The ANR-DFG PhraseoRom Project

Manual for the stylistic annotation of motifs (French and English corpus)

Summary

1. Description of the annotation file

The file that has been developed for the stylistic annotation has the purpose of recording the data related to the stylistic analysis of motifs and thus of contributing to the elaboration of the database “PhraseoBase” within the PhraseoRom project. In its current state, it appears in excel format.

1.1. Structure and denomination of the columns

The columns of the excel file have been designed to include information that is relevant to the stylistic analysis of phraseological motifs and to make this information accessible to the users of the database. Each column designates the type of information that is annotated; the columns do not all have to be filled in in every single case: only the information that is actually useful for the analysis of the motif has been provided. The individual columns offer the following information:

1) Group Id: the identifier that is assigned to each motif, which typically combines several, similar RLTs (= recurring lexico-syntactic trees), which are specific to one or more subgenres;

2) Motif: the label that is assigned to the motif, which consists of its “headwords” (pivot nominal and verbal pivot), extracted from the respective “leader” RLT (cf. below, § 2.1);

3) Specificity: the subgenre(s) to which the motif is specific;

4) Core syntax: the syntactic components of the motif, i.e., the motif in its minimal syntactic configuration;

5) Position: a criterion that takes the position of the RLT beyond the individual sentence into consideration; this proves to be useful for identifying the discursive function(s) of a motif. We pay attention to the position the motif occupies within the larger textual structures (e.g. occurring at the beginning/at the end of a chapter or paragraph, in the environment of direct speech, etc.);

6) Distribution: a criterion referring to the sentence level; we examine the context in which the motif appears on the sentence level: in paratactic (coordination/juxtaposition) or hypotactic (in a subordinate clause) constructions, in an independent sentence, in gerund constructions, as infinitive complement, as interrogative sentence, in a prepositional phrase, etc.;

7) Optional component: syntagmatic extensions of the motif, i.e., all of the constituents that are not included in the core syntax but that can be important for the definition of the discursive function (adjectives, adverbs, prepositional groups functioning as adverbials, gerund, etc.);

8) Example: transcription of several examples extracted from Lexicoscope (see below, § 1.3.4) illustrating the syntactic configuration as well as the distribution of the motif and its discursive function(s);

9) Sentence identifier: identifier of the quoted occurrence(s) as they have been assigned by Lexicoscope; each identifier consists of the letter *s* (= “sentence”), followed by a number;

10) Reference: metadata of the examples extracted from Lexicoscope, i.e., the name of the author, followed by the initial(s) of his/her first name, the title of the work, and the year of publication;

11) Stylistic label: the discursive function(s) of the motif;

12) Stylistic interpretation: a more detailed analysis of the discursive function of the motif and of its role in the context of the text/subgenre;

13) Comments: a column that can be used for additional comments by the annotators; this column does not appear in the database.

1.2 Structure of the lines

Each line in the file contains the information demanded by the columns, but the number of the lines varies, depending on how complex a motif is, with respect to both the syntactic configuration (syntagmatic variation, distributions, optional constituents) and the discursive function(s). The identification of discursive functions has priority when it comes to the question of whether lines should be split or combined in order to avoid redundancy. This means that we chose to add a line if a motif appears in the same configuration, but has different discursive functions, while data that highlights differences with respect to the parameters of the position, the distribution, or the optional component but has no repercussions on the definition of the discursive function(s) will be collected within one and the same line.

1.3 Format of recording the data

1.3.1 Motif

The designation of the motif is represented in the following manner: *Constituent$_1$_Constituent$_2$_Constituent$_3$_Constituent$_n$*, for instance: *apparaître_sur_écran*, *appear_on_screen*

1.3.2 Specificity

The subgenre(s) to which each annotated motif is specific are indicated by means of the abbreviations used in the framework of the PhraseoRom project:
GEN (general fiction), SENT (romance; the abbreviation is derived from the French term 'romans sentimentaux'), POL (crime fiction; the abbreviation is derived from the French term romans policiers), HIST (historical novels), SF (science fiction), FY (fantasy)

1.3.3 Syntactic core

The abbreviations used to indicate the syntactic constituents of the motif and/or their grammatical category are those that are generally used by linguists and stylisticians, e.g.: SN (nominal syntagma), V (verb), PRO (pronoun), SNPrep (prepositional syntagma), DET (determiner), etc.

1.3.4 Examples

In the column "examples", the extractions are transcribed only as texts, without referring to the metadata, which is provided in another column. To pay tribute to the paradigmatic variation with regard to the constituents of each motif, several examples are provided (on average in between three and six); each example is separated from the following one by three @@@.

Examples: *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.*
The room was quiet for almost a minute as images of Mars at different resolutions appeared on the giant screen on the wall.@@@The four biots appeared on the screen and the video slowed.

1.3.5 Sentence identifier

The sentence identifier provided by Lexicoscope is used; each extract has an identifier consisting of the letter *s* ("sentence") and of a number. For example: s56597.
Just as the quotations, the identifiers are separated from one another by means of three @@@; the order of the identifiers corresponds to that of the examples from the literary texts.

1.3.6 Metadata

The metadata have been extracted from Lexicoscope. They are provided in the following format: *name of the author, initials of the author's first name, title of the novel, year of publication*. Examples:
Werber B., 2 La trilogie des fourmis 2 Le jour des fourmis, 1992.
Gibson G., Stealing Light, 2007.

Each group of metadata is separated from the following one by three @@@. If the motif is shared by two subgenres (or more), the subgenre is indicated in front of the name of the author, in capital letters.

1.3.7 Discursive functions (DF)

Labels used for classifying the discursive functions:

- Narrative function: *narrative*
- Infranarrative function: *infranarrative*
- Descriptive function: *descriptive*
- Infradescriptive function: *infradescriptive*
- Indirectly descriptive function: *indirectly_descriptive*
- Affective function: *affective*
- Cognitive function: *cognitive*
- Cognitive-commentative function: *cognitive:commentative*
- Cognitive mnemonic function: *cognitive:mnemonic*
- Pragmatic function: *pragmatic*
- Interactive function: *interactive*

When the same occurrence of a motif can be associated with two discursive functions, the two stylistic labels are separated by a + (without spaces). If, by contrast, a motif is associated with either one DF or another one (depending on the context), the alternatives are indicated by placing a | between the two labels.

1.3.8 Symbols used to annotate several pieces of information within the same cell

+ → meaning “and” marks the concomitance of two different pieces of information.
| → meaning “or” marks the existence of alternatives (two pieces of information present in different contexts and not concomitant).
: → indicates sub-categorisation; for instance, the commentative DF is a sub-category of the cognitive DF, which is reflected in its label *cognitive:commentative*.
@@@ → separators
_ (underscore) → replaces the spaces within labels consisting of more than one term

1.4 Discursive functions

The **narrative** and **descriptive DF**s are expected to occur most frequently in novels. The **narrative function** contributes to the structuring of the text by connecting a sequence of actions, events, words, and thoughts, while the **descriptive function** introduces and supports descriptive passages.
Examples:

- *Le conducteur consulta sa montre: 8h15* (role within the action)
- *I sat down and he slammed the door smartly and got behind the wheel.*
- *Il regarda de nouveau par la fenêtre. Des couleurs de cuisine, voilà ce qu'étaient les couleurs de l'Italie […]* + descriptive sequence (description)
- *They rounded the corner and wandered past the bookshop; Edith, half-heartedly, hung back to look in the window, where Le Soleil de Minuit, in its paper cover, made a modest appearance.* + descriptive sequence (description)
- *Hervé fronça les sourcils, intrigué* (description of emotions)
- ‘Savages,' he whispered, wrinkling his nose at the smell of their fires, of their burned food, of their unwashed bodies. (description of emotions)

In the case of the description of affects and emotions, we propose the specific category of the **affective DF**. Examples:

- *Sarah écrasa nerveusement sa cigarette.*
- *He hesitated for a moment and bit his lip. Bruno thought he was going to start crying and couldn't understand why.*

In the pilot studies, the descriptive DF has often turned out to be affective.

The **indirectly descriptive DF** corresponds to a repeated action or a gesture that characterizes a literary figure (gros fumeur [POL]; à la sensibilité accrue [FY], She took a tentative sip of wine [SENT])

The **“infra-narrative” DF** is specific to motifs which remain in the background of the action. These contribute to fleshing out a conversation or constitute a sequence of small actions in a script without narrative consequences for the main action. Examples:

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

- *"What are you doing Friday?" he persists, raising his eyebrows and taking a drag of his cigarette.*

The **"infra-descriptive" DF** is fulfilled by those motifs that introduce a descriptive detail, which often corresponds to a stereotype. Example:

- *Maintenant ils se taisaient, regardant par la fenêtre les reflets d'un ciel sinistre dans les eaux de la lagune.*
- *I sat in the vast kitchen, under a vast, grimy window, and watched Charlie making breakfast in a cloud of fat-smoke.*

The **cognitive DF** is used for motifs that presuppose cognitive processes (hypotheses, fear of events, reflections…). Examples:

- *Je ressentais une joie que je ne pouvais ressentir – dont j'avais l'impression de ne plus posséder les instruments pour la ressentir.*
- *I didn't think anyone had noticed I had an almighty crush on that riding instructor until I overheard him laughing about it with his friends – and his wife!*

We also propose two sub-categories for this DF: a cognitive **mnemonic** DF, for expressing memories and a cognitive **commentative** DF. The latter translates a reflexive stance (*reflective writing*) of a character on the very activity of writing, or the action of commenting, reflecting or the expression of a judgement.

The **pragmatic DF** refers to the motifs that express speech acts involving the characters of the novel (essentially in direct speech). When it fulfils this function, the motif establishes the relationships between the characters, within the speech that is integrated into the narrative text. Examples:

- *-N'en faites rien, Madame, je vous en prie, s'écria Eudeline.*
- *"No names", Newman promised.*

Finally, the term **interactive DF** is used when the motif triggers an interaction and introduces a dialogue sequence. Examples:

- *Un Mokranien casqué apparat sur l'écran.\n– Pourquoi nous avoir stoppés? s'indigna-t-il.*
- *A bleary-eyed Karl, who had obviously not slept since their last conversation, appeared on his screen.\n "Here it is," he said, exhaustion and triumph competing in his voice.*

2. Steps of the annotation

2.1 Identification of the "leader" RLT on the basis of the file comparing the 6 subgenres

The first step in the stylistic annotation is the identification of the RLTs (recurring lexico-syntactic trees) that are apt to constitute the same phraseological motif. These RLTs, which have been annotated syntactically and semantically, have been extracted automatically from the PhraseoRom corpus and collected in a comparative file where all of the six subgenres we have examined appear. The selection of the RLTs is based on two criteria: a quantitative criterion and a criterion of minimal significance: for each subgenre, we have chosen the RLTs that have a very high threshold of specificity and whose semantic and syntactic completeness – resulting from the presence of at least a verbal pivot and a nominal pivot – made it possible to consider them as significant units. This means that, for instance, the RLT <mis au point>, highly specific to SF (its threshold of specificity, calculated by applying the *loglikelihood ratio* method, reaches 207,53, while the minimal threshold was fixed at 10,83) nevertheless was not retained for the stylistic annotation due to it being incomplete and lacking a nominal pivot. The selected RLTs have been called "leaders", since their basic syntactic components have been used as a starting point for identifying other, similar RLTs, which represent syntagmatic and/or paradigmatic variation and, together with the "leader" RLT, constitute a phraseological motif.

2.2 Table of the RLTs specific to each subgenre selected for the 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	<dégainé 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

GEN	Log	SENT	Log	POL	Log	HIST	Log	SF	Log	FY	log
<standing at the window>	44,35	<blew a kiss>	51,63	<slammed the door>	48,27	<bit his lip>	116,54	<float in the air>	47,66	<wrinkled his nose>	31,96
<opening the windows>	46,08	<took a sip>	170,54	<picked up the phone>	236,10	<clenched his fists>	51,34	<appeared on the screen>	55,28	<raised his hands>	102,23
<playing the piano>	55,67	<had a crush>	109,02	<shut the door>	184,32	<broke the seal>	56,58	<the screens showed>	59,67	<drew his sword>	357,41
<sat in the kitchen>	80,97	<raised his eyebrows>	148,93	<replaced the receiver>	138,12	<bowed his head>	33,54	<tilted her head>	29,92	<open his eyes>	120,04
<came into the room>	55,01	<take her hand>	166,81	<found the body>	117,13	<raised his voice >	109,34	<was at the edge>	21,75	<drummed his fingers>	38,94

2.3 Verification of the automatic groups

After the selection of the "leader" RLT, we checked the automatic groups in the file comparing the six subgenres. This step is made necessary by the fact that the threshold fixed for grouping the most similar RLTs by comparing them to each other does not allow us to identify all of the occurrences, since it does not take into consideration all of the criteria that one would apply in the course of a manual regrouping. In fact, similar RLTs, which constitute the same motif, sometimes appear in different groups, while RLTs that constitute different motifs are subsumed under the same identifier. By doing this, the annotator also identifies more quickly the paradigmatic and/or syntagmatic variation relevant to the annotation of the phraseological motif.
Example:

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 Finding examples in Lexicoscope

As soon as the "leader" RLT and the similar RLTs have been identified, the annotator enters the regular expression that characterizes the RLT into Lexicoscope to examine the context in which the motif – established by the combination of several RLTs – can appear. The scrutiny of the textual extracts provided by the software constitutes a vital step for determining the DF associated with a motif and, more generally, for the stylistic interpretation of the role played by the phraseological segment both in the texts where it appears and in the subgenre to which it is specific.
The following table recapitulates the total number of occurrences of the annotated motifs in each subgenre based on 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_Siege	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

GEN	**Nb. Occ.**	**SENT**	**Nb. Occ .**	**POL**	**Nb. Occ.**	**HIST**	**Nb. Occ.**	**SF**	**Nb. Occ.**	**FY**	**Nb. Occ**
stand_at_window	113	blow_kiss	59	slam_door	182	bite_lip	214	float_in_air	38	wrinkle_nose	85
open_window	131	take_sip	314	pick_up_phone	221	clench_fist	104	appear_on_screen	54	raise_hand	549
play_piano	72	have_crush	69	shut_door	341	break_seal	53	screen_show	47	draw_sword	340
sit_in_kitchen	93	raise_eyebrow	680	replace_receiver	97	bow_head	170	tilt_head	115	open_eye	692
come_into_room	265	take_hand	645	find_body	137	raise_voice	264	be_at_edge	46	drum_finger	82

2.5 Recording the data and the analyses in the annotation file

The information that has been gathered on the basis of reading and examining the examples provided by Lexicoscope is inserted into the different columns of the annotation file, following the structure of recording the information that was explained above. This allows the annotator to define the discursive function(s) of a motif and to provide a comment on stylistic features that is more detailed with respect to the functions of a motif in its context as well as its contribution to the narrative structure.