Digital Stylistic Analysis in *PhraseoRom* Methodological and Epistemological Issues in a Multidisciplinary Project

Clémence Jacquot, Ilaria Vidotto, and Laetitia Gonon

Abstract This article is based on the literary corpus of the ANR-DFG[1] *PhraseoRom* project (https://phraseorom.univ-grenoble-alpes.fr/?language=en), which analyzes a large annotated corpus of novels (about 2,500 items) from the twentieth and twenty-first centuries in French, English and German, composed of historical novels, science fiction, fantasy, romance, crime fiction, and 'general literature' novels. The methodology used to build and explore this corpus is semi-automated by the interrogation tool Lexicoscope, based on automatic language processing methods and a corpus-driven approach. In this article, we present the stylistic annotation methodology of this corpus which links phraseological analysis of a large literary corpus together with stylistic issues concerning its formal and literary implications, through the concept of *motif*. We discuss the definition of *motif* and its methodological and epistemological implications on the contributions of digital tools for stylistic analysis.

Keywords *PhraseoRom*, stylistic annotation, *motifs*

1. Introduction

The development of multidisciplinary and interdisciplinary projects in the digital humanities and the growing importance of literary sources spawned by the massive digitization of archives and libraries in recent years has generated interest in the methods employed for their tool-based exploration. *Textual data* extracted from literary works and analyzed with various computational tools (statistical calculations, lexicometry or textometry) merit special attention and recognition of the intentional and stylistic

1 Agence Nationale de la Recherche & Deutsche Forschungsgemeinschaft.

Jacquot, Clémence, Vidotto, Ilaria and Gonon, Laetitia: Digital Stylistic Analysis in *PhraseoRom*. Methodological and Epistemological Issues in a Multidisciplinary Project, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 261–278.
DOI: https://doi.org/10.17885/heiup.1157.c19375

specificities of the text, statistically speaking (see Garric and Maurel-Indart 2010; 2011). This will ensure that their characteristics (textuality and the discursive dimension, for example) receive due consideration, particularly in quantitative studies and in comparisons with other qualitative studies.

Progressing toward a "reconquest of expression" (Rastier 2001: 69), text enrichment opens the way to new objects, new observable facts and, eventually, to theory construction. This in turn prompts the growth of new disciplines to take their place alongside corpus linguistics, discourse analysis, lexicometry and textometry. This is precisely what is happening with *digital stylistics*, in its early stages based partly on corpora of literary works.

Digital stylistics addresses a variety of questions that at times are distinctly shaped by national traditions (see Herrmann et al. 2015). Leaving this aside, for now we note that a significant digital stylistic terminology has already been accumulated, offering grounds for thinking that this novel discipline is ripe for taking its place in the continuity of stylistic topics (whose particulars would still need to be specified according to different uses and national academic practice). Digital stylistics is also conceived of as methodologically close to other *digital* disciplines, given its use of structuring, annotation and, more broadly, its conception of the scientific artifact for quantitative and, especially, statistical processing borrowed from lexicometry and textometry, from corpus linguistics, and other distant reading methods.

This raises questions such as what the need to call it *digital* stylistics implies, what it tells us about how what seems initially to be a methodology relates to its mother discipline, i.e. stylistics, or, for that matter, to the other above-mentioned disciplines, and, finally, if it really is merely a methodology.

Research in recent years has pointed out the need for defining the contours of digital stylistics, particularly in the context of projects situated at the intersection of linguistics and stylistics such as *PhraseoRom*.[2] Since this multidisciplinary project had linguists from diverse fields including syntax, semantics, and natural language processing collaborating with specialists in literary stylistics, the contribution made by stylistics to the joint effort needs sorting out.

Based on the methodology developed for annotating a large digitized corpus[3] in the *PhraseoRom* project, we propose to conduct a broader examination of how and by what means stylistics functioned in it. The questions to be answered include how the project parameters shaped the stylistic inquiry and how the project's multidisciplinary approach contributed to the interpretation of literary texts and to our knowledge of the literary genre.

2 https://phraseorom.univ-grenoble-alpes.fr/descriptif-projet.

3 In a first stage of the project, before extracting on lexico-syntactic recurrences and identifying the motifs of our literary corpus, we have already compared it with a non-literary contrast corpus of 65 million words in French.

2. Investigating the Novel Genre through Extended Phraseology

As already noted, *PhraseoRom* is an interdisciplinary project where linguistics meets literary studies and phraseology, stylistics, theory of literary genres, corpus linguistics and natural language processing NLP in particular. Given its research focus (the phraseology of the novel) and its—digital linguistic—methodology, the project falls within the domain of digital humanities in the humanities and social sciences.

Assuming that literary language is characterized by the statistically significant over-representation of lexemes (keywords), collocations or phraseologisms (see Siepmann 2015) that statistically characterize it, the project goal is to highlight and analyze these over-represented *patterns* or *motifs* from a linguistic and stylistic point of view. As such, it takes its place in the continuity of research carried out in recent years on the specificities of literary language (see Maingueneau and Philippe 1997; Philippe and Piat 2009; Vaudrey-Luigi 2011).

The corpus was developed to explore the French-, English- and German-language fiction discourse of the second half of the twentieth century because the novel is the literary genre with a remarkable, dynamic variety of subgenres and the widest readership. The French corpus is constituted as shown in Table 1.

Table 1 Quantitative information (authors, numbers of novels and of tokens) in the French corpus

Subgenres	Authors	Novels	Tokens
Fantasy (FY)	43	104	13,323,976
General (GEN)	170	445	34,334,554
Historical novels (HIST)	39	114	14,868,273
Crime fiction (CRIM)	84	194	17,859,351
Romance (ROM)	40	112	9,802,410
Science fiction (SF)	39	147	13,173,618
TOTAL	**365**	**1,116**	**103,362,182**

For these large textual corpora, the *PhraseoRom* project seeks, first, to establish what role extended phraseological units play in the construction of the literary text and, second, to create a typology of these units. The linguistic analysis of data on the semantic, syntactic and discursive levels is articulated for comparative purposes by a stylistic examination of different novelistic genres.

Working on the specific language of the novel requires an investigation of its generic boundaries and the *values* that derive from them (see Jouve 2010). This dictates

the inclusion of works in the corpus that, by editorial tradition and ideological representation of the novel's subgenres, are categorized by French literature specialists as *paralittérature* (see Couégnas 1992; Boyer 2002), i.e. popular fiction.

This pejoratively termed *paraliterature* contrasts a priori with a so-called *general* literature, which however represents a valued aesthetic project and is accorded pride of place in the field of artistic productions.[4] *Paraliterature* instead exists on the margins of this field. However, this also operationalizes it for a discussion on classifying novel subgenres, either through criticism or consequent on editorial and commercial conventions (see Boyer 2002; Genette 1987).

We do not subscribe to this axiological bias of devaluing popular works as stereotypical, poorly conceived, and as only intended for immediate cultural consumption (e.g., romance novels, science fiction, crime novels). Instead, we treat popular subgenres as literary works. This allows placing these novels in a broader set of contemporary fiction productions suitable for probing the relevance of the boundaries between subgenres, i.e., to critically examine what formally distinguishes (linguistically, for one) a novel released by a publishing house renowned for the high literary and aesthetic standards of its books (Minuit or Gallimard, for example) from a novel published by a less prestigious, institutionally less ambitious house or, for that matter, as part of a big publisher's clearly identifiable collection (e.g., the Folio SF collection by Gallimard).

Questions on what crossover margins can be identified between subgenres at the phraseological level despite obvious differences in thematic content, particularly in respect to plot or representation of a universe[5]—although by no means exhaustive—open new perspectives on the theory of genres (see Beauvisage 2001; Rastier 2011). They especially invite a reconsideration of the notions of *stereotype* and *cliché* (see Amossy and Herschberg Pierrot 2016) in the linguistic construction of literary works. In addition to the structural characteristics of popular seriality, its stylistic definition, as reflected in the arrangement of certain textual sequences and, above all, by a form of *constancy* of expression[6] also merit further study.

4 This literary fiction is covered in the *PhraseoRom* corpus by the GEN subcorpus (for *general* literature).

5 For example, the description of emotional states in romance novels (see among others Gymnich, Neumann, and Nünning 2007; Zymner 2003; Frow 2006; Duff 2000 or Monte and Philippe 2014 on textual genres).

6 "In fact, repetition, in all its forms, is, in both oral tradition and popular fictions, a generic marker, a formal mechanism expected by the public, i.e. a fundamental element of the reading contract, based on the interplay of the similar and the variation" (Boyer 2002: 76; translation by the authors).

3. Stylistic Corpus Annotation: Methodological and Epistemological Issues

3.1 Traditional Stylistics vs. *Digital Stylistics*

The lexicometric and textometric heritage of current *digital stylistics* mentioned in the introduction influences its definitions as a disciplinary field. The *digital* dimension came to be emphasized as such in recent years because it articulates the computational and statistical analysis of style (pattern recognition, authorship attribution, etc.) and because of its modeling according to the languages, genres and periods under examination. However, while corpus linguistics and the statistical analysis of texts (literary or otherwise) have long since taken root in linguistic studies, *digital stylistics* still tends to finds itself on the margin of cultural studies, relegated to adapting the same stylistic analysis units (phrase, sentence, paragraph, verse, etc.) used in traditional text exploration.

The methodology developed by this type of tool-based computational approach requires pragmatic redefinitions of the notion of style (see Herrmann et al. 2015), which it achieves by incorporating a contrastive, empirical dimension. The participation of *digital stylistics* in interdisciplinary and multidisciplinary research projects seems to considerably influence not only its objects of study but also its corpus design and how its results are rendered visible and readable (see Jacquot 2016).

3.2 Stylistic Annotation

The following sections describe the steps in the stylistic annotation of *motifs* (see section 5, "The Definition of *motif* Adopted in the *PhraseoRom* project"), followed by highlighting the place of stylistics in *PhraseoRom* and its contributions to the project.

3.2.1 Step 1: Extracting RLTs

The *PhraseoRom* corpora were syntactically annotated using the Xip analyzer (see Aït Mokthar et al. 2002), allowing the automatic extraction of recurrent lexico-syntactic trees (RLTs) from them (see Tutin and Kraif 2016). These RLTs include related, syntactically-dependent lexical units and are built from statistically significant collocate series based on a statistical association measure. As the name implies, the RLT depicts extracted lexico-syntactic information in the form of a tree whose branches diagram the relationships between components (see Figure 1).

This first step in the extraction of raw data as an RLT is followed by a more refined analysis of the information necessitated either by the irrelevance of extracted forms

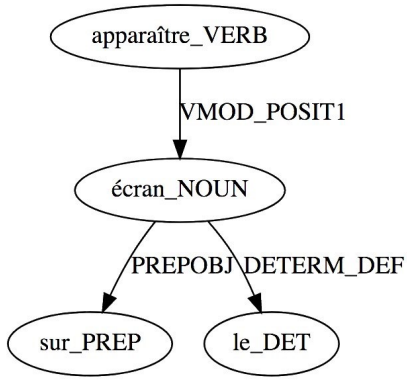

Fig. 1 Example of the RLT extraction <*apparaître sur l'écran*> ('appear on screen'), (Jacquot, Vidotto, Gonon, CC BY).

which produces *noise* (e.g. <*taken I have*>) or, in the pilot studies conducted so far, by the choice to expunge forms missing the verb (e.g. <*building inhabitants*> or <*and the king*>) in order to exclude solely referential expressions (e.g <*the king of France*>) (see Novakova and Siepmann 2019: 4–5).

3.2.2 Step 2: Selecting and Semantically Tagging LSCs

In this step, what we call recurrent lexico-syntactic constructions (LSCs) are isolated for study by retaining the LSC <*apparaître sur l'écran*> ('appear on the screen') in the RLT shown in Figure 1. The transition from RLT (row data) to LSC highlights a methodological progression in the phraseological analysis of literary texts: the LSC is therefore a culmination, an end product chosen for analysis by applying the criteria mentioned earlier (in particular, the requisite verbal pivot).

The retained LSCs (numbering some 6,450 items for the French corpus) then were annotated semantically and harmonized by applying a semantic grid developed by the team of semantics experts. A real breakthrough was achieved here just recently with the automation of semantic coding, thanks to a script written by the project's IT specialists. This has important ramifications for the stylistic annotation work because it ensures the transition from the LSC level to the level of the *motifs.* Making use of the lexical and syntactic similarity of the LSCs, the script on the one hand facilitates the automatic completion of semantic information as previously coded, i.e. by comparing the corpus of LSCs with a contrast file functioning as a dictionary, and, on the other hand, the coding primarily ensures the automatic grouping of similar but previously dispersed LSCs. In practice, this means that LSCs meeting a high threshold of similarity, for example <*monter les escaliers*> ('climbing stairs') and <*descendre les escaliers*> ('descending stairs'), will be clustered in the same group under a single numeric identifier.

This clustering affords the stylisticians a valuable, immediate, and reliable view of the data, with groupings showing satisfactory consistency, typically with zero noise. Above all, it makes quicker identification of the syntagmatic and paradigmatic variations specific to each LSC possible and, by extension, of the RLTs likely to form *motifs*. Thus, for each RLT specific to one or more subgenres, the stylistician upon spotting the identifier of the group under which it has been classified, immediately sees the other RLTs grouped by the script in this same set, letting scholars determine in turn if such occurrences constitute a *motif*.

A concrete example of this process is furnished by one of the most specific RLTs common to the CRIM French subgenre—<*prévenir la police*> ('alert the police'). Automatic grouping reveals that the same numeric group identifier (ID 1761) features the RLTs <*prévenir les flics*> ('alert the cops'), <*appelé les flics*> ('called the cops') and <*j'ai appelé les flics*> ('I called the cops'). The double paradigmatic variation on the pivot V (*appelé/prévenir*) and the N (*la police/les flics*) justifies the hypothesis that these occurrences represent the different expressions of the same *motif* specific to the crime novel. With these preliminary checks completed, annotation can commence. Stylistically annotating a *motif* by using a corpus-driven approach thus in essence means identifying the discursive function or functions that a particular *motif* is likely to assume in the context in which it appears. This step 3 requires further development.

3.2.3 Step 3: The Discursive Functions

The label *discursive function* (DF) was agreed on by the linguists and stylisticians at the start for use in the project. It means that a *motif*, i.e. a relevant grouping of LSCs, plays a role in the *textual coherence* (see Martin 1983: 100) of the fiction discourse. It could just as well be synonymously labeled a *textual function*, but for the sake of consistency the terminology initially adopted has been retained for all the studies that followed.

Baroni, for example, uses *discursive function* in discussing the meaning of verbal tenses: "It is important […] to keep in mind the dependence of the discursive function that a given textual structure can perform in its *context of use*, which naturally includes both the 'cotext,' the intertext and the genre of the story" (2015: 140; translation by the authors).

In a narrative text like a novel, the DF of *motifs* will be primarily narrative and descriptive: "A *predominantly narrative* text is generally composed of a series of actions, events, words and thoughts represented, but […] it also includes more or less developed descriptive moments" (Adam 2011: 267; translation by the authors). As more contrastive studies were conducted based on the statistical comparison of certain corpora (for example, CRIM vs. GEN) the stylisticians working on the project added more DFs to the initial.

The following examples illustrate the current state of our research. They are extracted from the French corpora, here translated into English. *Motifs* are in italics.

— Narrative and descriptive DFs are to be expected the most in novels.

(1) Le conducteur *consulta sa montre*: 8 h 15.[7]
'The driver *consulted his watch*: *8:15 am*'
narrative DF; the *motif* plays an active role in the plot

(2) Il *regarda de nouveau par la fenêtre*. Des couleurs de cuisine, voilà ce qu'étaient les couleurs de l'Italie.[8]
'He *looked out the window again*. Cooking colors, that's what the colors of Italy were.'[9]: descriptive function.

— Affective DF represents a special case of the descriptive function in which the *motif* refers to affects.

(3) Sarah *écrasa nerveusement sa cigarette*.[10]
'Sarah *nervously stubbed out her cigarette*.'

— Indirectly descriptive DF: a repeated action or a gesture in effect serve to describe the character (here as a *bad boy*).

(4) J'*écrasai ma cigarette contre un mur*, jetai le mégot sur le sol.[11]
'I *stubbed out my cigarette* against a wall, threw the butt on the floor.'

— Infranarrative DF: the term applied to DFs operating in the action's background. The *motifs* in this case serve to embellish the conversation without narrative consequences for the main action.

(5) —Tu feras mieux la prochaine fois, assure Alexandre *en allumant une cigarette*.[12]
'"You'll do better next time", Alexandre asserts, lighting a cigarette.'

7 M. Villard, *Cœur sombre*, 1997 (CRIM). N.B. This excerpt from French novel and all the following excerpts mentioned in this article have been translated by the authors and appear in quotation marks.
8 J.-Ch. Rufin, *Sauver Ispahan*, 1998 (GEN).
9 Then follows a descriptive sequence, triggered by the *motif*.
10 J.-C. Grangé, *Le Vol des cigognes*, 1994 (CRIM).
11 O. Gay, *Les Talons hauts rapprochent les filles du ciel*, 2012 (POL).
12 K. Giébel, *Juste une ombre*, 2012 (CRIM).

— Infradescriptive DF: here the *motif* provides a minimal, often stereotypical descriptive precision.

(6) Maintenant ils se taisaient, *regardant par la fenêtre* les reflets d'un ciel sinistre dans les eaux de la lagune.[13]
'Now they were silent, *looking out the window* at the reflections of a sinister sky in the waters of the lagoon.'

— Cognitive DF:[14] this variant covers *motifs* involving cognitive processes (hypotheses, apprehension of events, reflections, etc.).

(7) *Je sais pas* ce qu'il va devenir. J'ai pas les moyens de le changer d'école.[15]
'*I don't know* what's going to happen to him. I can't afford to have him change schools.'

— Commentary DF: denotes a special use of the cognitive function, when cognition relates to a reflection on writing activity (found only in GEN FR corpus).

(8) Bien sûr on aurait pu envisager *d'écrire un roman proustien jet set*…
Ça n'aurait eu aucun intérêt.[16]
'Of course, one could have considered *writing a jet-set Proustian novel*…; it would not have been interesting.'

— Pragmatic DF: this variation applies to *motifs* that express speech acts between the novel's characters (mainly direct speech). They establish coherent relationships between the characters, within the reported discourse integrated into the narrative text.

(9) – N'en faites rien, Madame, *je vous en prie*, s'écria Eudeline.[17]
'Do not do anything about it, Madam, *I beg you*, Eudeline cried.'

To reiterate, this typology was carried out progressively from the first experiments in text annotation to pilot studies and is invariably used in an empirical manner during the annotation process. Once the *motif* has been identified, the stylistician thoroughly

13 J. d'Ormesson, *San Miniato 1, Le vent du soir*, 1985 (GEN).
14 It seems that the cognitive function also supports a memory-related use, when cognition leads to the expression of memories.
15 D. Van Cauwelaert, *Hors de moi*, 2003 (GEN).
16 M. Houellebecq, *Les Particules élémentaires*, 1995 (GEN).
17 M. Druon, *Les Rois maudits* t. 3, 1956 (HIST).

reviews the textual examples provided by Lexicoscope (see Kraif 2016; Lexicoscope URL: http://phraseotext.u-grenoble3.fr/lexicoscope/).

The stylistician closely examines the *motif's* left and right cotext with special focus on certain parameters which, as shown by the pilot studies conducted throughout the project, may be relevant for determining the *motif*'s DFs, namely:

— the position of the *motif* on a transphrastic level (i.e., whether the *motif* appears in the surroundings of the direct speech, at the beginning or at the end of the sentence/paragraph/chapter);
— its intraphrastic distribution (i.e., whether the *motif* is coordinated or juxtaposed with or subordinated to other textual segments);
— and, finally, the possible presence of an *optional component* marking a significant syntagmatic variation, i.e. one or more terms that are not part of the minimum syntax of the *motif* but constitute its extended version.

The combination and recurrence of these parameters, in conjunction with the stylistician's reading expertise, contribute to the identification of the discursive function(s) of the *motif* in context. This is subsequently refined through stylistic interpretation, specifying why and how a particular *motif* is charged with a descriptive, narrative, emotional or cognitive value in the given subgenre.

Taking for example the above-mentioned *motif* formed around the RLT <*prévenir la police*>, the analysis of occurrences revealed that this *motif* has a cognitive function when it appears in direct speech, in interrogative (direct or indirect) or hypothetical modality—whereas it might conceivably be assigned a narrative function, which in fact it also has in other distributions:

(10) Que devait-elle faire? Décrocher son téléphone pour commencer. Et *prévenir… la police*?[18]
"What should she do? Pick up the phone for starters. And *alert…the police*?"

(11) – Je me demande si on ne devrait pas *appeler les flics*, suggéra Hélène, à court de plan C, D ou E.[19]
'I wonder if we shouldn't *call the cops*, suggested Helen, who was out of plan C, D or E.'

The character is portrayed as thinking, as weighing whether or not to take an action that, therefore, is still just virtual and not yet accomplished.

18 M. Chattam, *Le Cycle de l'homme 1*, *Les Arcanes du chaos*, 2013 (CRIM).
19 A. H. Japp, *Cinq filles, trois cadavres, mais plus de volant*, 2009 (CRIM).

4. Project Perspectives and Annotation Issues

The project's objective of stylistically annotating approximately 30 *motifs* for the French corpus (all subgenres) was achieved. For the English corpus, stylistic annotation is in progress, with *motifs* already having been selected.

One possible research direction that emerged from these early efforts is focusing on the *motifs* formed around LSCs specific to one or more subgenres, with the objective of providing a fertile contrast dimension for the stylistic interpretation in connection with determining DFs. This research could open up new perspectives on the generic and subgeneric configurations of the contemporary novel.

At this point, it can already be asserted that stylistic annotation is of key importance to the project. Although it represents its final stage and in effect is the culmination of an enormous amount of computer and linguistic processing of the corpus data, the stylistic aspect nevertheless crowns the entire effort. Clarifying the functioning of the *motif* in the narrow cotext and, more broadly, in the generic context, helps link the purely statistical and linguistic dimensions of the project to its textual dimension. In other words, stylistics provides an interpretation of the raw data collected by establishing, for example, whether the specificity of a quantitatively calculated *motif* is also the reflection or the guarantee of a stylistic specificity—i.e., of a salience (Fr. *saillance*).

Through the intervention of *motifs* and their DFs, stylistic analysis moreover can contribute to redefining the conventional and editorial contours of a subgenre. Furthermore, the contrastive analysis of *motifs* common to one or more subgenres invites us to rethink the sometimes fossilized lines drawn between the different *paraliterature* subgenres or the questionable distinction between *high* literature and so-called *popular* literature—a tricky issue if there ever is one.

True to its nature as a *hybrid* discipline, stylistics—positioned uncomfortably amid the sciences of language, literature and now also of the *digital humanities*—bridges the gap between linguistic issues and the more strictly literary questions raised by the *PhraseoRom* project. However, a major issue with the notion of *motif* that emerged from this initial phase of stylistic annotation still needs addressing, as discussed next.

5. The Definition of Motif Adopted in the *PhraseoRom* Project

Before concluding this section on stylistic annotation, a more precise definition of the concept of *motif* is called for. Up to this point, it has provisionally been defined as a "relevant grouping of LSCs." As we have seen, LSCs already provide interesting

information for work on differentiating fiction genres, but their description does not take into consideration the textual dimension of the corpus, i.e. their role in structuring texts. This is where the concept of *motif* comes in by making it possible to integrate its discursive component into the phraseological dimension of this analysis.

In this case, the *motif* is not, as it is characterized in thematic criticism,[20] "an imaginary object or a metaphorical term […] precisely because it constitutes one of these microsystems that is found 'assembled in a system' in a complete oeuvre" (Bellemin-Noël 1972: 26; translation by the authors).

In other words, a *motif* as conceived here is not a fictional, symbolic or constitutive element of the imaginary of a work, but an observable phraseological element characterized by continuous or discontinuous units combining several elements. However, our definition of the *motif* includes a dimension of syntagmatic variation that can be found in thematic criticism (see Richard 1979). Hence, this is the definition adopted for the *PhraseoRom* project:

> [Motifs] […] display lexico-syntactic regularities and variations at the syntagmatic and paradigmatic levels while simultaneously performing particular discursive/narrative functions. They are therefore recurrent linguistic units that can be described at the levels of lexico-grammar, semantics and pragmatics/discourse (Longrée and Mellet 2013; Legallois 2012). [They] furnish a link between linguistics and literary studies to the extent that they collaborate in the construction of scripts and schemas; and are situated—unlike traditional literary motifs—where social scripts and fictional scripts (Baroni 2007; 2009) intersect. Motifs as we understand them cannot be identified by fully automatic procedures, but instead require the linguist and the literary scholar to make a judgement (Novakova and Siepmann 2019: 9–10)

6. Stylistic Annotation and the Granularity of *Motifs*

Based on the three criteria of 1) syntactic and lexical regularities, 2) syntactic and paradigmatic variations and 3) the involvement of DFs, the *motif* is a productive concept enabling the gathering of more extensive phraseological units than can be collected with simple collocations analysis, while excluding fixed expressions thanks to the variation criterion. It facilitates recognition of salient sequences that otherwise would not be thought of *a priori* as candidates for systematic grouping and for having their role in the cohesion and structuring of novelistic texts examined. Starting from a given

20 See Bellemin-Noël (1972), Richard (1961) and (1979).

LSC specific to one or more subgenres—here <*regarder par la fenêtre*> ('looking out the window')—a *motif* is realized in a more or less diversified way, as shown in the following examples:

(12) La responsable commerciale *regarde par la vitre sale*, elle n'est pas très concentrée. Comme les gens marchent vite, se dit-elle, c'est parce qu'il pleut à torrents.[21]
'The sales manager *looks out through the dirty glass*, not really concentrating. People are walking fast, she thought, because it's pouring rain.'

(13) Comme d'habitude, je *contemplai par la fenêtre* le mouvement de la rue.[22]
'As usual, I *gazed out the window* at the movement in the street.'

(14) Estelle se lève, s'étire, *jette un regard par le hublot*: 'Tiens, tu es là, la mer?'[23]
'Estelle gets up, stretches, *looks out the porthole*: "Really, are you there, sea?"'

(15) Mais, juste avant de sortir, Blunt *regarda machinalement par la fenêtre* et, à travers les volets, vit que deux hommes semblaient surveiller la maison: il s'affola.[24]
'But just before leaving, Blunt *automatically looked out the window* and, through the shutters, saw that two men seemed to be watching the house: he panicked.'

As shown by these examples, the *motif* bundles several LSCs that are similar and can vary on the syntagmatic axis in an extended version of the *motif* (here by adding an epithet, a circumstantial adverb, a complementation, etc.) and on the paradigmatic axis (by a nominal pivot variation: *fenêtre/hublot/vitre*, and a verbal pivot variation: *jeter un regard/regarder/contempler*). This *motif* also illustrates the diversity of DFs as determined by the different contexts it appears in:

— Ex. (12): Cognitive function. The *motif* <*regarder par la fenêtre*> gives access to the character's thoughts.

21 G. Brisac, *Dans les yeux des autres*, 2014 (GEN).
22 E. Ionesco, *Le Solitaire*, 1973 (GEN).
23 J. Boissard, *Croisière*, 1988 (ROM).
24 G. Perec, *La Vie mode d'emploi*, 1978 (GEN).

— Ex. (13): Infradescriptive function: introduces a minimum descriptive precision updated by the presence of the complementation *le mouvement de la rue*.
— Ex. (14): Infranarrative function: part of a sequence of minimal actions and of a "wake-up" script.
— Ex. (15): By use of the adverb *machinalement* and, incidentally of the proposition *il s'affola*, the *motif* reflects the character's emotions, hence here it performs an affective function.

However, the stylistic annotation of the *motifs* raises questions about the granularity of the *motif* definition, such as what objective criteria are applied in grouping LSCs into *motifs*. The very diversity of the LSC forms (i.e. the extension and variation of the *motif*) can be problematic.

The following examples of LSCs specific to the science fiction subgenre <*apparut sur les écrans*> ('appeared on the screen'); <*inscrit sur l'écran*> ('written on the screen'); <*voir sur l'écran*> ('to see on the screen') and <*défilaient sur les écrans*> ('scrolling on the screen') meet the criteria spelled out above for defining the *motif*. They clearly represent paradigmatic variations of the verb, they have DFs in the various instances proposed by them in context, and they would therefore likely compose a single standard *motif* like <*apparaître sur l'écran*>. However, this result of the *motif* modeling does not allow for the aspectual and especially the actancial dimensions of the different LSCs proposed; in particular, grouping under the same model <*apparaître sur l'écran*> and <*voir sur l'écran*> presents a problem.

(16) Enfin, le visage redouté *apparut sur l'écran*. Ses traits étaient impassibles.[25]
'Finally, the feared face *appeared on the screen*. Its features were impassive.'

(17) Vivement intéressé par ce qu'il *avait vu sur l'écran télévisionneur*, le professeur Yegov, d'un ton légèrement doctoral, s'empressa d'ajouter: […].[26]
'Deeply interested in what he *had seen on the television screen*, Professor Yegov, in a rather bombastic tone, was quick to add: […]'

In example (16), the subject of the inchoative verb *apparut* is not the agent of what is happening. This contrasts with example (17), in which the subject *il*, referring by cataphor to *le professeur Yegov* refers to a human animate agent of the imperfective verb

25 J. Wintrebert, *Les Olympiades truquées*, 1987 (SF).
26 J. Guieu, *L'Homme de l'espace*, 1954 (SF).

voir. This semantic contrast tends to be confirmed in the different occurrences of the corpus, hence here it seems necessary to identify two distinct *motifs* stemming from these two types of LSC: <*apparaître sur l'écran*> and <*voir sur l'écran*>.

That this is a relevant distinction is corroborated by the fact that the two *motifs* perform different DFs in context. The appearance of an entity (face, person, object, message, numbers), for example, indicates the beginning of a new narrative sequence and moves the action forward (in accordance with the aspectual inchoative value of the verb). The *motif* <*apparaître sur l'écran*> seems more likely than <*voir sur l'écran*> to indicate progressive action and plot progression.

At this point, the two previously defined *motifs* <*apparaître sur l'écran*> and <*voir sur l'écran*> could be grouped to form a more abstract syntactic-semantic pattern, which would constitute a final stage of annotating the corpus: from the LSC to the *motif sensu stricto*. It could take the following form: [Verb of vision + Preposition + Inanimate object].

This solution offers twin advantages: for one, it preserves the theoretical coherence of the *motif*'s definition through inclusion of a finer granularity in the semantic and stylistic description, and, for another, it bundles syntactically identical constructions that, from a purely syntactic point of view, would not necessarily require contrasting. Clearly, stylistic analyses and the annotation work feed into the broader reflection on the phraseological notion of *motif*.

7. Conclusion: Stylistic Analysis as Starting and End Point

The *PhraseoRom* stylisticians tracked the evolution of the project and were involved in each of its stages; paradoxically, however, they performed their work both far upstream and far downstream in the project timeline.

To begin with, they built the French corpus on which most of the pilot studies were carried out. At the end of the pilot studies, the IT specialists of the project worked from September 2016 to August 2018, among other tasks, on compiling the lists of works to be included in each novelistic sub-corpus. The GEN corpus, for example, was sourced from the list of books awarded the Goncourt prize and other prizes since 1950. The difficulties the programmers encountered in pursuing his task were, for one, finding that not all these awarded books were novels (they included autobiographies, stage plays, collections of short stories, etc.), and, for another, not knowing where to find romance literature titles, historical novels, and so on.

In response, starting from the very incomplete files compiled by the project's IT specialists, the stylisticians combed through specialized sites to find suitable titles. This variously required reading book summaries or locating a particular novel in the

collection it was part of to make correct classifications. For instance, a novel dealing in detail with a historical period and published by Minuit would pose the question for the researchers on whether it should be classified in the GEN or HIST corpus. This is precisely what happened with books by Anna Gavalda, whom some consider a *literary* author (GEN) while others view as a writer of unrefined stories (ROM). Obviously, the criteria applied here at times had to be subjective or at least based on reading experience which, with enough practice, evolved into reading expertise. When the stylisticians were stumped, they looked for clues in collections or relied on intuition developed from reading excerpts (sometimes just the publisher's jacket blurb) from these ambiguous works to select the appropriate subgenre.

Thus, the skill set required for building coherent corpora in the first stage included reading skills and at least some background in popular literature. By contrast, the second stage called for competence in literary analysis for relating the micro context to its immediate environment but also for mastering the specificities of the subgenre a *motif* was a potential candidate for. Furthermore, to refine this recontextualization, the stylisticians also had to analyze *paraliterature*.

The foregoing tasks set the stage for stylisticians to critically examine in a contrastive manner the more or less permeable lines drawn between subgenres from fresh perspectives. The observations produced by the quantitative analysis of the corpora and the stylistic annotation of the selected *motifs* will be instrumental in this effort.

Returning to the initial question on the place of stylistics in multidisciplinary digital projects such as *PhraseoRom*, we can assert that it plays an active role in performing the following vital functions:

— Upstream: it creates the literary coherence of the corpora and provides elements of literary problematization of the data, for example diegetic stereotypy versus linguistic stereotypy. Furthermore, stylistic annotation contributes to the process of building up a corpus by conceptualizing and complexifying *motifs* through the identification of DFs.
— Downstream: stylistic analyses facilitate the study of extracted data and stimulate critical reflection on conventional boundaries (literary criticism, academic work, publishing house collections, etc.) between subgenres by shedding light on the very definitions—linguistic, phraseological, stylistic—of the boundaries.

Digital tools like these and the research they enable change the center of gravity of stylistic thinking by letting us shift the focus from the auctorial and the definition of the author's style. Instead, they sensitize us to the French stylistic concept of *salliance* (as significant recurrence, see Jacquot: 2016) as redefined by the insights gained with this digital tool-enabled stylistics research into both recurrence and specificity within a subgenre, as here, or within any other desired ensemble.

References

Adam, Jean-Michel. 2011. *Les Textes: types et prototypes: récit, description, argumentation, explication et dialogue*. Paris: Armand Colin.

Aït Mokhtar, Salah, Jean-Pierre Chanod, and Claude Roux. 2002. "Robustness beyond Shallowness: Incremental Deep Parsing." *Natural Language Engineering* 8 (2-3): 121–44.

Amossy, Ruth, and Anne Herschberg Pierrot. 2016. *Stéréotypes et clichés*. Paris: Armand Colin.

Baroni, Raphaël. 2007. *La Tension narrative: suspense, curiosité et surprise*. Paris: Seuil.

Baroni, Raphaël. 2009. *L'Œuvre du temps. Poétique de la discordance narrative*. Paris: Seuil.

Baroni, Raphaël. 2015. "Temps, mode et intrigue: de la forme verbale à la fonction narrative." *Modèles linguistiques* 71: 125–42.

Beauvisage, Thomas. 2001. "Exploiter des données morphosyntaxiques pour l'étude statistique des genres : application au roman policier." *TAL* 42 (2): 579–608.

Bellemin-Noël, Jean. 1972. "Le motif des orangers dans 'La Chartreuse de Parme.'" *Littérature* 5 (1): 26–33.

Boyer, Alain-Michel. 2002. *Les Paralittératures*. Paris: Armand Colin.

Couégnas, Daniel. 1992. *Introduction à la paralittérature*. Paris: Seuil.

Duff, David, ed. 2000. *Modern Genre Theory*. London, New York: Routledge.

Frow, John. 2006. *Genre. The New Critical Idiom*. London, New York: Routledge.

Garric, Nathalie, and Maurel-Indart Hélène. 2010–2011. "Vers une automatisation de l'analyse textuelle." *Texto! Textes et Cultures* 15 (4), and 16 (1): 3–13.

Genette, Gérard. 2002. *Seuils*. Paris: Seuil.

Gymnich, Marion, Birgit Neumann, and Ansgar Nünning, eds. 2007. *Gattungstheorie und Gattungsgeschichte*. Trier: WVT Wissenschaftlicher Verlag.

Herrmann, Berenike, Christof Schöch, and Karina van Dalen-Oskam. 2015. "Revisiting Style, a Key Concept in Literary Studies." *Journal of Literary Theory* 9 (1): 25–52.

Jacquot, Clémence. 2016. "Rêve d'une épiphanie du style: visibilité et saillance en stylistique et en stylométrie." *Revue d'Histoire Littéraire de la France* 116 (3): 619–39.

Jouve, Vincent. 2010. *Pourquoi étudier la littérature?* Paris: Armand Colin.

Kraif, Olivier. 2016. "Le lexicoscope: un outil d'extraction des séquences phraséologiques basé sur des corpus arborés." *Cahiers de lexicologie* 108: 91–106.

Legallois, Dominique. 2012. "La colligation: autre nom de la collocation grammaticale ou autre logique de la relation mutuelle entre syntaxe et sémantique?" *Corpus* 11: 31–54.

Longrée, Dominique, and Sylvie Mellet. 2013. "Le motif: une unité phraséologique englobante? Étendre le champ de la phraséologie de la langue au discours." *Langages* 189: 68–80.

Maingueneau, Dominique, and Gilles Philippe. 1997. *Exercices de linguistique pour le texte littéraire*. Paris: Dunod.

Martin, Robert. 1983. *Pour une logique du sens*. Paris: PUF.

Monte, Michèle, and Gilles Philippe, eds. 2014. *Genres et textes: Déterminations, évolutions, confrontations*. Lyon: Presses universitaires de Lyon.

Novakova, Iva, and Dirk Siepmann. 2019. *Phraseology and Style in Subgenres of the Novel. A Synthesis of Corpus and Literary Perspectives*. Basingstoke: Palgrave Macmillan.

Philippe, Gilles, and Julien Piat. 2009. *La Langue littéraire: une histoire de la prose en France de Gustave Flaubert à Claude Simon*. Paris: Fayard.

Rastier, François. 2011. *La Mesure et le grain: sémantique de corpus*. Paris: Honoré Champion.

Richard, Jean-Pierre. 1961. L'Univers imaginaire de Mallarmé. Paris: Seuil.

Richard, Jean-Pierre. 1979. *Microlectures*. Paris: Seuil.

Siepmann, Dirk 2015. "A Corpus-based Investigation into Key Words and Key Patterns in Post-War Fiction." *Functions of Language* 22 (3): 362–99.

Tutin, Agnès, and Olivier Kraif. 2016. "Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines: l'apport des arbres lexico-syntaxiques récurrents." *Lidil. Revue de linguistique et de didactique des langues* 53: 119–41.

Vaudrey-Luigi, Sandrine. 2011. "Ce que la linguistique dit des textes littéraires – De la signature stylistique à la reconnaissance d'un style d'auteur." *Le français aujourd'hui* 175 (4): 37–46.

Zymner, Rüdiger, ed. 2003. *Handbuch Gattungstheorie*. Stuttgart: J. B. Metzler.