Projet ANR-DFG PhraséoRom

Manuel d'annotation stylistique des motifs (corpus français)

Sommaire

Laetitia Gonon (Université Grenoble Alpes), Marion Gymnich (Universität Bonn), Ilaria Vidotto (Université de Lausanne)

1. Description du fichier d'annotation

Le fichier élaboré pour l'annotation stylistique a été constitué pour stocker les données relatives à l'analyse stylistique des motifs en vue de l'élaboration de la base de données « PhraseoBase » du projet Phraséorom. Dans son état actuel, il se présente au format excel et compte 13 colonnes et 106 lignes.

1.1. Structuration et dénomination des colonnes

Les colonnes du fichier excel ont été construites dans le but d'inclure et rendre accessibles aux usagers de la base de données les informations pertinentes pour l'analyse stylistique des motifs phraséologiques. Chaque colonne détaille le type d'information annoté ; leur remplissage n'est pas obligatoire : on n'annote que les informations utiles pour l'analyse du motif. Voici le contenu des colonnes dans le détail :

1) Id groupe : identifiant attribué à chaque motif, lequel agrège plusieurs ALR similaires, spécifiques à un ou plusieurs sous-genres ;

2) Motif : étiquette attribuée au motif, consistant en ses « mots-vedettes » (pivot nominal et verbal), extraits de l'ALR « leader » (cf. *infra*, § 2.1) ;

3) Spécificité : le ou les sous-genres au(x)quel(s) le motif est spécifique ;

4) Syntaxe cœur : composantes syntaxiques du motif, soit le motif dans sa configuration syntaxique minimale ;

5) Position : critère transphrastique, susceptible de contribuer à l'identification de la fonction discursive du motif. On tient compte de la place que le motif occupe dans l'enchaînement textuel (en début/en fin de chapitre ou de paragraphe, dans l'entourage du discours direct, etc.)

6) Distribution : critère intraphrastique ; on observe le contexte d'apparition du motif au niveau de la phrase : enchaînement parataxique (coordination/juxtaposition) ou hypotaxique (dans une proposition subordonnée), phrase indépendante, construction au gérondif, complément infinitif, phrase interrogative, syntagme prépositionnel, etc.

7) Composante facultative : extensions syntagmatiques du motif, soit tous les constituants qui ne sont pas inclus dans la syntaxe cœur mais qui peuvent être significatifs pour la définition de la fonction discursive (adjectifs, adverbes, groupes prépositionnels en fonction de compléments circonstanciels, gérondif etc.) ;

8) Exemple : transcription de plusieurs exemples extraits du Lexicoscope (cf. *infra*, § 2.5) illustrant la configuration syntaxique ainsi que la distribution du motif, et rendant compte de sa ou ses fonctions discursives ;

9) Identifiant de phrase : identifiant de l'occurrence citée tel qu'il s'affiche dans le Lexicoscope ; chaque identifiant se compose de la lettre *s* (« sentence »), suivi d'un chiffre ;

10) Référence : métadonnées de l'exemple extraites du Lexicoscope, soit le nom de l'auteur, suivi de l'initiale du prénom, le titre de l'œuvre et l'année de publication ;

11) Étiquette stylistique : désignation de la/des fonction(s) discursive(s) du motif ;

12) Interprétation stylistique : analyse plus détaillée des FD du motif et du rôle de ce dernier dans l'économie du texte/du sous-genre ;

13) Commentaires : colonne libre à l'usage des annotateurs ; elle ne figure pas dans la base de données.

1.2 Nombre et structuration des lignes

Le fichier compte actuellement 106 lignes. Chaque ligne contient les informations catégorisées par les colonnes, mais leur nombre varie en fonction de la richesse du motif, et ce aussi bien en termes de configuration syntaxique (variations syntagmatiques, distributions, composantes facultatives) que des FD. Afin d'éviter les redondances, l'identification des fonctions discursives est prioritaire dans le choix de dédoubler ou d'unifier les lignes. Cela signifie que si un motif présente la même configuration mais des fonctions discursives différentes, on choisit d'ajouter une ligne, alors que les données mettant en évidence des différences au niveau des paramètres de la position, de la distribution ou de la composante facultative, mais n'ayant pas de répercussions sur la définition des FD seront rassemblées à l'intérieur d'une même ligne.

1.3 Format de saisie des valeurs

1.3.1 Motif

Le nom du motif est saisi de la façon suivante : *Constituant$_1$_Constituant$_2$_Constituant$_3$_Constituant$_n$*, par exemple : *apparaître_sur_écran*

1.3.2 Spécificité

Le ou les sous-genre(s) au(x)quel(s) chaque motif annoté est spécifique sont indiqués en recourant aux abréviations utilisées dans le cadre du projet PhraseoRom :
GEN (littérature générale), SENT (romans sentimentaux), POL (romans policiers), HIST (romans historiques), SF (romans de science-fiction), FY (romans de fantasy)

1.3.3 Syntaxe cœur

Les abréviations utilisées pour noter les constituants syntaxiques du motif et/ou leur catégorie grammaticale sont celles en vigueur dans la communauté des linguistes et des stylisticiens, par exemple : SN (syntagme nominal), V (verbe), PRO (pronom), SNPrep (syntagme prépositionnel), DET (déterminant), etc.

1.3.4 Exemples

Dans la colonne « exemples », les extraits sont transcrits uniquement sous forme de texte, sans faire référence aux métadonnées, qui sont indiquées dans une autre colonne. Afin de rendre compte de la variation paradigmatique au niveau des constituants de chaque motif, on cite plusieurs exemples (entre trois et six, en moyenne) ; chaque exemple est séparé du suivant par trois @@@.
Exemple : *Le ronronnement d'un mécanisme se fit entendre et un être monstrueux apparut sur l'écran.\n @@@Une phrase s'inscrivit peu à peu sur l'écran de l'ordinateur.\n Simultanément, un synthétiseur vocal la prononça.*

1.3.5 Identifiant de phrase

On retranscrit l'identifiant de phrase fourni par le Lexicoscope ; chaque extrait possède un identifiant composé de la lettre *s* (« sentence ») et d'un chiffre. Par exemple : s56597
De même que pour les citations, chaque identifiant est séparé du suivant par trois @@@ ; l'ordre de transcription des identifiants suit celui des exemples textuels.

1.3.6 Métadonnées

Les métadonnées sont extraites du Lexicoscope. Leur transcription se fait dans le format suivant : *Nom de l'auteur, Initiale du prénom, Titre de l'ouvrage, date de publication*, en petites capitales. Exemple : Werber B., 2 La trilogie des fourmis 2 Le jour des fourmis, 1992.
Chaque groupe de métadonnées est séparé du suivant par trois @@@. Si le motif est commun à deux sous-genres (ou plus), on précise le sous-genre avant le nom de l'auteur, en lettres capitales.

1.3.7 Fonctions discursives (FD)

Étiquettes utilisées pour noter les fonctions discursives :

- Fonction narrative : *narratif*
- Fonction infranarrative : *infranarratif*
- Fonction descriptive : *descriptif*
- Fonction infradescriptive : *infradescriptif*
- Fonction indirectement descriptive : *indirectement_descriptif*
- Fonction affective : *affectif*
- Fonction cognitive : *cognitif*
- Fonction cognitive commentative : *cognitif:commentatif*
- Fonction cognitive mémorielle : *cognitif:mémoriel*
- Fonction pragmatique : *pragmatique*
- Fonction interactive : *interactif*

Lorsqu'une même occurrence de motif présente deux fonctions discursives, on sépare les deux étiquettes stylistiques par un + (sans espaces). Si en revanche un motif assume, en fonction des contextes, soit telle FD, soit telle autre, l'alternative se marque par un |entre les deux étiquettes.

1.3.8 Symboles utilisés pour annoter plusieurs informations dans la même cellule

+ → équivaut à « et », marque la concomitance de deux informations différentes.
| → équivaut à « ou », marque l'alternative (deux informations présentes dans des contextes différents et non concomitants).
: → indique la sous-catégorisation ; par exemple, la FD commentative est une sous-catégorie de la FD cognitive, d'où son étiquette *cognitif:commentatif*.
@@@ → séparateurs.
_ (underscore) → remplace l'espace typographique dans les étiquettes composées de plus d'un terme.

1.4 Fonctions discursives

Les **FD narrative** et **descriptive** sont les plus attendues dans le roman. La fonction narrative contribue à la structuration du texte en reliant une suite d'actions, d'événements, de mots et de pensées, alors que la fonction descriptive introduit et supporte un passage descriptif.
Exemples :

- *Le conducteur consulta sa montre : 8h15* (rôle dans l'action)
- *Il regarda de nouveau par la fenêtre. Des couleurs de cuisine, voilà ce qu'étaient les couleurs de l'Italie […]* + séquence descriptive (description)
- *Hervé fronça les sourcils, intrigué* (description des affects)

Dans le cas de la description d'affects, sentiments, émotions, on propose une catégorie spécifique de **FD : affective**. Exemples :

- *Sarah écrasa nerveusement sa cigarette.*

Dans les études pilotes, la FD descriptive est en réalité souvent affective

La **FD indirectement descriptive** rend compte d'une action répétée, d'un geste qui caractérisent le personnage (gros fumeur [POL] ; à la sensibilité accrue [FY])

La **FD « infra-narrative »** est propre aux motifs qui demeurent à l'arrière-plan de l'action. Ceux-ci contribuent à meubler la conversation ou forment un enchaînement d'actions menues, dans un script sans conséquence narrative pour l'action principale. Exemple :

- *-Tu feras mieux la prochaine fois, assure Alexandre en allumant une cigarette.*

La **FD « infra-descriptive »** est assumée par les motifs qui introduisent une précision descriptive minimale, souvent stéréotypée. Exemple :

- *Maintenant ils se taisaient, regardant par la fenêtre les reflets d'un ciel sinistre dans les eaux de la lagune.*

La **FD cognitive** s'emploie pour les motifs impliquant des processus cognitifs (hypothèses, appréhension des événements, réflexions…). Exemple :

- *Je ressentais une joie que je ne pouvais ressentir – dont j'avais l'impression de ne plus posséder les instruments pour la ressentir.*

On propose également deux sous-catégories pour cette FD : une FD cognitive **mémorielle**, pour l'expression du souvenir et une FD cognitive **commentative**. Elle traduit une réflexivité (*reflective writing*) d'un personnage sur l'activité même d'écriture, ou bien une activité de commentaire, de réflexion ou l'expression d'un jugement.

La **FD pragmatique** concerne les motifs qui expriment des actes de langage (*speech acts*) entre les personnages du roman (au discours direct essentiellement). Lorsqu'il fait état de cette fonction, le motif permet la cohérence des rapports entre les personnages, au sein du discours rapporté intégré au texte narratif. Exemple :

- *-N'en faites rien, Madame, je vous en prie, s'écria Eudeline.*

On parle enfin de **FD interactive** lorsque le motif déclenche une interaction et introduit une séquence dialogale. Exemple :

- *Un Mokranien casqué apparut sur l'écran.\n – Pourquoi nous avoir stoppés ? s'indigna-t-il.*

2. Étapes de l'annotation

2.1 Repérage des ALR « leader » à partir du fichier de comparaison des 6 sous-genres

La première étape de l'annotation stylistique consiste à repérer des ALR (arbres lexico-syntaxiques récurrents) susceptibles de former un même motif phraséologique. Ces ALR, annotés syntaxiquement et sémantiquement, ont été extraits automatiquement du corpus PhraséoRom et réunis dans un fichier de comparaison où figurent tous les six sous-genres étudiés. La sélection des ALR s'est faite selon un double critère : un critère quantitatif et un critère de signifiance minimale : pour chaque sous-genre, on a sélectionné les ALR présentant un seuil de spécificité très élevé et dont la complétude sémantico-syntaxique – résultant de la présence d'au moins un pivot verbal et un pivot nominal – était suffisante pour les constituer en unités signifiantes. Cela signifie que, par exemple, l'ALR <mis au point>, hautement spécifique dans la SF (son seuil de spécificité, calculé en appliquant la méthode du *loglikelihood ratio*, atteint 207, 53, alors que le seuil minimal a été fixé à 10,83) n'a pas été retenu pour autant pour l'annotation stylistique car incomplet et dépourvu de pivot nominal. En raison de leur spécificité, les ALR retenus ont été appelé « leaders », dans la mesure où c'est à partir de leurs composantes syntaxiques de base qu'on identifie ensuite d'autres ALR similaires, présentant des variations syntagmatiques et/ou paradigmatiques et pouvant constituer, avec l'ALR « leader », un motif phraséologique.

2.2 Tableau des ALR spécifiques à chaque sous-genre retenus pour l'annotation

GEN	log	SENT	log	POL	log	HIST	log	SF	log	FY	log
<fumer une cigarette>	229,06	<elle prit dans ses bras>	154,78	<découvert le corps>	159,45	<dit avec un sourire>	299,03	<voyager dans le temps>	129,12	<dégaine son épée>	344,13
<j'eu le sentiment>	99,82	<éclata de rire>	138,77	<prévenir la police>	153,07	<baiser les mains>	135,96	<apparut sur les écrans>	104,33	<reporta son attention>	253
<le soleil brûle>	95,47	<elle posa sa main>	110,68	<ouvrait le coffre>	134,13	<font la guerre>	124,31	<l'écran montre>	99,25	<poussa un cri>	238,29
<regardait la mer>	92,36	<il lui adressa un sourire>	85,98	<se passa la main>	89,73	<met le siège>	117,86	<ouvrir le sas>	94,79	<encoché une flèche>	218,07
<écrit un roman>	89,44	<elle prit par main>	23,97	<allumer une nouvelle cigarette>	19,28	<il dit d'une voix>	105,17	<lancé un coup d'œil>	80,12	<arrachant un cri>	61,21

2.3 Vérification des regroupements automatiques

Après la sélection des ALR « leader », on procède à la vérification des regroupements automatiques dans le fichier de comparaison des six sous-genres. Cette opération est rendue nécessaire par le fait que le seuil fixé pour grouper les ALR les plus similaires en les comparant les uns aux autres ne permet pas de repérer l'intégralité des occurrences, car il ne prend pas en compte tous les critères qu'on adopterait au cours d'un regroupement manuel. En effet, des ALR similaires, susceptibles de former un même motif, figurent parfois dans des groupes différents, alors que des ALR susceptibles de former des motifs différents sont regroupés sous le même identifiant. Ainsi faisant, l'annotateur repère en outre plus rapidement les variations paradigmatiques et/ou syntagmatiques pertinentes pour l'annotation du motif phraséologique.
Exemple :

ID groupe : 1226	**ID groupe : 923**	**ID groupe : 5497**
elle prit dans ses bras (log 154, 78) ***elle le prit bras*** (log 101,82) ***il me prit bras*** (log 32,73) ***la prendre par la main*** (log 21,54) ***elle le prend par la main*** (log 17,78) ***il la prit par la main*** (log 17, 78)	***elle prit bras*** (log 149,42) ***il prise main*** (log 48, 14) ***elle prit par main*** (log 23, 97) ***elle prit par la main*** (log 23,45) *elle prend ses mains* (log 23,45) ***il les prit main*** (log 17,35) ***elle prend entre mains*** (log 17,21) ***il prend visage mains*** (log 16,58) ***elle prit visage entre ses mains*** (log 15,28) ***il prit entre ses mains*** (log 15,04) *il prit le bras* (log 13,95)	***prend par la taille*** (log 109,76) ***la prendre par la taille*** (log 84,57) ***prend par la taille*** (log 82,66)

2.4 Collecte des exemples dans le Lexicoscope

Une fois l'ALR « leader » et les ALR similaires identifiés, l'annotateur renseigne l'expression régulière qui caractérise l'ALR dans le Lexicoscope, afin d'observer le contexte dans lequel le motif – formé par l'agrégation de plusieurs ALR – est susceptible d'apparaître. Le passage en revue des extraits de texte fournis par le logiciel constitue une étape essentielle pour la détermination des FD associées au motif et, plus largement, pour l'interprétation stylistique du rôle que le segment phraséologique joue aussi bien dans les textes où il figure que dans le sous-genre dont il est spécifique.
Le tableau ci-dessous récapitule le nombre total d'occurrences des motifs annotés dans chaque sous-genre proposées par le Lexicoscope :

GEN	**Nb. Occ.**	**SENT**	**Nb. Occ .**	**POL**	**Nb. Occ.**	**HIST**	**Nb. Occ.**	**SF**	**Nb. Occ.**	**FY**	**Nb. Occ**
fumer_ cigarette	364	prendre_ dans_bras	344	découvrir_ corps	123	dire_avec_ sourire	229	voyager_ dans_ temps	47	dégainer_ épée	108
avoir_ sentiment	636	éclater_rire	347	prévenir_ police	100	baiser_ main	146	apparaître_ sur_écran	69	reporter_ attention	140

soleil_ brûler	99	poser_ main	314	ouvrir_ coffre	89	faire_ guerre	188	écran_ montrer	35	pousser_ cri	588
regarder_ mer	141	adresser_ sourire	157	passer_ main_sur_ visage	58	Mettre_ siège	40	ouvrir_sas	25	encocher_ flèche	65
écrire_ roman	123	prendre_ par_main	96	allumer_ cigarette	331	dire_voix	376	lancer_ coup_d'œil	86	arracher_ cri	55

2.5 Saisie des données et des analyses dans le fichier de l'annotation

Les informations recueillies à partir de la lecture et de l'observation des exemples proposés par le Lexicoscope sont renseignées dans les différentes colonnes du fichier de l'annotation, suivant le protocole de saisie indiqué plus haut. Elles amènent l'annotateur à définir la ou les fonctions discursives du motif et à livrer un commentaire stylistique plus détaillé à propos du fonctionnement du motif en contexte, ainsi que de sa contribution à la structuration narrative.