis, J. (1998) A study of polysemy judgements and inter-annotator agreement. In Véronis, J. (Ed.), *Program and advanced papers of the Senseval workshop*, Herstmonceux Castle, Great Britain, 2-4 September 1998.

Véronis, J. (2000) From the Rosetta Stone to the information society : A survey of parallel text processing. In Véronis, J. (Ed.), *Parallel Text Processing*, Dordrecht, Netherlands, Kluwer Academic Publishers, § 1, 24 p.

Véronis, J. (2000) Sense Tagging : Don't look for the meaning but for the use. In *Proceedings of Comlex 2000*, Patras, Greece, University of Patras, pp. 1-9.

Véronis, J., Langlais, P. (2000) Evaluation of parallel text alignment systems – The ARCADE project. In Véronis, J. (Ed.), *Parallel Text Processing*, Dordrecht, Netherlands, Kluwer Academic Publishers, § 19, 20 p.

Victorri, B., Fuchs, C. (1996) *La polysémie, construction dynamique du sens*, Paris, Hermès.

Vinay, J.-P, Darbelnet, J. (1958) *Stylistique comparée du français et de l'anglais*, Paris, Didier.

Vogel, S., Ney, H., Tillman, C. (1996) HMM-Based Word Alignment in Statistical Translation. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, Copenhagen, 5-9 August 1996, pp. 836-841.

Volle, M. (1997) *Analyse des données*, Paris, Economica.

Walter, H. (1997) *L'aventure des mots français venus d'ailleurs*, Paris, Laffont.

Watanabe, H. (1992) A Similarity-driven Transfer System. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING-92*, Nantes, 23-28 août 1992, pp. 770-776.

Wehrli, E. (1991) Pour une approche interactive au problème de la traduction automatique. In Clas, A., Safar, H. (sous la dir. de), *Actes du colloque de Mons, L'environnement traductionnel ; La station de travail du traducteur de l'an 2001*, Mons, Belgique, 25-27 avril 1991, pp. 59-68.

Wehrli, E. (1993) Vers un système de traduction interactif. In Bouillon, P., Clas, A. (sous la dir. de), *La traductique*, Montréal, Les presses de l'université de Montréal, pp. 423-432.

Weischedel, R. (1993) Coping with Ambiguity and Unknown Words through Probabilistic Models. *Computational Linguistics*, vol. 19, n. 2, pp. 359-382.

Whorf, B. L. (1958) *Language, thougth and reality*, New-York, Wiley and sons.

Wittgenstein, L. (1961) *Tractatus Logico-Philosophicus, suivi de investigations philosophiques*, Paris, Gallimard.

Wu, D. (1994) Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, ACL-94*, Las Cruses, New Mexico, New Mexico State University, June 1994, pp. 80-87.

Wu, D. (1995) Grammarless Extraction of Phrasal Translation Examples from Parallel Texts. In *Proceedings of the Sixth International Conference on Theoretical and Methological Issues in Machine Translation of Natural Languages,TMI-95*, Leuven, Belgique, pp. 354-372 (disp. à l'adresse : http://www.cs.ust.hk).

Wu, D. (1996) Stochastic Inversion Transduction Grammars, with application to Segmentation, Bracketing, and Alignment of Parallel Corpora. In *Proceedings of IJCAI 96*, Montréal, 20-25 août 1996, vol. 2, pp. 1328-1335 (disp. à l'adresse : http://www.cs.ust.hk)

Wu, D., Xia, X. (1994) Learning an english-chinese lexicon from a parallel corpus. In *Proceedings of the Association for Machine Translation in the Americas, AMTA-94*, Columbia, MD, octobre 1994, pp. 206-213.

Wu, D., Xia, X. (1995) Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon. *Machine Translation*, Dordrecht, Netherlands, Kluwer Academic Publishers, 9(3-4), pp. 285-313.

Yang, Y., Brown, Ralf D., Frederking, R. E., Carbonell, J. G., Geng, Y., Lee, D. (1997) Bilingual-corpus based approaches to translingual information retrieval. *2nd Workshop on Multilinguality in Software Industry: The AI Contribution (MULSAIC'97)* Nagoya, Japan, August 25, 1997.

Zavrel, J., Daelemans, W. (1997) Memory-based learning: Using similarity for smoothing. In *Proceedings of the Annual Meeting of the ACL and Conf. of the European Chapter of the ACL 1997*, pp. 436-443.

Zinglé, H. (1993) Evolution de la Traduction Automatique de 1947 à nos jours. *Actes du 3ème Colloque Histoire de l'Informatique*, Sophia Antipolis, 13-15 octobre 1993, pp. 398-414.

Zinglé, H. (1993) The Zstation Development and the modelling of linguistic knowledge. In *International Conference on Mathematical Linguistics, Report*, Tarragona, Spain, 30-31 march 1993, pp. 105-106.

Zinglé, H. (1996) ZART : un linguiciel d'aide à la rédaction scientifique et technique en langue étrangère, in *Travaux du LILLA*, Presses universitaires de Nice, n° 1, pp 111-113.

Zipf, G. K. (1935) *The Psychobiology of Language, an Introduction to Dynamic Philology*, Boston, Houghton-Mifflin.