DE GRUYTER
MOUTON

Dominique Legallois, Thierry Charnois, Meri Larjavaara (Eds.)

THE GRAMMAR OF GENRES AND STYLES

FROM DISCRETE TO NON-DISCRETE UNITS

TRENDS IN LINGUISTICS

Dominique Legallois, Thierry Charnois, Meri Larjavaara (Eds.)
The Grammar of Genre and Styles

Trends in Linguistics Studies and Monographs

Volume 320

The Grammar of Genres and Styles

From Discrete to Non-Discrete Units

Edited by
Dominique Legallois
Thierry Charnois
Meri Larjavaara

DE GRUYTER
MOUTON

ISBN 978-3-11-058968-9
e-ISBN (PDF) 978-3-11-059586-4
e-ISBN (EPUB) 978-3-11-059284-9
ISSN 1861-4302

Library of Congress Control Number: 2018009946

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Typesetting: Integra Software Services, Pondicherry
Printing and binding: CPI books GmbH, Leck
♾ Printed on acid-free paper
Printed in Germany

www.degruyter.com

Contents

Grammar of genres and styles: an overview

1 Investigating genres with grammar

In a book that enjoyed some success in the US, the psychologist J. Pennebaker (2011) analyzed *the secret life of the most forgettable words*, that is to say, grammatical words and their use in speech, arguing that grammatical words and how we use them are indicative of our personalities. The book gives several examples, among which this one: in daily conversations, e-mails, informal conversations and blogs, women use first-person singular pronouns more than men do. The reason is that women, on average, are more self-aware and self-focused than men. On the contrary, men use articles more than women do. Since articles are used with nouns, and especially concrete, highly specific nouns, men talk about objects and things more than women do. The issue here is not to discuss the relevance of such a study or to declare a war of the sexes: Are women more self-focused? Are men more materialistic and trivial? The goal is instead to show, with this reference to a supposedly serious book, that the domain of the analysis of the grammar of textual genres, registers and styles, can be addressed by disciplines other than linguistics. What is more, the reader of the book by Pennebaker can easily check the researcher's findings for himself, using his own corpus and the various tools available. In short, ever since the work of Bakhtin, the investigation of genres and registers has expanded considerably in different directions and orientations. The main motivation is obviously understanding the functioning of genres themselves, and their lexical and grammatical content. Lately, however, some studies have been conducted in which genre analysis is an essential parameter for understanding:

- syntax and its variation. The distinction here is between a grammar *of* genres on the one hand, and an analysis of grammar that is genre-based (grammar *in* genres) on the other hand. Studies that emphasize the latter orientation take a specific phenomenon of syntactic variation as their starting point and examine how its patterns of occurrence can be explained in relation to genre. One can speak of genre effects on syntax (Dorgeloh and Wanner 2010);
- sociolinguistic variation. The aim is to analyze the influence of registers and genres on the shape of language forms and on the patterns of language use by groups of complex speech communities (Biber and Finegan 1994);
- teaching of grammar. In educational contexts, a genre-based approach to teaching grammar is opposed to naturalistic models of language learning (i.e., considering language as a stand-alone set of rules). A genre-based approach, on the contrary, emphasizes the social constructedness of language, and holds that grammar is a set of resources which varies according

https://doi.org/10.1515/9783110595864-001

to usage (Knapp and Watkins 2005). This applies both to first and to second language learning;
- contrastive linguistics, which studies the ways in which linguistic features vary across genres/registers cross-linguistically (Neumann 2014; Lefer and Vogeleer 2016);
- text classification in Natural Language Processing (NLP). Automatic classification methods are becoming more and more efficient. Whether it is for the language industry, or for linguistic research, clustering techniques (e.g. factor analysis, Naive Bayes Multinomial Classifier) are used to calculate intertextual distances. They are applied to various issues such as author identification, plagiarism detection, Spam filtering, etc. (Oakes 2014; Fang and Cao 2015).
- genre evolution. "Distant reading" (Moretti 2013) and "Macroanalysis" (Jockers 2013) reveal, for example, variations of the literary genre during history. The use of data mining and corpus linguistics techniques can help to better understand how a particular genre, but also a cultural and economic category such as the bestseller (Archer and Jockers 2016) develops, not only in terms of the themes it expresses, but also in the linguistic forms (especially grammatical forms) that characterize it.

As one can see, motivations and applications are numerous.

This book is about genres and their grammars, but also about the grammar of style, since some of the chapters study the style of authors, and even the "oratory" style of characters. It is indeed possible to use the term "grammar of styles" whenever stable sets of features that characterize a style are identified and described. The link between grammar and style was first put forward by Bakhtin:

> One might say that grammar and stylistics converge and diverge in any concrete language phenomenon. If considered only in the language system, it is a grammatical phenomenon, but if considered in the whole of the individual utterance or in a speech genre, it is a stylistic phenomenon. And this is because the speaker's very selection of a particular grammatical form is a stylistic act. But these two viewpoints of one and the same specific linguistic phenomenon should not be impervious to one another and should not simply replace one another mechanically. They should be organically combined [...] on the basis of the real unity of the language phenomenon. Only a profound understanding of the nature of the utterance and the particular features of speech genres can provide a correct solution to this complex methodological problem. (Bakhtin 1986: 66–67).

The present book provides new methods and new findings about the grammar of genres and styles. Since Biber's early studies, multi-dimensional analyses have become increasingly common in many different fields such as linguistics, didactics, or automatic processing. At present, however, new analyses are emerging that draw on these earlier studies, but also propose new methodological solutions.

2 Goals and motivations

In his foreword to *Genres on the Web* (Mehler et al. 2010), Martin writes:

> As a reader, I'm looking for two things from a new book on genre. First, does it offer some new tools for analysing genres; and second, does it explore genres that haven't been much studied before? (Martin 2010: V).

We can claim that the present book meets these two criteria: the majority of chapters – this was the reason for their selection – offer new tools and methods, based on precise empirical studies. This innovative approach is also reflected in the genres studied here, several of which have seldom been examined: social reports on at-risk children, machine-translated texts, theater characters' dialogues, texts produced (by a socio-technical device) during visits of tourist attractions, for example. But above all, it is the types of units taken into account which make the originality of this book. These units are of various kinds, but in every chapter the units considered share one essential characteristic: they are non-discrete, that is to say they are sequential or syntagmatic. In fact, the vast majority of previous studies on genres concern units (or "descriptors") that can be considered as discrete: generally, they are simple, individual and therefore not syntagmatic. They occur in a distinct temporal or spatial order. They may be lexemes annotated on the basis of semantic characteristics, or grammatical words, or morphosyntactic categories: for example, adverbial subordinators, attributive adjectives, existential *there*, gerunds, modals, the perfect aspect, nominalizations, first person pronouns, proper nouns, infinitives, punctuation, etc. To this can be added data such as word and sentence length, lexical density, and type-token ratio. Studies that only consider discrete units belong to a "paradigmatic" approach. The paradigmatic approach rests upon the quantification of morpho-syntactic categories. For instance, in his work on oral discourse in the academic community, Biber (2006) revealed the overuse (in comparison with written discourse) of first person pronouns, evaluative expressions ("mental" verbs, modal adverbs, etc.), WH-questions, etc. By means of factorization, it is possible to determine a set of properties particular to a specific genre and to find correlations between linguistic features. Needless to say, discrete units are not neglected in this book, as their statistical treatment is highly significant. However, all the chapters insist on the need to also consider, along with discrete units, the units that we call non-discrete. While discrete studies analyze tables of quantified units, non-discrete studies aim to investigate chunks of speech or texts, and in particular three types of units: 1- combinations of linguistic units, that is to say *sequential patterns* (e.g. lexical

bundles, or syntactic schemes); 2- production units (e.g. prosody); 3- relations between units (e.g. reference chains). These "non-discrete" studies belong to a "syntagmatic" approach. It is only very recently that researchers have become interested in these units.

Interestingly, these three types of non-discrete units are associated with different types of analysis. The study of the relations between units, for example, is particularly relevant for characterizing genres in terms of textual coherence or informational organization. Generic specificities may thus emerge: different genres prefer different types of organization. Moreover, in order to assess the peculiarities of a genre or a style, analyses based on the automatic classification of texts can rely on *sequential patterns* (cf. the notion of "motif" below). Finally, in process-oriented approaches, linguists can also take into account production units, e.g. prosodic patterns, since these can be very valuable indices for describing a particular genre.

The chapters deal mainly with Romance languages (French and Spanish), one with Latin, and two articles with English (one with translations to German). This choice of languages can be seen as complementary to English, which is the language mostly studied from this perspective, and those interested in these languages will certainly gain new insights. However, the methodological and theoretical issues remain our main concern here; the choice of languages under investigation is of secondary importance.

In the rest of this introduction, we briefly develop these different categories, while also presenting the articles dealing with each type of non-discrete units.

3 Non-discrete units for the grammar of genres and styles

3.1 Relational units and textual and informational organization

Relational units concern two or more units that create by their relationship a texture, binding parts of continuous text together. Information structure, reference chains, and coherence relations are the relational units discussed in this book in three articles that promote a textual perspective on genre. This perspective seeks to determine what distinguishes a text from a group of unrelated sentences, providing insights into the linguistic features that cause a text to be interpreted as a coherent and cohesive unit.

3.1.1 Information structure and thematic organization

Information structure is generally studied in relation to sentence forms and textual progression. Although some researchers have for a long time highlighted the influence of genre on certain informational structures (for example, Reinhart 1982), the analysis of these non-discrete units as a discriminating feature of genres remains to be done. Very interesting perspectives can be found, however, in Biber and Gray (2010) on the increase in compressed structures in English academic writing over the last 100 years, and the paper by Kerz (2012) on the comparison of information structuring and "compression" strategies in research article abstracts, and fiction.

From a Systemic Functional Linguistics perspective, **A. E. Martínez-Insua and J. Pérez-Guerra**, in this volume, base their analysis on the distinction between contentlight and contentful subject themes in order to describe medical texts (1500–1800) and news (1661–1791) genres in Modern English. They confirm in their article the hypothesis that the context and the text genre play an important role in language users' choices of information structure. The themes are more contentful in medical texts addressed to learned readers vs. unlearned audiences, and in the news genre "hard" topics have more contentful themes than other topics. The degree of textual formality, the audience addressed and the whole context thus play an important role in language users' choice of themes and hence information structure and thematic organization.

3.1.2 Reference chains

Lexical chains are sometimes identified in texts for the purpose of summaries. Reference chains (Chastain 1975), however, are more complex. Studies on the detection of textual genres on the basis of how co-reference is constructed are very sparse, no doubt because of the difficulty in conducting a sufficiently convincing quantitative analysis (parsing problems are numerous and annotation is necessary). However, in a functional perspective based on Halliday and Hasan (1976), Swanson (2003) showed that the ways in which co-reference is made are very good indicators of genres (in this case, academic journal texts, fictional narrative texts, and news magazine texts).

In this book, **C. Schnedecker** shows how a study of reference chains is essential in defining the main characteristic of a genre. To illustrate and support this argument, the author makes use of an existing corpus of incipits of fairy stories and news briefs, which are closely related to each other in terms of the production

situation, codification, length and content. This analysis enables the author to highlight the limitations of a paradigm approach which aims to identify genres by focusing on categories of nouns and pronouns and then on the expression of anaphora. Secondly, she advocates a method which "transcends" paradigm approaches and syntagmatic approaches: this "configurational" approach combines multi-dimensional features, namely discrete units (quantification of grammatical and lexical categories) with non-discrete units based on reference chains and the recurrence of their motifs.

3.1.3 Coherence relations

Just like reference chains, the structural organization of a text can be a very good indicator of genre. Coherence relations between various textual segments can correspond not only to the local but also to the global organization of the text. Again, probably because the annotation of coherence relations is extremely time-consuming, there are very few studies on the relationship between coherence and textual genres. One of the few is *Discourse on the Move* (Biber, Connor and Upton 2007), a convincing investigation into the relationship between moves (Swales 1990) and overall textual organization.

On similar lines, but adopting a bottom-up approach and using Rhetorical Structure Theory (RST) concepts developed by Mann and Thompson (1988), **O. Méric** offers here an analysis that shows how coherence relations can characterize a specific discursive genre, namely discourse produced in a guided tour (French and Spanish visits assisted by a socio-technical device, and French and Spanish visits guided by an education and visitor service officer). The method is grounded on segmenting the texts into units, named contributions. These units are based on the constraint of relevance and completeness. Their relations are tagged (according to RST). In this way, scholars can discover the structural organization of texts, making it possible to suggest a specific set of features that describe the representative prototype of the text genre studied.

3.2 Sequential patterns and textual classification for grammar of genres and styles

Studies using automatic classification techniques generally focus only on the computational aspects of the analysis. On the contrary, the contributions here address different linguistic aspects related to sequential patterns. They

empirically demonstrate the usefulness of these non-discrete units for the tasks of identification or characterization.

Word-based n-grams, or more simply *N-grams*, or *lexical bundles* (in the corpus linguistics tradition), or *repeated segments* (in the French lexicometric tradition, cf. Léon and Loiseau 2016) are strings of n contiguous words (2 words, 3 words, etc.) from a given corpus. They have proved highly relevant over the last 10 years in the teaching of Language for Specific Purposes. Here are some examples, pertaining to academic prose: *it should be noted, as we have seen, on the other hand, at the same time, it is clear that* (examples from Biber and Conrad 2009). Thanks to the availability of specific tools, these units which are most often prefabricated, become very easily identifiable. They are mentioned or under investigation in several articles in the present book, either to compare them with other units such as "bursts" (see the chapter by G. Cislaru and T. Olive), or to show their analytical limits. Scholars can also take into account not only word-grams, but also Part-Of-Speech-grams (POS-grams) – that is, n-grams of grammatical tags – in order to determine characteristic syntactic patterns. In their study, E. Lapshinova-Koltunski and M. Zampieri show the effectiveness of such non-discrete units, in the discrimination of genres, whereas, for D. Longrée and S. Mellet, this method does not really account for the sequential structure of the text's linearity.

But a new type of unit, called "motifs", is taken into account in several of the chapters in this book. *Motifs* are both more schematic than n-grams, and more specific than POS-grams: they are lexically open syntactic patterns. They are considered to be characteristic of a discourse genre or an author. D. Longrée and S. Mellet (this volume) give a more formal definition of motif:

> What is a "motif"? In a formal way the "motif" is defined as an ordered subset of the textual ensemble, formed by the recurring combination of n elements provided with its linear structure. Thus, if the text is formed by a certain number of occurrences of elements A, B, C, D and E, a "motif" can be the recurring micro-structure ACD or AAA, etc., without here prejudging the nature (lexical, grammatical, metrical, etc.) of the elements A, B, C, D and E in question: the 'motif' is only the framework – or the collocational pattern – accommodating a range of parameters to be defined and capable of characterizing the diverse texts of a corpus, or even the different parts of a text.

In a word, motifs are sequential patterns that combine different levels of abstraction (word forms, lemmas, POS-tags, linear order). Therefore they have a multidimensional nature (Quiniou et al. 2012). For example, in their contribution, F. Frontini, M. Amine Boukhaled and J.-G. Ganascia identify a statistically significant motif that is typical of Harpagon, the main character in Molière's play L'Avare:

[PRO:PER=*on*] [PRO:PER=*me*] [VER:pres=*avoir*] [VER:pper]:
on m' a privé ... [they have deprived me ...]
on m' a dérobé ... [they have robbed me ...]
on m' a volée ... [they have stolen from me ...]
on m' a pris ... [they have taken from me ...]

As this example shows, motifs are more than just POS-grams, because they underline the discursive unity of the dialogue of a character: a structure with verbs denoting movement, namely, removal or deprivation.

Motifs are automatically identified by the so-called unsupervised method, which means that the scholar did not look for something predetermined. They are thus different on this point from the syntactic sequences analyzed by S. Augendre, A. Kupść, G. Boyé and C. Mathon in this book. These syntactic sequences (e.g. proper names followed by a relative clause) – which of course are also non-discrete units – were manually annotated.

Four chapters base their work on sequential patterns:

E. Lapshinova-Koltunski and M. Zampieri focus on text classification techniques to discriminate methods and registers in translations per se and to identify their specific characteristics and relevant systemic differences in a single study. For this purpose they use linguistically motivated features representing texts, and more precisely, non-discrete ones, here sequential patterns which combine part-of-speech tags arranged in bigrams, trigrams, and 4-grams. The classification method used in this study is a Bayesian classifier with Laplace smoothing. The output of the classifiers is then used to carry out an extensive feature analysis on the main difference between genres and methods of translation.

F. Frontini, **M. Amine Boukhaled and J.-G. Ganascia** propose a new methodology for the study of characterization in French plays from a syntactic point of view, using the bottom up extraction of morpho-syntactic sequential patterns (again, "motifs" or extracted patterns of 3-4-5 grams of POS-tags, with at most one gap). The major and classical problem of the automatic extraction of motifs is the proliferation of patterns and the difficulty for humans to make sense of the huge number of resulting dimensions of variation between texts. To tackle this issue, the authors use a type of statistical analysis called Correspondence Analysis. They apply this method to the study of characterisation, namely to automatically find characterizing traits in the discourse of different theatrical characters by the same playwright. The corpus comprises the dialogues of Molière's most memorable protagonists.

D. Longrée and S. Mellet note that the paradigmatic approach provides little new information for the linguist or philologist, and, on the other hand, the syntagmatic approach does not really take the sequential structure of the text's

linearity into account. The authors therefore propose the new concept of *motif* in order to handle the different tokens of a given structure and to model them in a single pattern whose identification is based on its unified text dynamics function, disregarding surface variations. As a general pattern, the motif is able to characterize a genre; but its different realizations or tokens may be specific to different authors in a given genre. This claim is exemplified by a contrastive analysis of the style of two Latin historians, Caesar and Tacitus. The authors mine sets of characteristic motif tokens and observe their distribution along the text and their meaningful collocations. This leads to building a new type of paradigmatic list: the list of syntagmatic structures that characterize the grammar of an author's style, making it possible to go beyond the opposition between paradigmatic and syntagmatic approaches.

The study by **D. Legallois**, **T. Charnois and M. Larjavaara** focuses on literary stylistics. Their contribution illustrates both the complementarity between discrete units and non-discrete units, and between a stylistics of *identification* (based on stylometry techniques), and a stylistics of *characterization* (adopting a qualitative approach). They analyze 60 novels by twelve 19th century French novelists (Balzac, Dumas, Flaubert, Gaboriau, Hugo, Huysmans, Maupassant, Sand, Stendhal, Sue, Verne, Zola). The authors present in detail the methodology for extracting motifs (abstract lexico-grammatical patterns) that can be called characteristic of an author's style. Very often, these features have not been identified by traditional stylistics.

3.3 Production units and genre characterization

According to the process-oriented approaches to text genres, the non-discrete dimension of the units under consideration is fundamental, even more so because they consist of units of performance. These units are well suited for real time communication analysis; they contribute to a "*real-time grammar*" of genres.

Undoubtedly, these units are genre-sensitive, even if, in order to establish the description of a genre or a style, it is necessary to take other units into account. Two types of performance units are examined in this book.

3.3.1 Prosodic patterns

One is often struck, when one is in a room adjoining another one in which a program is being broadcast on radio or television, by the fact that it is quite possible to identify the message style or genre (sports commentary, weather

report, etc.), even when the sounds are not intelligible. This ability to identify styles on the basis of prosodic information alone was tested in the 1970s on French by the Hungarian I. Fónagy. Fónagy (1978) showed that intonation is genre-sensitive, with an oxytonic trend in conversation, and a barytonic trend in broadcast information. However, the relationship between prosody and genre is still rarely studied today. An exception is Obin et al. (2008) who showed that a small number of prosodic patterns are sufficient to discriminate discourse genres: the oral genres investigated by the authors (political discourse, radio news, radio interview, map task, life story) are linked to specific prosodic strategies in terms both of phonological structure and of acoustic features[1].

In this book, **S. Augendre, A. Kupść, G. Boyé and C. Mathon** investigate, together with syntactic structures, the weight of prosodic features (such as rhythm, pitch and intensity) in the identification of sports commentary as a genre. They distinguish two different kinds of sports commentary discourse in their corpus – simultaneous narration of moves and non-activity tied parts – and investigate the correlations between the game and the sports commentary both syntactically and from the prosodic point of view. They also conducted two perceptual studies. According to the authors, sports commentary is a constrained discourse genre that must be defined by taking into account not only thematic and medium dimensions as well as textual and stylistic dimensions but also the emotional level and the global context which influence the commentary. Exploring the oral dimension is necessary to define the genre.

3.3.2 Bursts

The remark by Halliday*: writing exists whereas speech happens* (Halliday 1985: xvii) is not entirely accurate although it is very often cited. Today's technical means enable the recording of acts of writing. It is thus possible to identify production units such as bursts. Bursts are continuous units that show the writing process in a real-life situation. They are also relevant units for examining discursive genres. More precisely, the term "bursts" of writing refers to strings of text that are produced without any major interruption during the writing process. In other words, bursts are segments of text that are produced between two consecutive pauses in a single writing episode. For example (from G. Cislaru and T. Olive):

1 See also Pršir, Goldman and Auchlin (2013) and Belialo, Lacheret and Kahane (2015).

un désir de partager du temps avec sa sœur et le petit ami de celle-ci
lit. a desire to spend time with her sister and her sister's boyfriend
Ainsi, il ramène des tableaux ou
lit. So he brought tables back or

are bursts. Bursts are identified by a computer tool that records all the actions that a writer performs when composing a text using a word processor and keyboard.

In view of the fact that some linguistic units are predetermined by the discursive genre, the question arises whether some of the bursts correspond to specific linguistic structures (for example, collocational frameworks or repeated segments) and whether there are linguistic regularities (prefabricated or not).

G. Cislaru and T. Olive analyze the writing and re-writing processes which jointly participate in the configuration of a written genre, for instance social reports on at-risk children, which are examples of professional writing that are both institutionally and socially constrained. They are "routinized discourse genres" as their form and content are pre-defined and their production is supposedly routinized and anonymized. In order to determine the way discourse and genre routines are processed and to detect specific or recurrent linguistic structures, they compare the content of the bursts of writing with the repeated segments (or n-grams) in the finished texts. The real-time data analyzed show that lexical strings that are repeatedly used in finished written discourse are generally not produced as blocks or bundles. More often than not, the social workers automatically produce specific non-routinized strings. This means that the strings usually called "prefabs" and considered as memorized formulae are not part of a discursive stock, but – at least partly – the result of adaptation strategies.

All these non-discrete units are therefore given special consideration in the contributions of this volume. Some have obviously long been known (e.g., reference chains), but have not been systematically compared from the perspective of genre analysis and stylistics. Others, on the contrary, are new in linguistics (e.g. motifs and bursts). The methods of extraction and interpretation still require further development, but they already enable unprecedented linguistic descriptions to be made.

4 Before starting to read

As should be clear by now, the present book provides new methods and new findings about the grammar of genres and styles – in some articles genres or styles, in some contributions both. Since Biber's early studies, multi-dimensional analyses have become increasingly common in many different fields such as linguistics, didactics, or automatic processing. At present, however, new analyses are emerging

that draw on these earlier studies, but also use new methodological solutions, test new methods, suggest new ways of seeing the linguistic variation between genres and styles and the ways in which belonging to a genre predetermines some of the linguistic choices.

The grammar of genres and styles: from discrete to non-discrete units is designed as a set of proposals, to be discussed, debated and criticized. We hope that the reader will find in these pilot studies the necessary inspiration to develop his/her own analyses.

References

Archer, Jodie and Matthew L. Jockers. 2016. *The Bestseller code: Anatomy of the blockbuster novel*. New York: St. Martin's Press.

Bakhtin, Mihail M. 1986. *Speech genres and other late essays*. Translated by Vern W. McGee. Austin: University of Texas Press.

Belialo, Julie, Anne Lacheret and Sylvain Kahane. 2015. Marqueurs intonosyntaxiques en français parlé et genres : compter pourquoi, compter quoi, compter comment? *Langages* 197, 129–152.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, Douglas and Susan Conrad. 2009. *Register, genre, and style*. Cambridge University Press.

Biber, Douglas, Ulla Connor and Thomas A. Upton. 2007. *Discourse on the move: Using corpus analysis to describe discourse structure*. Amsterdam-Philadelphia: John Benjamins.

Biber, Douglas and Edward Finegan, (eds).1994. *Sociolinguistic perspectives on register*. New York: Oxford University Press.

Biber, Douglas and Bethany Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9:1. 2–20.

Chastain, Charles. 1975. Reference and context. In Keith Gunderson (ed.), *Language, mind, and knowledge*, 194–269. Minneapolis: University of Minnesota Press.

Dorgeloh, Heidrun and Anja Wanner (eds.). 2010. *Syntactic variation and genre*. Berlin-New York: De Gruyter Mouton.

Fang, Chengyu Alex and Jing Cao. 2015. *Text genres and registers: The computation of linguistic features*. Springer.

Fónagy, Ivan. 1978. A new method of investigating the perception of prosodic features. *Language and Speech* 21: 34–49.

Halliday, Michael A. K. 1985. *An Introduction to Functional Grammar*. London: Edward Arnold.

Halliday, Michael A. K. and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Jockers, Matthew L. 2013. *Macroanalysis: Digital methods and literary history*. Urbana (Illinois): University of Illinois Press.

Kerz, Elma. 2012. The role of genre in information structuring in English. *Belgian Journal of Linguistics* 26: 143–159.

Knapp, Peter and Megan Watkins. 2005. *Genre, Text, Grammar: Technologies for Teaching and Assessing Writing*. University of New South Wales Press Ltd.

Lefer, Marie-Aude and Svetlana Vogeleer (eds.). 2016. Genre- and register-related discourse features in contrast. Special issue of *Languages in contrast* 14:1.

Léon, Jacqueline and Sylvain Loiseau. 2016. *History of Quantitative Linguistics in France*. RAM Verlag.

Mann, William and Sandra Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8 (3). 243–281.

Martin, James R. 2010. Foreword. In Mehler, Alexander, Serge Sharoff and Marina Santini (eds.), *Genres on the web: Computational models and empirical studies*. Dordrecht: Springer.

Moretti, Franco. 2013. *Distant reading*. London: Verso.

Neumann, Stella. 2014. *Contrastive register variation: A quantitative approach to the comparison of English and German*. Berlin: De Gruyter Mouton.

Oakes, Michael P. 2014. *Literary Detective Work on the Computer*. Amsterdam: John Benjamins.

Obin, Nicolas, Anne Lacheret-Dujour, Christophe Veaux, Xavier Rodet and Anne-Catherine Simon. 2008. A Method for automatic and dynamic estimation of discourse genre typology with prosodic features. *Proceedings of Interspeech 2008, Brisbane*. 1204–1207.

Pennebaker, James W. 2011. *The secret life of pronouns: What our words say about us*. New York: Bloomsbury Publishing.

Pršir, Tea, Jean-Philippe Goldman and Antoine Auchlin. (2013), Variation prosodique situationnelle : étude sur corpus de huit phonogenres en français, in Mertens, P. and A. C. Simon (eds.), *Proceedings of the Prosody-Discourse Interface Conference 2013 (IDP-2013)*. Leuven, September 11–13, 2013, 107–112 [http://wwwling.arts.kuleuven.be/franitalco/idp2013/Proceedings.html].

Quiniou, Solen, Peggy Cellier, Thierry Charnois and Dominique Legallois. 2012. What about sequential data mining techniques to identify linguistic patterns for stylistics? *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent text Processing*, Mar. 11–17, CICLing, New Delhi, India. 166–177.

Reinhart, Tanya. 1982. *Pragmatics and linguistics: An analysis of sentence topics*. Bloomington, IN: Indiana University Linguistics Club.

Swales, John M. 1990. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swanson, Wendy. 2003. *Modes of Co-reference as Indicator of Genre*. Bern: Peter Lang.

Ana Elina Martínez-Insua and Javier Pérez-Guerra

Text types, audience and thematic organisation in the recent history of English

Abstract: This study is couched within a larger project which deals with the relation between the target audience of a given genre (or a text type within a genre) and the thematic choices made by the writer, Theme being conceived as in Berry (1995, 2013a). In particular, we have focused on just one part of the Theme: the syntactic Subject or Subject Theme and, in particular, its meaning, in connection with Berry's (2013a) distinction between contentful and contentlight Subject Themes and her hypothesis that most Subject Themes are contentful in formal written texts, while most of them are contentlight in informal spoken texts. This chapter also tests to what extent the target audience may be a factor affecting the category and the content weight of the thematic position or what the Subject Themes refer to. We have taken data from two genres: medical texts (from the electronic corpus of Modern English Medical Texts) and newspaper discourse (from the Zurich English Newspaper corpus), in Early and Late Modern English. The data reveal, first, that 'learnedness' constitutes the distinctive factor that correlates with the distribution of the thematic content and with the thematic reference in the Subject Themes in the medical texts. Second, textual formality has proved to be the distinctive factor that correlates with thematic content in the news. Third, style has been the distinctive factor that correlates with thematic reference in the Subject Themes in the medical texts.

1 Introduction

This study takes the initial assumption that the grammar of texts can be approached not only through an investigation of the frequency of discrete units which serve as indices of a broader functional characterisation, but also by

Note: We are grateful to the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (grants no. FFI2013-44065-P and FFI2016-77018-P) for generous financial support. We gratefully acknowledge Margaret Berry's most generous help on many aspects of this study.

Ana Elina Martínez-Insua and Javier Pérez-Guerra, University of Vigo

https://doi.org/10.1515/9783110595864-002

considering larger, non-discrete open sets of segments and their relevance to the informative, referential content and target audience of those texts. More specifically, in this chapter we investigate the way in which sentences are initiated in the textual discourse by focusing on the content (see Section 4.2 for the multifaceted concept of 'content' to be used here) of the syntactic subjects in the textual samples analysed. The ultimate goal is to address the thematic organisation of texts in Modern English (ModE), as part of a larger project on the degree of textual variation in the recent history of the English language in terms of the organisation of clausal constituents.

We follow Matthiessen (2015: 5) when he argues that "a **text** is language functioning in a situation" and that "[a]s we move further up in the cline of instantiation, we can group texts that are alike in crucial aspects into **text types**, characterizing a text type as language functioning in a situation type" (author's emphasis). We will explore the connection between the linguistic (structural and informative) organisation of the clause in two text types and the consequences which such a design has for the actual function of the texts in their corresponding situation types. More specifically, this study aims to analyse the thematic organisation of two genres or text types in Early and Late Modern English (eModE and lModE, respectively), namely, medical and newspaper texts. Variation in linguistic usage and variation in the context of the communicative event have been claimed to go hand in hand (Hymes 1974; Biber 1988; Biber and Conrad 2009), and this study is an attempt to find possible specific correlations (i) between content weight of Subject Themes in medical texts and target audience, and (ii) between content weight of Subject Themes in newspaper texts and textual formality. We hypothesise that texts addressed to learned audiences should have contentful Themes, and that clauses in texts targeted to other audiences should contain 'contentlight' Subject Themes to a greater extent (see Section 4.2 for an explanation of this concept).

2 Historical background

The period covered in this study is characterised by major sociocultural changes in England. At a general level, the English language is not only affected by the social context but also by the increasing literacy and the development of printing.

Regarding the domain of medicine, the period 1500–1700 witnessed the so-called "third phase of vernacularization" (Taavitsainen 2010a: 38), in which Latin gives way to English, and English finally becomes the language of science. As described by Pahta and Taavitsainen (2010: 5), by 1500 "English had emerged

in medical writing from the shadow of Latin in the first wave of a pan-European process of vernacularization in science and medicine that began in non-institutional contexts". The language of medicine was "a forerunner for other types of Fachprosa in the vernacular" (Taavitsainen 2001: 189), which may well have led it to become one of the particular points of focus in the linguistic standardisation that took place in the early period. By 1700, this process of vernacularisation had reached a stage where English had already challenged Latin as the language of institutional medicine; it "had become the dominant language of medical writing in England, and was used as the original medium for communicating new scientific". The evolution of English medical texts, however, has to be seen against the standard of Greco-Roman texts, given the large number of translations and adaptations from Latin or French sources (Taavitsainen 2001: 193).

Following the appearance of print technology in England towards the end of the fifteenth century, more heterogeneous audiences began to have access to medical texts. Such writing became more widely available and accessible to the public, "including increasingly literate non-professional readers" (Pahta and Taavitsainen 2010: 5). This, together with the changing spectrum of known diseases, the introduction of new substances from the New World, and the development of physical and chemical methods of treatment, led to a transition from the thought styles of the earlier periods to more modern approaches to medicine. The natural world was seen in a very different light by the end of the seventeenth century than it had been two centuries earlier (Taavitsainen 2010b: 12). In brief, the human being was no longer regarded as the centrepiece of the universe, and considerable advances in the knowledge of anatomy and physiology had been made, all of which led to significant changes in the linguistic practices in the medical writing of the Early Modern period (Taavitsainen 2010a: 30). The Modern periods thus offer a fascinating picture of the language in terms of the connection between situational circumstances, as illustrated by the taxonomy of target audiences in the case study reported in Section 5.1, and the shaping of texts, here investigated through the thematic design of clauses, that is, of the clausal constituents acting as sentence openers.

Turning to newspapers, the eModE and lModE periods witnessed major changes affecting this text type, which at the time was heterogeneous in shape, audience and function. First, as Brownlees (2006: 8) has claimed, "[n]ews discourse was as heterogeneous in presentational format, function, linguistic features and content in early modern Britain as it is in the modern era. It is very difficult if not impossible to encapsulate the characteristics of seventeenth and eighteenth century news discourse within tidy, easily definable categories". In 1620 the 'corantos' consisted of small folios in which news reporting was very factual, with a flat, impersonal style devoid of editorial comment, typically about events occurring on continental

Europe and often translated from German publications that had been released a matter of days previously. Only a few years later, corantos ceased to be mere translations and gained their own editorial voice; in Brownlees' (2005: 69–76) words, "[a] new tenor of discourse was established between news writer and reader which led to a more familiar, oral mode of discourse". As Bös (2012) has summarised, by the early nineteenth century there was competition between the 'respectable' daily press (*The Times*, *The Observer*, etc.) and the so-called 'radicals' or independent newspapers (*Political register*, *Black Dwarf*, *Poor Man's Guardian*, etc.), that is, publications which employed working class voices and class-conscious language. The cost of printing technology was such that a weekly alternative now emerged, often published on Sundays (newspapers such as *British Gazette and Sunday Monitor*, *Bell's Life in London*, *News of the World*, *Lloyd's Weekly Newspaper*), and typically a synthesis of respectable and more radical newspapers for those readers who could not afford a daily newspaper. These new titles, with an interest in gossip and sensationalism, came to overpower radical publications in the marketplace. It would not be until the beginning of the 20th century that a 'new journalism' would be identified with the tabloidisation and popularisation of news content in newspapers such as *The Daily Mail* and *The Daily Mirror*. The selection of the newspaper genre in this study is justified precisely by the parallelism between, on the one hand, the type of newspaper (daily vs. weekly) and the type of news (see Section 4.1), and, on the other hand, the target audience of the texts. This connection will be analysed through an investigation of the thematic design of clauses in daily/weekly newspapers and the different types of news, in an attempt to establish some sort of correlation between linguistic shaping and the situational circumstances of the text types in question.

In sum, we are concerned here with the link between text types, that is, medical texts and (printed) news, in a period in which these genres were under development both in their external format and regarding their intended readerships. As noted above, we will approach this form/function relationship by focusing on the way in which the design of sentences (or clauses) reveals the functional condition of texts, and hence, of their audience.

3 Theoretical assumptions

In this section we will summarise the guiding theoretical principles of the study. In Section 3.1 we focus on the consequences of the social context for the linguistic realisation of discourse, and the advantages of the theoretical framework adopted are discussed in Section 3.2.

3.1 Relevance of context as a factor informing the users' decisions. Context and choice

Accepting as a point of departure the postulate that the context (of culture and discourse) "plays a significant role in determining the actual choices" made by users (Halliday 2009: 55), we assume that factors such as the context, the participants and their roles, their social, regional and cultural background, or the purpose of the communicative act are all capable of causing variation in the linguistic choices made by writers and speakers (see, among others, Hymes 1974; Biber 1988; Biber and Conrad 2009). Together with this context-embeddedness of real language (Thompson 2004: 11), a certain degree of freedom in the user is also taken for granted, in that it is assumed that whenever one wants to make linguistic choices appropriate to the context, one has to make decisions, either conscious or unconscious, about the current context (Berry 2013b: 375). As Berry (2013a: 245) observes, "if language users are to choose appropriately from the semantic options available, they need to make decisions about the nature of the situations in which they find themselves".

We aim here to test the premise that linguistic use accommodates to variations in the context of the communicative event in which a text is produced. Consequently, the study is designed as an attempt to verify qualitatively and quantitatively the interconnection between tenor[1] (in this case, the target audience of the texts) and intra-genre variation in the weight and reference of the Subject Themes. In essence, we seek to associate both the variationist assumption that variability in language may be structured (Milroy and Milroy 1997: 47), and the Hallidayan (Halliday 1985, 2014; Halliday and Matthiessen 2004) functional multi-tiered perspective on English clauses.

As is customary in historical variationist studies, since the connection between language change and language-external factors is assumed (Weinreich et al. 1968: 188), in addition to the relevance of the target audience of texts and the periodicity of the publications, this study acknowledges the influence of other factors over time (see Section 2). As Pahta and Taavitsainen (2010: 3) argue with reference to medical texts,

1 In Martin and Rose's (2008: 11) words, "[t]enor refers to who is taking part, to the nature of the participants, their statuses and role: what kind of relationship obtain (…) and the whole cluster of socially significant relationships in which they are involved". In this respect, see Matthiessen (2015: 43) for a study of fields of activity which represent, among other issues, variation in the medical text type from an expounding ("reference material, in print (…) with general information about the human body") to an exploring ("[reference material, in print with] advice about how to deal with the health problem") primary type, variation according to the field of activity, and also according to the tenor (and texts directed to a professional public or to the general public).

> [w]ith changes in the scientific paradigm, i.e. in the ways of scientific thinking and in the ways of doing science, the ways of communicating science also change [...] Like any other ideology, scientific thought-styles are mediated to us through language and connected with particular ways of using language, which also construct those thought-styles

(see also Taavitsainen and Pahta 1995; Taavitsainen et al. 2002; Pahta and Taavitsainen 2011). Turning to the news text type, in Section 2 we mentioned the evolution from translated corantos to new editorial products, which as Brownlees (2005: 69–76) has pointed out gained "editorial voice and an audience in the sense that the editor now began to address the readership directly. A new tenor of discourse was established between news writer and reader which led to a more familiar, oral mode of discourse".

3.2 The choice of the theoretical framework

The major role of linguistic choice as the manifestation of variation, specifically at the sentence level and compromising situational or contextual differences, has led us here to adopt as a theoretical framework Systemic Functional Linguistics (SFL), originally developed by Michael Halliday in the late 1950s and the 1960s. SFL focuses on the functions of language, assuming that language operates as a system of human communication and users construe meanings by making choices from linguistic systems. Such choices are, in turn, conditioned by the genre (or text type) to which texts belong, and the social contexts in which they are produced. From this perspective, language is seen to be organised as a system of options, and enables speakers to create meaning by selecting relevant options from the system (Fontaine 2013: 5). The principles and assumptions underlying the term Systemic Functional (Montemayor-Borsinger 2009: 79) include the structure of language being seen as "the outward form taken by systemic choices" (Halliday and Matthiessen 2004: 23), and the consideration of function as the "driving force" in language use (Fontaine 2013: 5).

A few words on the framework seem in order at this point, especially about textual metafunction, which will be the focus of the current analysis. From a functional perspective, communication is a very important function of language, but it is not the only one. Clauses, as instances of language, are multifunctional units (Fontaine 2013: 9) and their various meanings relate to the so-called metafunctions of language. These include (i) the function of language which consists in expressing content (the 'experiential metafunction'), (ii) the expression of the speaker's own involvement or attitude towards the message, as well as the establishment of given interrelations between speaker and hearer (the 'interpersonal

metafunction'), and (iii) the function which is responsible for enabling speakers and writers to create texts, that is, coherent passages of language in use, whether spoken or written, which constitute connected and contextualized pieces of discourse (the 'textual metafunction').

Ordering within the clause itself has meaning (Thompson 2004: 7), and if we take meaning as the sum of what the speaker/writer wants the hearer/reader to understand, then understanding how a given message fits into its given context is clearly part of the meaning of such a message. In general, SFL "sets out to investigate what the range of relevant choices are, both in the kinds of meanings that we might want to express (or functions that we might want to perform) and in the kinds of wordings that we can use to express these meanings; and to match these two sets of choices" (Thompson 2004: 8). This being so, when approaching language from a functional perspective, one has to look outwards at the context (see Section 2) and, at the same time, to identify the linguistic options (i.e. lexical and structural possibilities offered by the language system) and to explore the meanings that each option expresses. It may be said that what a functional analysis aims to uncover are the reasons that lead the user to produce "a particular wording rather than any other in a particular context" (Thompson 2004: 9). The textual structure of language is fundamental to the creation of a text, as it allows the user to distribute information in the clause. This component permits users to control the part of the language system that enables them to interact with their interlocutors, and to structure what they are saying in such a way as to transmit the message successfully.

One of the structures that falls within the textual component and helps to create texture is the thematic structure, which constitutes the object of this study. As will be explained below, when analysing the thematic structure of a clause, one identifies the Theme of each clause, that is, the point of departure for each message. The thematic organisation of a clause reveals and expresses how the text develops and, by analysing the thematic structure of a text clause by clause, one can gain insights into its texture and understand how writers/speakers made clear to their addressees the nature of their underlying concerns or communicative purposes (see Halliday 1985: 67).

Even though the Systemic Functional framework has been applied mainly to the description of contemporary English over the last fifty years, our aims here are to contribute to the characterisation of Modern English text types. This chapter is thus inspired by Cummings' (1995) study of the thematic structure of Old English and his claim that it is natural to wonder whether functional analyses of the thematic area of the clause not only illuminate the structure of real texts in recent historical dialects, but may also be applied usefully to texts at all stages of historical development (Cummings 1995: 275).

4 The data and the variables

In this section we will describe the corpora and the data (Section 4.1) and the research variables used (Section 4.2). The null hypothesis in the analysis of both medical and newspaper texts is that linguistic form, here focussing on the category and content of Subject Themes, is not conditioned by either the audience (medical texts) or the audience and topic (newspapers) of the texts.

4.1 The corpora

The medical texts were selected from the second and third components of the three-part *Corpus of Early English Medical Writing*, which covers the period 1375–1800 (see Pahta and Taavitsainen 2010: 1–2). The compilers' aim was to collect evidence of "stylistic change in medical English in a long diachronic perspective in a multifaceted sociohistorical framework" (Pahta and Taavitsainen 2010: 2), the general framework being variationist. Hence, the sample selected for analysis contains extracts from the corpus of *Early Modern English Medical Texts* (EMEMT) (1500–1700) and the corpus of *Late Modern English Medical Texts* (LMEMT) (1700–1800). The newspaper texts analysed are drawn from the *ZEN Zurich English Newspaper Corpus*, a corpus of 1.6 million words containing newspapers published between 1661 and 1791 (Lehmann, Keller, and Ruef 2006).

In selecting the samples, we have avoided translations, particularly frequent in EMEMT, mostly from Latin and French, and in the first periods of ZEN from German (see Section 2), since the conventions of (un)marked themes may vary in different languages and the translator may be unaware of the trends affecting thematic progression and organisation in the source language, leading to their preservation in translation.

The study is based on extracts amounting to approximately 2,000 clauses, with texts dating from the 1550s, 1650s and 1690s (EMEMT), the 1760s, 1770s and 1780s (LMEMT), and from the eighteenth and nineteenth centuries (ZEN). The medical texts are organised in six textual categories in EMEMT and LMEMT (see Taavitsainen and Pahta 2010; Taavitsainen et al. 2014), of which we have focused on:

- General treatises and textbooks (category 1)
- Treatises on specific diseases (category 2)
- Regimens and health guides (category 4).

The other textual categories in both EMEMT and LMEMT, including Recipe collections (category 3), Surgical treatises (category 5) and Philosophical Transactions of the Royal Society of London (category 6), are either more specific or addressed

to very specific audiences, and as such have not been included. The newspaper material in ZEN has been classified following the textual categories recognised in studies such as Fries (2001, 2009) and Bös (2012):

– type of newspaper: 'respectable' versus 'Sunday/weekly' papers
– type of news: 'hard' news (politics, economics, diplomacy, appointments, commercial information, etc.), with a neutral, formal and distant style, versus 'soft' news (births, marriages, diseases, deaths, crimes, court trials, accidents, etc.) in an involved, personal and colloquial style.

We have analysed samples assigned to the following combinations of categories: hard news in weekly papers, soft news in weekly papers, hard news in daily papers and soft news in daily papers.

4.2 The variables: Theme and content weight

As we have already noted, this study pays special attention to the meanings conveyed by Themes, hypothesising that they may be particularly significant for each text type, and assuming that they relate, ultimately, to their broader socio-cultural context (Forey and Thompson 2008: 3). However, we must also acknowledge the ongoing debate concerning the boundaries and interpretation of Theme itself. Taking as a basis the Hallidayan consideration of the concept, a great deal of work has been done on its identification in texts, looking at which factors are involved in deciding what to include in Theme, and/or proposing specific guidelines for establishing boundaries between Theme and Rheme. For our purposes, we have adopted Berry's (2013a, 2013b, among others) working definition, according to which "the Theme of a clause is everything up to the main verb of the clause" (2013a: 248).[2] Following Berry herself, we concentrate here on just one part of the Theme: the syntactic Subject or Subject Theme (SubjTh).[3] In an attempt to replicate Berry's (2013a) approach to 'contentlight' and 'contentful' Subject Themes in very different types of texts, only Themes of independent clauses were considered for inclusion in the database, in that the main contribution here comes from the thematic structure of independent clauses (see also Halliday 1985, and Brown

2 For the controversy on the notion of Theme in SFL, see also Berry (1995) and Downing (1991), among others. Other views on the extent of the Theme can be found in the literature (for example, Berry 1996: 29–31) but a detailed discussion would go beyond current space constraints.

3 The label 'Subject Theme' is a term that Berry (2013a: 248) takes from Fawcett (2008: 182) to refer to what the former had previously labelled 'Basic Theme'. Her label 'Subject Theme', or simply SubjTh, will be adopted here.

and Yule 1983). More specifically, only declarative main clauses were explored, which constitute the vast majority of the clauses contained in our texts, given their explanatory, descriptive, formative nature (treatises, guides, etc.).

Also following Berry, we will discuss "content weight in terms of meanings, in terms of what the Subject Themes refer to" (Berry 2013a: 249). Weight is seen as the amount of meaning or reference. The lesser the meaning of a Subject (for example, of a pronominal form), the lighter it is regarded to be in terms of content. Conversely, the greater the amount of semantic content of the Subject (a noun or a nominal group), the more contentful it is taken to be. Besides content weight, the informational reference of SubjTh has also been taken into consideration. Thus, as in Berry (2013a: 252), (i) SubjTh consisting of personal pronouns with antecedents are assumed to refer to 'Given Topics' (GivTop); (ii) SubjTh consisting of a noun/nominal group that refers to an aspect of a discourse topic previously mentioned are assumed to refer to 'Resumed Topics' (ResTop), a label used for references to topics that have already been introduced some time ago and which therefore require reactivation; and (iii) SubjTh referring to aspects of the discourse topic which have not previously been introduced are assumed to refer to 'New Topics' (NewTop). As explained by Berry, this classification draws on Dik's (1997) distinction between New Topic (NewTop), Given Topic (GivTop) and Resumed Topic (ResTop), and is reinforced by the adoption of labels related to content weight: while SubjTh referring to GivTops can be regarded as 'content-light', those referring to ResTops and NewTops are labelled as 'contentful' (Berry 2013a: 258). In addition to these three informational categories, a fourth, 'Qualified Resumed Topic' (QResTop), will be used in the analysis of the corpus[4] with SubjTh that refer to aspects of the discourse topics which have been mentioned before and contain some kind of qualification bringing semantic nuances to the resumption of the topic (Berry p.c.). The SubjTh in (1) below, from *Method and means* (1683), illustrate the four categories (SubjTh are italicised and the tags qualify only the subject in such Themes):

(1) 1. *If heat exceeds; the natural moisture*$_{\text{NEW}}$ dries up, *the spirits*$_{\text{NEW}}$ evaporate, and *the body*$_{\text{NEW}}$ withers. *If cold; the faculties*$_{\text{NEW}}$ are torpid and benumbed, the spirits being frozen up to a cessation from their duties.
2. *If moisture prevails; the spirits*$_{\text{RES}}$ *are* clogged, suffocated and drowned in the channels of the body.

4 Berry brought to our attention the possibility of including this fourth category. Brown and Yule (1983: 174) refer to expressions of the form *the* + *property* + *noun* and the fact that they are "used almost exclusively in identifying displaced entities [that is, those labelled as *resumed* here]".

3. If siccity and dryness; the organical parts are stubborn, unpliable and incapable of their regular motions and due actions; the vital streams being drunk up that should irrigate, refresh, and supple them.
4. *Were the body*$_{\text{RES}}$ always taking in and sending nothing forth, *it*$_{\text{GIV}}$ *would* either increase to a monstrous and vast magnitude; or fill up, suffocate and stifle the soul: were it always in excretion and emission, the body would waste away and be reduced to nothing.
5. Nor is the receiving in of any thing, sufficient and satisfactory to the body for its preservation; but that which is appointed by Nature, proper and suitable: nor emission or ejection of any thing, but that which is superfluous and unnecessary to be retained.
6. *If Sleep prevails contrary to the Law of Nature; the body, in a lethargic soporiferous inactivity, stupefied and senseless*$_{\text{QRES}}$, lies at the gates of death.
7. *If Watching exceeds the limits, transgresses and steals away the due time for sleep; the faculties*$_{\text{RES}}$ *are* debilitated and enervated, the spirits tired, worn out, and impoverished.

The natural moisture, *the spirits*, *the body* and *the faculties* in the first paragraph are all NewTops, referring to aspects of the discourse topic mentioned for the first time. *The spirits* in the second paragraph, *the faculties* in paragraph 7, and *the body* in paragraph 4 all refer to ResTops, as they have been previously mentioned in the discourse before the introduction of other entities. It is true that, although there is not a great distance between the first and second mentions of these topics, distance is not the only relevant matter here, but rather the very existence or not of other entities between the two mentions. Brown and Yule's (1983: 173) alternative term for (Berry's) ResTop, 'displaced given entity', is helpful here in that it neatly encapsulates the fact that an entity can only be regarded as a GivTop if it has not been displaced by another entity (Berry p.c.). In *Were the body always taking in and sending nothing forth, it would either…*, the body is alluded to twice, by the nominal group *the body* and by the pronoun *it*. This second mention, the third person pronoun *it*, is thus a GivTop because no other entity has occurred between it and the first reference that displaces it from the GivTop status. Indeed, as already pointed out, personal pronouns are by default GivTop. The first reference in this sentence cannot be regarded as a GivTop, however, because although *the body* has been mentioned previously in paragraph 1, several other entities have been mentioned between paragraph 1 and paragraph 4 to displace *the body* from the GivTop status. This itself explains why it has to be referred to by a noun/nominal group and not by a pronoun, which would be ambiguous. Similarly, even though the mentions of *the spirits* in paragraphs 1 and 2 are quite close together, other entities are introduced between them, thus displacing them from the GivTop

status, and as a consequence they have to be regarded as ResTops and referred to by nouns/nominal groups and not by pronouns, which once again would lead to ambiguity. The case of *the body* in paragraph 6 is slightly different. The fact that it is qualified (*in a lethargic soporiferous inactivity, stupefied and senseless*) means that it is giving us a new view of the body, and as such has to be regarded as a QResTop.

Let us now consider some examples from the news database:

(2) 1. <essay s 3.2>*Edgar, sir-named the Peaceable*$_{NEW}$, was crowned at Kingston by Otho, Archbishop of Canterbury.
2. <essay s 3.3>*To rid the Land of Wolves, which then were very plenty, instead of the Tribute imposed on the Prince of Wales by King Athelstan, he*$_{GIV}$ appointed Luduall Prince of Wales, to pay yearly 300 Wolves.
3. <essay s 3.4>*His Navy Royal, consisting of 3600 Ships, he*$_{GIV}$ employ'd in securing the Coasts of Pirates and foreign Enemies, wherein himself would sail every Summer.
4. <essay s 3.5>*And in the Winter he*$_{GIV}$ would circuit the Country, taking an Account of the Administration of his Laws, and a Demeanour of his great Men, especially his Judges, whom he would punish severely, if he found them to have been guilty of Bribery, or Partiality, insomuch that there was never less Robbery, Deceit or Oppression than in the Reign of this King.

In (2) *Edgar, sir-named the Peaceable*, the person dealt with in the essay, is named only in paragraph 1, where his name occupies the thematic position. The following paragraphs use *he*, a clear example of a contentlight given SubjTh.

(3) 1. <letter s 5.1>*In the Interim the good old Lady Bank, who had on many Occasions serv'd her Country, by opening her Coffers in Times of Distress, and had many honest and faithful Servants about her*$_{NEW}$, was now neglected, and her Rival, by large Bribes and Presents, ran away with the Prize, tho' she had bid what the Thing wou'd honestly bear, provided it was faithfully and honestly manag'd, and to the Interest and Advantage of those concern'd.
2. <letter s 6.1>*Lady South-Sea*$_{NEW}$ gilded her Pretensions over with a Shew of Zeal for the publick Good, by her Readiness to reduce the publick Debts.
3. <letter s 6.2>*Lady Bank*$_{RES}$, without any of those specious Shews, made a plain honest Proposal; but her Probity did her no Service, she was fain peaceably to sit still and see her Adversary flourish.

4. <letter s 6.3>Her whole Family was in deep Mourning, and look'd all very much dejected.
(…)
5. <letter s 8.1>*In the Interim Lady South-Sea*$_{RES}$ improv'd ev'ry Day in Splendor and Magnificence: If you ask'd how the Stock went? the Servants answer'd, with an insulting Joy, Money for the Refusal at a 1000; and with a little good Management, they did not question but to blow up their Stock as high as their Sister Lady Missisippi in France, and to see the useless Metals of Gold and Silver abolish'd, and to establish Paper in the room; which would save the Labour of counting Money.

Unlike in the essay in (2), the letter in (3) does not focus on a single person, but rather there are two ladies that most frequently occupy the position of SubjTh, Lady Bank and her rival Lady South-Sea. Once they have both been named (Lady Bank in paragraph 1, and Lady South-Sea in paragraph 2), using a personal pronoun to refer to either of them would be ambiguous. Hence, resuming their full names, in paragraphs 3 and 5, respectively, disambiguates references to the SubjTh in these paragraphs.

(4) 1. <foreign.news s 89.2>On the 31st at Eleven in the Morning we saw a Sail, which came out of the Harbour of St. Sebastian, we immediately gave her Chace, and at four in the Afternoon came up with her and took her: *She*$_{GIV}$ proved to be a Privateer called the Biscaia belonging to St. Sebastian, mounting ten Carriage and two Swivel Guns, had on board 150 Pistols, 19 Blunderbusses, 140 Muskets, 166 Cutlasses, 20 Pikes, and a great Number of Powder-Flasks, Hand Granades and Pole-Axes; *she*$_{GIV}$ had but 119 Men on board, tho' by the Roll of the Ship's Company she should have had 140, so that it is presumed the rest were killed in the Action and thrown over-board: *This Privateer*$_{QRES}$ came out of St. Sebastian at Five that Morning.
2. <foreign.news s 89.3>The Engagement was very smart on both Sides, we having been obliged to fire 31 Chace Guns with Round and Partridge Shot, and a great Number of small Arms at her, before she struck: Nay, after the Captain, who was a Frenchman, had struck his Colours, the two Lieutenants, who were Dutchmen, hoisted them, and continued the Fight 'till we obliged them to strike.
3. <foreign.news s 89.4>*The above Privateer*$_{QRES}$ has taken 23 Prizes since the Beginning of the War.

Example (4) illustrates the qualified resumption of entities already mentioned in the discourse as an alternative to the use of personal pronouns. In this case,

the ship that constitutes the centre of attention in this item of foreign news is first introduced as a NewTop in the Rheme of the initial clause (*On the 31st at Eleven in the Morning we saw a Sail*), and further characterised as a privateer in the Rheme of a later clause (*She*$_{GIV}$ *proved to be a Privateer called*...). In subsequent clauses, the entity is mentioned on at least three other occasions. The use of QResTop to refer to the ship helps the writer to resume an entity whose first mention is now quite distant, and to avoid the ambiguity that might arise from the use of personal pronouns once other entities have been introduced in the discourse.

Taking this classification into account, for the purposes of somehow correlating content weight and informational reference of the Themes, a cline of contentfulness seems more plausible than a clear-cut distinction between contentlight and contentful. The cline goes from the contentlight GivTop towards the increasingly more contentful ResTop, QResTop and NewTop, as Figure 1 below represents.

Importantly, according to Berry's (2013a: 264) findings, there seems to be a strong tendency for SubjTh in informal spoken English to realize contentlight options and for those of formal written English to realize contentful options. Bearing this in mind, and taking into account the expected different degrees of formality of the texts in our sample, the hypotheses to be tested in the present study are:

(i) SubjTh in medical texts addressed to unlearned audiences should be contentlight, while texts addressed to learned audiences should contain contentful SubjTh more frequently.
(ii) As regards the degree of referentiality conveyed by SubjTh, the medical texts targeted to learned audiences should contain more QResTops and ResTops (nouns and noun phrases).
(iii) SubjTh in newspapers addressed to learned audiences ('respectable' newspapers) should be contentful, while newspapers addressed to other audiences should contain contentlight SubjTh more frequently.
(iv) SubjTh in the so-called 'hard' news should be more contentful than those in the so-called 'soft' news.
(v) The degree of referentiality of SubjTh in news reports should also correlate with the target audience and/or the textual formality of the texts.

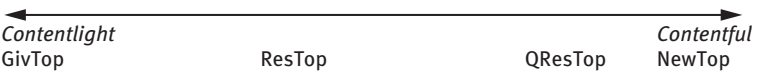


Figure 1: Contentfulness scale.

5 Case studies

The broad research question which will be addressed here is: do variation in linguistic usage and variation in the context of the communicative event go hand in hand? As already pointed out, we will address this by focussing on the potential association between both content weight and informational reference in SubjTh, and by, first, target audience in the medical texts (case study I in Section 5.1) and, second, textual formality in newspapers (case study II in Section 5.2).

5.1 Case study I: Medical texts

In this section we deal with the research question 'do, on the one hand, content weight and informational reference of SubjTh, and, on the other hand, target audience, correlate in medical texts?' The target audiences of the texts in EMEMT and LMEMT include "both specialists and lay audiences" for whom physicians, surgeons and other learned professionals tended to write (Taavitsainen 2010a: 34). In this vein, we have considered the following target audiences:

- learned audience, prototypically of the textual category 2 (Treatises on specific diseases). As pointed out in Pahta and Ratia (2010: 74), the readers of these texts range "from strictly academic specialists to the widest general readership", just as the authors of the texts themselves may "come from different social and educational backgrounds". In this study we have concentrated on texts addressed to academic specialists.
- unlearned audience, as the prototypical one of the texts in category 1 (General treatises and textbooks). The readers are lay people of the middling classes, including "the illiterate who may hear it [the book] read aloud or explained" (Taavitsainen and Tyrkkö 2010: 70).
- intermediate audience, which in the textual category 4 (Regimens and health guides) comprises "readers (…) not assumed to have specialised medical knowledge, though references to authorities or Latin terms are not completely absent (…) The buyers (…) were literate people with enough money to spend on ensuring their health by following the advice in the books" (Suhr 2010: 117).[5]

5 The texts which have served as the empirical basis of this study are:
– Against the sweatyng sicknesse (1550), Gutta Podagrica (1633), Little Venus unmask'ed (1670, 2nd ed.) and An inquiry into the nature, cause, and cure of the croup (1765) in category 2 (learned audience)

Table 1: Clause count in medical texts.

	learned	intermediate	unlearned	Total
16thc	105	101	101	307
17thc	232	231	232	695
18thc	121	121	121	363
Total	458	453	454	1,365

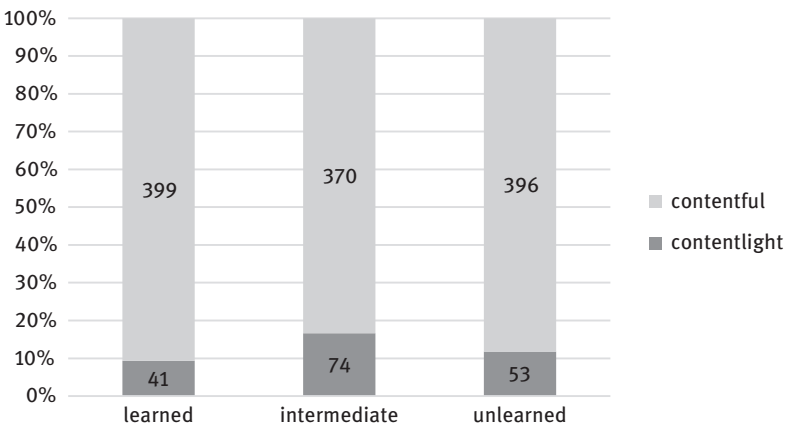


Figure 2: Content weight of SubjTh in medical texts.

Table 1 sets out the number of clauses registered in our database per period and textual category, totalling 1,365. Figure 2 plots the percentages of contentful and contentlight SubjTh in the three textual categories under investigation.

The data in Figure 2 reveal, first, that most SubjTh are contentful in the three audience categories, since all the texts contain written specialised texts. Second, the differences in the distribution of contentful and contentlight SubjTh between the intermediate and the unlearned categories are not statistically significant ($\chi^2(1) = 3.94$, $p = .0472$). Third, the degree of variation regarding thematic content weight is statistically significant between the learned and the intermediate categories ($\chi^2(1) = 9.91$, $p < .0016$). The frequency of contentful SubjTh is higher in the texts addressed to a learned audience, which implies that the proportion of

– Boke for to lerne a man (1550), Skilful Physician (1656), Method and Means (1683) and An easy way to prolong life (1775) in category 1 (unlearned audience)
– Prognostication (1554?), Marrow of Physicke (1640), Every man his own doctor (1671, 1st ed.) and Every patient his own doctor (1785) in category 4 (intermediate audience).

contentlight Themes is greater in the unlearned and intermediate textual categories. The previous observations lead to the conclusion that 'learnedness' constitutes the distinctive factor that correlates with the distribution of the thematic content in the medical texts.

Let us now deal with the referential content of the Themes investigated. Table 2 contains the frequencies of GivTops, NewTops, QResTops and ResTops per textual category. Since we have shown that the intermediate and unlearned categories act alike with respect to the content weight of Themes, we have grouped these two categories together in Figure 3, which displays the percentages of the Topics in the learned and the intermediate+learned textual categories in a more visual way.

The following remarks with respect to the referential content of the SubjTh in our data seem in order here. First, there are no statistically significant differences between audience categories as regards the introduction of NewTops ($\chi^2(1) = .09$, $p = .7642$). Second, variation between the learned and the intermediate+unlearned categories in terms of the resumption of Themes

Table 2: Reference of SubjTh in medical texts.

	GivTop	NewTop	QResTop	ResTop
learned	41 (24.4%)	216 (33.5%)	28 (28.6%)	155 (36.7%)
intermediate	74 (44%)	213 (33%)	46 (47%)	111 (26.3%)
unlearned	53 (31.6%)	216 (33.5%)	24 (24.4%)	156 (37%)

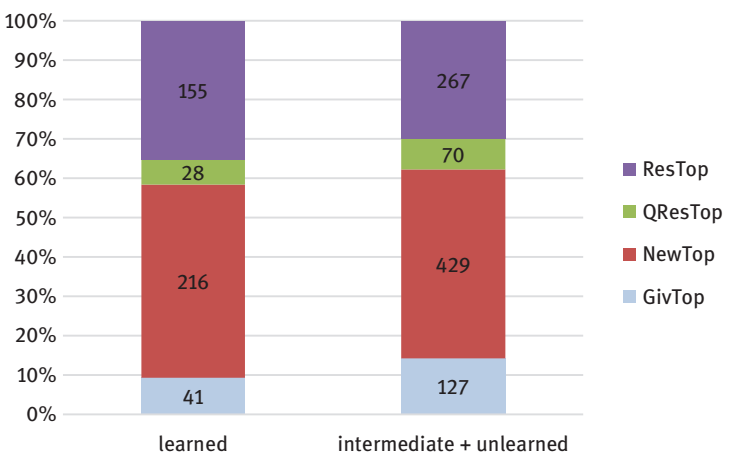


Figure 3: Reference of SubjTh in medical texts.

previously introduced in the discourse reveals a significant difference: in the texts addressed to a learned audience the writer uses more nouns and noun phrases, that is, QResTop and ResTop ($\chi^2(1) = 1.68$, $p = .1949$), whereas in the unlearned+intermediate texts the writer opts for more pronominal themes or GivTop ($\chi^2(1) = 6$, $p = .0143$). As a conclusion, 'learnedness' constitutes the distinctive factor that correlates with thematic reference in the SubjTh in the medical texts.

5.2 Case study II: Newspapers

As an instantiation of the general research question raised in the introductory paragraph, 'do variation in linguistic usage and variation in the context of the communicative event go hand in hand?', this section deals with the degree of correlation between both the content weight and the informational reference of SubjTh, and textual formality in newspaper texts. For this purpose, we have analysed SubjTh in the following textual categories of the news, already discussed in Section 4.1: hard news in weekly papers (samples of foreign news from *The Flying Post*, *The London Gazette* and *The Craftsman; or Say's Weekly Journal*, and home news from *The London Gazette*), soft news in weekly papers (from obituaries in *The Flying Post*, essays in *The Flying Post* and *The Weekly Journal: or, British Gazetteer*, crime news from *The Flying Post*, and letters to the Editor in *The Flying Post* and *The Weekly Journal: or, British Gazetteer*), hard news in daily newspapers (foreign news published in *The Daily Courant* and *The London Daily Post*, home news from *The Daily Courant* and *The London Daily Post*, and shipping news from *The London Daily Post*) and, finally, soft news published by daily newspapers (letters in *The Daily Courant* and *The Daily Post*, accident news in *The Daily Post*, crimes reports in *The Daily Post* and obituaries from *The Daily Post*).

The database is described in Table 3 and the proportions of contentful and contentlight SubjTh per newspaper category and news type are plotted in Figures 4 and 5.

Table 3: Clause count in news.

frequency	news	GivTop	NewTop	QResTop	ResTop
daily	hard	38 (23.2%)	49 (24.1%)	5 (31.3%)	9 (45%)
daily	soft	55 (33.5%)	35 (17.2%)	7 (43.8%)	3 (15%)
weekly	hard	26 (15.9%)	69 (34%)	3 (18.8%)	3 (15%)
weekly	soft	45 (27.4)	50 (24.6%)	1 (6.3%)	5 (25%)

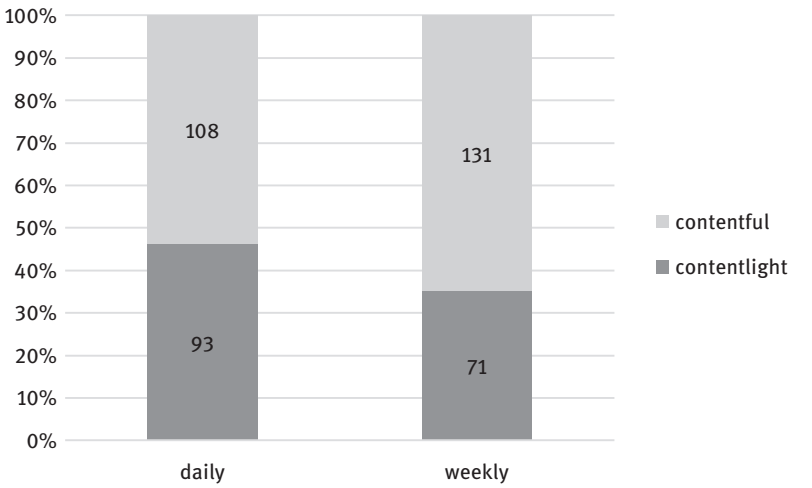


Figure 4: Content weight /daily-vs.-weekly in news.

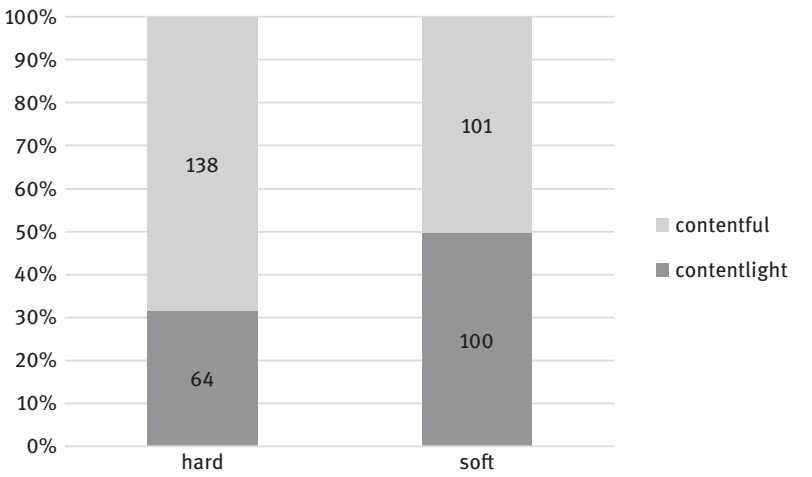


Figure 5: Content weight / hard-vs.-soft in news.

These are the findings with respect to the cross-tabulation of the variables content weight of SubjTh and type of news and newspaper. First, since all the texts are representative of formal language, most SubjTh are contentful in all the newspaper samples. Second, the difference in the empirical distribution of contentful and contentlight Themes per frequency (daily vs. weekly) of the newspaper is not statistically significant ($\chi^2(1) = 4.71$, $p = .03$). Since, as noted in Section 2, the frequency of publication strongly correlates with the target audience of the

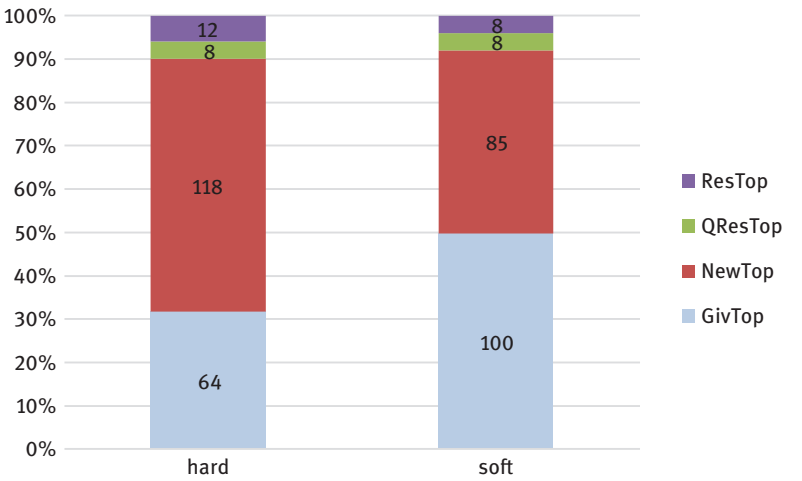


Figure 6: Reference / hard-vs.-soft in news.

newspaper, in light of the previous finding, the audience cannot be claimed to play a role as regards the content weight of SubjTh. Third, the difference between contentlight and contentful themes according to the hard vs. soft status of the texts is indeed statistically significant ($\chi^2(1) = 12.89$, $p = .0003$), so textual formality does play a role in the content weight of SubjTh. Hence, the major conclusion in this study on the connection between content weight and news/newspaper category is that the degree of textual formality constitutes the distinctive factor correlating with thematic content in the news.

In what follows we will consider the referential links conveyed by SubjTh. Figure 6 illustrates the relative proportions of the different types of Themes in so-called hard and soft news.

As for the referential content of SubjTh, no statistically significant differences are observed between textual categories (hard vs. soft news) as regards the introduction of ResTop and QResTop themes ($\chi^2(1) = .82$, $p = .6637$). By contrast, differences between hard and soft news in terms of the reference of Themes in discourse are statistically significant in our database. First, hard news contains more NewTops ($\chi^2(1) = 9.85$, $p = .0017$) since, on the one hand, the style is less involved and hence fewer personal pronouns are attested, and on the other hand the texts in this textual category are shorter than those in the soft news, so fewer thematic referential chains can be established. Second, soft news contains more GivTops ($\chi^2(1) = 12.89$, $p = .0003$) since in this textual variant the style is more involved and personal pronouns are used more widely; the news texts are also longer, which gives rise to the potential for more thematic referential chains. These facts lead

to the conclusion that style constitutes the distinctive factor correlating with thematic reference in SubjTh in the medical texts, and thus supports the hypotheses in Section 4.2.

6 Summary and concluding remarks

In this study we investigated a selection of textual samples retrieved from two corpora of medical texts (EMEMT and LMEMT) and a corpus of newspaper texts (ZEN), with a focus on the Themes realising the syntactic category of Subject (or SubjTh) in declarative clauses. We tested the following hypotheses, which received support from the relevant literature:

(i) SubjTh in medical texts addressed to unlearned audiences are contentlight, while texts addressed to learned audiences should contain contentful SubjTh more frequently. As regards the degree of referentiality conveyed by SubjTh, the medical texts aimed at learned audiences should contain more QResTops and ResTops (nouns and noun phrases).

(ii) SubjTh in newspapers addressed to learned audiences ('respectable' newspapers) should be contentful, while newspapers addressed to other audiences should contain contentlight SubjTh more frequently. SubjTh in the so-called 'hard' news should be more contentful than those in the so-called 'soft' news. The degree of referentiality of SubjTh in news should also correlate with the target audience and/or the textual formality of the texts.

The vast majority of SubjTh in the Modern English medical texts analysed turned out to be contentful, as might be expected in specialised written discourse. As hypothesised, a certain resemblance was attested between our learned category of medical texts and Berry's (2013a) formal written texts, in that contentful (as regards content weight) and lexically complex (informational reference) SubjTh are frequent in such learned texts. On the other hand, resemblance between our unlearned+intermediate target-audience categories and Berry's (2013a) 'speechy' discourse was given support by our data, in that contentlight pronominal SubjTh were frequent in both categories.

As in medical writing, in Modern English newspaper texts, most SubjTh were contentful, something to be expected in formal written discourse. In the case of newspapers, the findings gave support to our hypothesis in (ii) above in that the degree of textual formality was found to play a role in the theme's content weight. Thus, while contentful (content weight) and lexically complex (informational reference) SubjTh were frequent in hard news, contentlight and pronominal SubjTh

were frequent in soft news. The type of (natural) audience, as expressed by the newspapers' frequency of publication (weekly/popular vs. daily/'respectable') did not prove to be relevant to the content weight of the SubjTh, thus leaving our second hypothesis unconfirmed.

In essence, then, significant degrees of correlation were attested between (i) Berry's formal written texts, including medical texts addressed to learned audience, and hard news, and (ii) Berry's more informal 'speechy' texts, including medical texts addressed to unlearned and intermediate audiences, and soft news.

The above findings lead to the following qualitative conclusions. First, 'learnedness' constitutes the distinctive factor that correlates with the distribution of the thematic content and with the thematic reference in the Subject Themes in the medical texts. Second, our analysis of the crosstabulation of content weight and news/newspaper category revealed that textual formality constitutes the distinctive factor that correlates with thematic content in news texts. Third, style has proved to be the distinctive factor that correlates with thematic reference in the Subject Themes in medical texts.

As is usual with studies based on manual analyses, the small number of texts that can be analysed means that firm conclusions cannot easily be drawn. Rather, the aim of studies such as these is to confirm hypotheses which can help researchers characterise historical discourse and serve as pointers for future work (see Berry, Thompson, and Hillier 2014: 108 in this respect). In future research we would like to develop a fine-grained 'contentfulness' scale that takes into consideration Prince's (1981: 237) well-known 'Assumed Familiarity Scale' and to apply it not only to historical samples but also to Present-Day English texts, and to a larger range of text types and textual categories than those under analysis here. Finally, it is also our intention to assess possible correlations between the contentfulness of themes and their syntactic organisation and complexity.

References

Berry, Margaret. 1995. Thematic options and success in writing. In Mohsen Ghadessy (ed.), *Thematic development in English texts*, 55–84. London: Printer.

Berry, Margaret. 1996. What is Theme? A(nother) personal view. In Margaret Berry, Christian Butler, Robin Fawcett, and Guowen Huang (eds.), *Meaning and form: Systemic Functional interpretations. Meaning and choice in language. Studies for Michael Halliday. Advances in discourse processes, vol. LVII*, 1–64. Norwood, NJ: Ablex.

Berry, Margaret. 2013a. Contentful and contentlight subject themes in informal spoken English and formal written English. In Gerard O'Grady, Tom Barlett, and Lise Fontaine (eds.), *Choice in language. Applications in text analysis*, 243–268. Sheffield: Equinox.

Berry, Margaret. 2013b. Towards a study of the differences between formal written English and informal spoken English. In Lise Fontaine, Tom Barlett, and Gerard O'Grady (eds.), *Systemic Functional linguistics. Exploring choice*, 365–383. Cambridge: Cambridge University Press.

Berry, Margaret, Geoff Thompson, and Hilary Hillier. 2014. Theme and variations. In María Ángeles Gómez-González, Francisco Ruiz de Mendoza, and Francisco Gonzálvez-García (eds.), *Theory and practice in Functional-Cognitive space*, 107–126. Amsterdam: John Benjamins.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas and Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

Bös, Birte. 2012. From 1760 to 1960: Diversification and popularization. In Roberta Facchinetti, Nicholas Brownlees, Birte Bös, and Udo Fries (eds.), *News as changing texts: Corpora, methodologies and analysis*, 91–144. Newcastle upon Tyne: Cambridge Scholars.

Brown, Gillian and George Yule. 1983. *Discourse analysis*. Cambridge: Cambridge University Press.

Brownlees, Nicholas. 2005. Spoken discourse in early English newspapers. In Joad Raymond (ed.), *News networks in seventeenth-century Britain and Europe*, 67–83. London: Routledge.

Brownlees, Nicholas. 2006. Introduction. In Nicholas Brownlees (ed.), *News discourse in Early Modern Britain*, 7–13. Bern: Peter Lang.

Cummings, Michael. 1995. A systemic functional approach to the thematic structure of the Old English clause. In Ruqaiya Hasan and Peter H. Fries (eds.), *On Subject and Theme. A discourse functional perspective*, 275–316. Amsterdam: John Benjamins.

Dik, Simon C. (edited by K. Hengeveld). 1997. *The theory of Functional Grammar. Part I: The structure of the clause*. Berlin: Mouton de Gruyter. 2nd edition.

Downing, Angela. 1991. An alternative approach to theme: A systemic functional perspective. *Word* 42(2). 119–143.

Fawcett. Robin P. 2008. *Invitation to Systemic Functional Linguistics through the Cardiff Grammar: An extension and simplification of Halliday's Systemic Functional Grammar*. London: Equinox. 3rd edition.

Fontaine, Lise. 2013. *Analysing English grammar. A systemic functional introduction*. Cambridge: Cambridge University Press.

Forey, Gail and Geoff Thompson. 2008. Introduction. In Gail Forey and Geoff Thompson (eds.), *Text type and texture*, 1–7. London: Equinox.

Fries, Udo. 2001. Text classes in Early English newspapers. *European Journal of English Studies* 5(2). 167–180.

Fries, Udo. 2009. Crime and punishment. In Andreas H. Jucker (ed.), *Early Modern English news discourse. Newspapers, pamphlets and scientific news discourse*, 13–30. Amsterdam: John Benjamins.

Halliday, M.A.K. (edited by Jonathan J. Webster). 2009. *The essential Halliday*. London: Continuum.

Halliday, M.A.K. 1985. *Spoken and written language*. Victoria: Deakin University Press. Reprinted in Oxford University Press 1989.

Halliday, M.A.K. (revised by Christian M.I.M. Matthiessen). 2014. *Introduction to Functional Grammar*. London and New York: Routledge.

Halliday, M.A.K. and Christian M.I.M. Matthiessen. 2004. *An introduction to Functional Grammar*. London: Arnold, 3rd edition.

Hymes, Dell. 1974. *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press.

Lehmann, Hans Martin, Caren auf dem Keller, and Beni Ruef. 2006. ZEN Corpus 1.0. In Roberta Facchinetti and Matti Rissanen (eds.), *Corpus-based studies of diachronic English*, 135–155. Bern: Peter Lang.

Martin, J.R. and David Rose. 2008. *Genre relations. Mapping culture*. London: Equinox.

Matthiessen, Christian M.I.M. 2015. Register in the round: registerial cartography. *Functional Linguistics* 2/9.

Milroy, James and Lesley Milroy. 1997. Varieties and variation. In Florian Coulmas (ed.), *The handbook of Sociolingistics*, 47–64. Oxford: Blackwell.

Montemayor-Borsinger 2009. *Tema: Una perspectiva funcional de la organización del discurso*. Buenos Aires: Eudeba.

Pahta, Päivi and Maura Ratia. 2010. Treatises on specific topics. In Irma Taavitsainen and Päivi Pahta (eds.), *Early Modern English medical texts. Corpus description and studies*, 73–99. Amsterdam: John Benjamins.

Pahta, Päivi and Irma Taavitsainen. 2010. Introducing Early Modern English medical texts. In Irma Taavitsainen and Päivi Pahta (eds.), *Early Modern English medical texts. Corpus description and studies*, 1–7. Amsterdam: John Benjamins.

Pahta, Päivi and Irma Taavitsainen. 2011. An interdisciplinary approach to medical writing in Early Modern English. In Irma Taavitsainen and Päivi Pahta (eds.), *Medical writing in Early Modern English*, 1–8. Cambridge: Cambridge University Press.

Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In Peter Cole (ed.), *Radical pragmatics*, 223–256. New York: Academic Press.

Suhr, Carla. 2010. Regimens and health guides. In Irma Taavitsainen and Päivi Pahta (eds.), *Early Modern English medical texts. Corpus description and studies*, 111–118. Amsterdam: John Benjamins.

Taavitsainen, Irma. 2001. Language history and the scientific register. In Hans-Jürgen Diller and Manfred Görlach (eds.), *Towards a history of English as a history of genres*, 185–201. Heidelberg: C. Winter.

Taavitsainen, Irma. 2010a. Discourse and genre dynamics in Early Modern English medical writing. In Irma Taavitsainen and Päivi Pahta (eds.), *Early Modern English medical texts. Corpus description and studies*, 29–53. Amsterdam: John Benjamins.

Taavitsainen, Irma. 2010b. Expanding the borders of knowledge. In Irma Taavitsainen and Päivi Pahta (eds.), *Early Modern English medical texts. Corpus description and studies*, 11–12. Amsterdam: John Benjamins.

Taavitsainen, Irma and Päivi Pahta. 1995. Scientific 'thought-styles' in discourse structure: Changing patterns in a historical perspective. In Brita Wårvik, Sanna-Kaisa Tanskanen, and Risto Hiltunen (eds.), *Organization in discourse: Proceedings from the Turku Conference Anglicana Turkuensia* 14, 519–529. Turku: University of Turku.

Taavitsainen, Irma, Päivi Pahta, Noora Leskinen, Maura Ratia, and Carla Suhr. 2002. Analysis scientific thought-styles: What can linguistic research reveal about the history of science? In Helena Raumolin-Bunberg, Minna Nevala, Arja Nurmi, and Matti Rissanen (eds.), *Variation past and present: VARIENG studies on English for Terttu Nevalainen. Mémoires de la Société Néophilologique de Helsinki* 61, 251–270. Helsinki: Société Néophilologique.

Taavitsainen, Irma and Jukka Tyrkkö. 2010. General treatises and textbooks. In Irma Taavitsainen and Päivi Pahta (eds.), *Early Modern English medical texts. Corpus description and studies*, 65–72. Amsterdam: John Benjamins.

Taavitsainen, Irma, Turo Hiltunen, Anu Lehto, Ville Marttila, Päivi Pahta, Maura Ratia, Carla Suhr, and Jukka Tyrkkö. 2014. Late Modern English medical texts 1700–1800: A corpus for analysing eighteenth-century medical English. *ICAME Journal* 38. 137–153.

Thompson, Geoff. 2004. *Introducing Functional Grammar*. London: Arnold. 2nd edition.

Weinreich, Uriel, William Labob, and Marvin I. Herzog. 1968. Empirical foundations for a theory of language change. In W.P. Lehmann and Yakov Malkiel (eds.), *Directions for historical linguistics: A symposium*, 97–195. Austin: University of Texas Press.

Catherine Schnedecker

Reference chains and genre identification

Abstract: In this article, I demonstrate that the expression of co-reference is dependent on the text genre in which it is present, which means that the arguments for and against cognitive approaches need to be reconsidered and examined. As it has been demonstrated by the creators of inductivist genre typologies that genre characterization is based on a multi-dimensional analysis taking into account various sets of linguistic traits, a "configurational" approach to reference has major advantages compared to approaches limited to an analysis of the categories of the referential expressions present. This kind of approach also makes it necessary to bring together various methods of genre classification, because they are in a way complementary.

To illustrate and support this argument, I make use of an existing corpus of incipits of fairy stories and news briefs, which are closely related to each other in terms of the production situation, codification, length and content. The analysis shows that a "paradigmatic" approach, based on the quantification of grammatical categories, fails to account for their differences, whereas a study of the reference chains does so highly efficiently.

1 Introduction

In this study we advocate a method of differentiating between discourse genres that is based on the kinds of referential expression employed. This idea is not new in itself (cf. for example Biber & Conrad, 2009, Tutin, 2002 and Condamines, 2005). The novelty of the method however stems from the underlying conception of the proposed analysis of referential expressions, which combines a paradigmatic approach (based on the quantification of referential expressions) and the "syntagmatic" approach that forms the general subject of this monograph. In other words, the idea is to exploit at the same time the discrete units (linguistic

Note: Article written as part of the DEMOCRAT project (DEscription et MOdélisation des Chaînes de Références:outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique (ANR-15-CE38-0008)).

Catherine Schnedecker, Université de Strasbourg, LiLPa EA1339, Fonctionnements discursifs & Traduction

https://doi.org/10.1515/9783110595864-003

forms dedicated to the expression of the reference) and the not-discrete units which are the reference chains.

We will begin by reviewing the results of "paradigmatic" approaches, which are intended to identify genres by focusing on categories of nouns and pronouns and then on the expression of the anaphora. Secondly we will show the limitations of such approaches by reference to a case study. We will thirdly defend the idea of a "configurational" approach to the phenomena, before subsequently setting out in detail how it works. This approach seems more operationally effective, even though it raises theoretical and technical problems from the viewpoint of automatic and quantified analysis.

2 "Generic" constraints affecting expression of the reference: initial approaches

2.1 A paradigmatic approach to grammatical categories used to express reference

Approaches that characterise discourse genres on the basis of grammatical clusters never fail to point out the role played by noun and pronoun phrases. Biber & Conrad (2009: 56 *et seq*.)[1] for example make use of the disparity between nouns and pronouns to show the difference between passages from a geology textbook and from a lecture. But the exploitation of these cues goes no deeper than morpho-syntactical aspects, and they are merely one type of cue amongst others (cf. the list of Biber & Conrad, op. cit: 78 *et seq*.) (cf. Table 1).

This is understandable, as reference theories – whether French or Anglo-Saxon – have hardly ever stressed the role played by these lexical categories in differentiating between discourse genres, and this includes Accessibility Theory (Ariel, 1990), even though it is based on a corpus of authentic texts. With this notable exception, the examples referred to in the other theories are rudimentary to say the least, and are considered totally independently of genre, as shown in (1) and (2) below:

1 See also the list of linguistic characteristics that can be taken into account in genre analysis (pp. 78 *et seq*.), which includes content word classes and pronouns features. In addition, see the section entitled "Noun phrases", pp. 80–81.

Table 1: Normed rates of occurrence (per 100 words) for selected linguistic features (Biber & Conrad, 2009: 65).

Linguistic features	Textbook	Lecture
Pronouns	2.5	15.3
Nouns	29.1	10.8
Mental/desire verbs (e.g. *feel*, *want, believe*)	0.0	4.5
Clause-initial *and/but*	1.3	4.5
Finite relative clause	2.5	3.2
Nonfinite relative clause	6.3	0.0

(1) 1. Susan a offert un hamster à $Betsy_1$ / Susan gave Betsy a pet hamster
    2. Elle lui a rappelé que les hamsters étaient sauvages / She reminded her that such hamsters were quite shy
    3. $Betsy_2$ lui a dit qu'elle aimait beaucoup ce cadeau (Walker *et al.*, transl. by Cornish, 2000) / Betsy told her that she really liked the gift

(2) Un chasseur est arrivé hier. Cet Allemand a manqué tous ses tirs. (Milner, 1982: 24, his ex. (7a)) / A hunter arrived yesterday. This German missed with all his shots.

2.2 A paradigmatic approach to anaphora forms

As far as the French language is concerned, two seminal studies – taking a different view of NP which are now considered in their referential dimension – demonstrate the influence of genre on the use of anaphors. By comparing four different textual genres, Tutin (2002) shows that they vary in terms of anaphoric density (cf. Table 2). The highest anaphoric density occurs in narratives, and the lowest in procedural texts[2]. Tutin also considers the distance, in number of sentences, between the anaphora and its source. This distance is shorter in technical and scientific texts, and greater for newspaper and literary texts.

Tutin thus concludes that:

> anaphora resolution evaluations cannot ignore the textual genre since this variable is essential in the behaviour of discursive phenomena (Tutin, 2002).

2 We have not included all the categories of expressions studied by the author.

Table 2: Anaphoric density among several textual genres (based on Tutin, 2002).

	Human sciences Scientific texts	**Newspaper *Le Monde économique***	**Technical manual**	**Novel (J. Verne, *De la terre à lune*)**
Categories of expressions,				
Total no. of words	24 516	21 784	10 539	18 130
No. of anaph. expr.	621	663	54	783
Anaphoric density	2.53	3.04	0.5	**4.32**
"Clitic" p. pronouns	46	52	**68.5**	54
Possess. determiners	26	31	4	**36.5**
Demonst. pronouns	10	6	**18.5**	3
NHE[a]	**7**	3	4	2
Indef. pronouns	3	1.5	**5.5**	2

[a] Noun head ellipses.

Table 3: Impact of genre on anaphora type distribution (Condamines, 2005: 45).

Corpus	**Hy**	**Supp**	**Syn**	**Dév**	**Déadj**	**Dén**	**Fig**	**total**
Géo[a]	26%	50%	10%	9%	2%	0	3%	100% (266)
GDP[b]	32%	55%	5.5%	5.5%	2%	0	0	100% (246)
Moug[c]	**60%**	31.5%	4.5%	4%	0	0	0	100% (107)
LMD[d]	19%	**64.5%**	9%	1%	1%	0.5%	5%	100% (415)
Bel A[e].	15.5%	47%	**14.5%**	1%	0	0	**22%**	100% (305)

[a] Précis of geomorphology of 206,700 words.

[b] Planning guide of 148,000 words.

[c] *Méthode et outils de génie logiciel pour l'informatique scientifique* (45,100 words).

[d] *Le Monde Diplomatique* (1989). 110,700 words.

[e] *Bel Ami* by Maupassant. 170,200 words.

Abbreviations: Hy = hyperonymes (hypernyms); Supp = supplétifs (suppletives); Syn = synonymes (synonyms); Dév = déverbaux (deverbals); Déadj = dérivés d'un adjectif (derivatives of an adjective); Dén = dérivés d'un nom (derivatives of a noun); Fig = figures.

By considering a larger number of text genres and other anaphoric phenomena involving NP lexical heads, Condamines (2005) shows, in a way that complements the study mentioned above, that certain categories of anaphors predominate in certain genres (cf. Table 3): the hypernym in procedural texts, suppletive anaphors in press articles, synonyms and figurative anaphors in novels, etc.

These results were subsequently backed up by many other articles on:

– a specific genre (e.g. the journalistic portrait (Jenkins, 2002; Schnedecker, 2005), technical documents (Dupont & Bestgen, 2006), instructional discourse (Maes, Arts & Noordman, 2004));

– the comparison of various genres: press, novels, administrative texts (Longo & Todirascu, 2010; Longo, 2013); short stories *vs* journalistic portraits (Baumer, 2012).

Similarly, Anglo-Saxon research on English and other Germanic languages has focused on the important influence of "text genre" on the ways in which reference is expressed (see Table 4: the work of Fox, 1987 (written *vs* conversation opposition); Lord & Dahlgren, 1997; Kirsner & Van Heuven, 1988, on the proximal *vs* distal demonstrative in Dutch; Kronrod & Engel, 2001, (press); Swanson, 2003, (press, narration, academic texts)).

These results are extremely important as they run counter to the findings of Anglo-Saxon theoreticians of reference (cf. Ariel, 2007 in particular) who take the view that:

> Both Accessibility theory (Ariel, 1985, 1990, 2001) and the Giveness Hierarchy (Gundel *et al.*, 1993) have proposed a general (albeit different) account for the use of referring expressions in general (...). *Both theories hardly address themselves to register differences,* although differences between languages are recognized, and there is nothing inherent blocking the assumption of such differences within these theories (in fact, see Ariel 1990: 6.1 and Section 5 below). (Ariel, 2007: 267; our emphasis)

> (...) *there is no direct, conventional association between the specific register and the forms frequently figuring in it*. In each register, it is the same Accessibility Theory/Givenness Hierarchy principles which mediate between the register expectations (re entities) and the resulting linguistic expressions (type of referring expressions) (Ariel, art. cit.: 283; our emphasis)

Table 4: List of studies of the impact of discourse genres on reference expression.

Authors	Genres studied	Language studied
Baumer, 2013	Journalistic portraits vs short stories	English
Fox, 1987	Written *vs* oral	English
Goutsos,	Expository texts	English-Greek
Kirsner & Van Heuven, 1988	Family magazines *vs* governmental publications	Dutch
Kronrod & Engel, 2001	Press	English
Lord & Dahlgren, 1997	Press article	English
Swansson, 2003	Academic journal, news magazines, narration in fiction	English
Condamines, 2005	19th cent. novel, Précis of geomorphology, Planning guide, press, procedural text	Fr.
Tutin, 2002	Scientific texts, press, technical manual, 19th cent. novel	Fr.

and who have shown no interest in questions related to the type or genre of the texts in which the coreferential phenomena described occur[3].

The research papers listed in the table clearly illustrate a paradigmatic approach of the kind that is defined in this monograph. The aim is thus to count the grammatical categories (types of pronouns, in the case of Tutin) and the lexical categories (cf. Condamines) and then assign differences in frequency to the differences in genre. The method has proven its effectiveness and the results are quite convincing.

That said, the choice of certain reference indicator rather than others for the identification of genres is not – to the best of our knowledge – either attributed to a motive or explained in detail.

3 Some limits of "paradigmatic" approaches to reference

To show the limits of this type of approach, we will consider a case-study. The task is to contrast two small corpora. One corpus consists of news briefs (*faits divers* – FD) (Schnedecker & Longo, 2012), and the other is experimental and consists of incipits of fairy tales (*contes de fées* – CF). Referential expressions are observed that are combined with those covered in this two earlier studies. The corpus characteristics are summarised in Table 5 below.

The representativity of the phenomena studied is not affected by the small size of the corpus: the considerable constraints that apply when writing news briefs, and the relatively "formulaic" nature of fairy tales, mean that inter-text discrepancies may be considered to be extremely limited.

Table 5: Characteristics of FD/CF corpus.

	News briefs (FD)	Fairy tales (CF)
No. of texts	46	7
No. of words	9838	1305
Number of referential expressions	905	393
Number of chains	89	23

3 This is perfectly understandable. As their task was already highly complex, they did not seek to increase the number of variables.

These two genres were chosen because of their situational proximity, to use a concept developed by Biber & Conrad (2009). They are written texts; they are produced for readers whom the creator of the text does not know, and who are not present in the production situation; and in both cases the aim is storytelling.

If we consider the linguistic cues usually taken into consideration, the differences between the two genres are few in number, and are limited to:

- grammatical tenses (cf. Table 6),
- temporal reference mode: deictic (FD) *vs* anaphoric (CF)
- the CF opening formula, which in itself suffices to identify the fairy tale: *il était une fois*.

The other items in Table 6 show similarities in terms of form (relative shortness), content ("extraordinary" dimension of what is related) and other microstructural aspects (temporal connectivity markers, NP/VP distribution not very pronounced).

The differences between noun and pronoun categories (cf. Tables 7 and 8) are very slight. The three most commonly occurring categories are: definite NP (35% for FD and 26% for CF), indefinite NP (18% for FD and 26.5% for CF) and personal pronouns (19% for FD and 16.5% for CF), although the frequency order differs in the two genres.

Table 6: Summary of points in common and differences (FD/CF).

	Incipit of fairy tale (CF)	**News brief (FD)**
Situational aspects		
Situation	Unknown recipient, *in absentia*	Unknown recipient, *in absentia*
Medium	Literary work ~	Media (press) Heading system
Communicational purposes	Storytelling & edification	Storytelling & edification
Linguistic aspects		
Content	Extraordinary	Extraordinary
Form	Basically short	Basically short
Type of text / super-structure	Predominantly narrative "narrative scheme"	Predominantly narrative "narrative scheme"
Types of sentences	Predominantly declarative	Predominantly declarative
Grammatical tenses	Perfect	Present/Perfect
Connectivity markers	Temporal dominant Anaphoric temp. ref.	Temporal dominant Deictic temp. ref.
Formula seq.	*Il était une fois*	
Lexical categories	(N ≈ V) > ADJ	(N ≈ V) > ADJ

Table 7: Categories of referential expressions in news briefs (FD) (adapted from Schnedecker & Longo, 2012).

Category	Pr. N	NP Ø	Indefinite NP		Poss. NP	Definite NP		Dem. NP	Pronouns					Poss. Determ.
			bare	expanded		bare	expanded		dem.[i]	reflex.	Ø	relat.	pers.	
No. of occurrences	29	8	71	75	25	175	114	11	5	63	14	29	173	104
Percentages	3.2%	1%	8%	9%	2.5%	20%	12.5%	1.3%	0.5%	7%	1.5%	3%	**19%**	11.5%
			18%			**35%**		**1.8%**			**11.5%**			**30.5%**

Table 8: Categories of referential expressions in incipits of fairy tales (CF).

Category	Pr N	QNP	NP0	Indef. NP		Definite NP		Possess. NP		Dem. NP	Pers. pron.	0	Reflex. pron.	Relative pron.	Misc.
				Bare	Exp.	Bare	Exp.	Bare	Exp.						
Subtotal	2	26	26	21	31	51	15	33	3	13	65	4	7	26	70
Total	2	26	26	52		66		36		13	65	4	7	26	70
	0.5%	6.6	6.6	13.23		16.79		9.1		3.3	16.5	1	1.7	6.6	17.8%
			26.43					**25.89**							

– FD: definite NP > Pronouns > indefinite NP;
– CF: indefinite NP > definite NP > Pronouns

which means it is difficult to interpret these figures in this form.

4 An alternative "configurational" approach

4.1 (Co-)reference chains

However, there are some differences between the two genres.

To highlight them, we will focus on co-reference and on the way in which referential expressions (hereafter RE) form what has become known, thanks to Chastain (1975), as a *reference chain*. This concept is defined as follows[4]:

> La suite des expressions d'un texte entre lesquelles l'interprétation construit une relation d'identité référentielle (Corblin, 1995: 123) (*The sequence of expressions in a text between which the interpretation builds a relationship of referential identity.*)
>
> (...) les suites d'expressions coréférentielles (...). Seules peuvent appartenir (donner lieu) à une chaîne les expressions employées référentiellement, c'est-à-dire toutes et rien que les expressions nominales (et pronominales) permettant d'identifier un individu (un objet de discours) quelle que soit sa forme d'existence (personne humaine, événement, entité abstraite) (Charolles, 1988 : 8) *(sequences of coreferential expressions (....). The only expressions that can belong (give rise) to a chain are expressions used referentially, i.e. all, and nothing but, the nouns (and pronouns) used to identify an individual (an object of discourse) whatever its form of existence (human person, event, abstract entity).*

By way of example, in the following the reference chains that correspond to distinct entities are identified by different marking:

(3) Il était une fois une petite fille de Village, la plus jolie qu'on eût su voir; sa mère en était folle, et sa mère-grand plus folle encore. Cette bonne femme lui fit faire un petit chaperon rouge, qui lui seyait si bien, que partout on l'appelait le Petit Chaperon rouge.
Un jour, sa mère, ayant cuit et fait des galettes, lui dit: Va voir comme se porte ta mère-grand, car on m'a dit qu'elle était malade. Porte-lui une galette et ce petit pot de beurre. Le Petit Chaperon rouge partit aussitôt pour aller chez sa mère-grand, qui demeurait dans un autre Village. En passant dans un bois elle rencontra compère le Loup, qui eut bien envie de la manger. (Charles Perrault: *Le Petit Chaperon rouge,* 1797)

4 Cf. also Schnedecker (1997) and Schnedecker & Landragin (2014).

Once upon a time there lived in a certain village **a little country girl**, the prettiest creature **who** was ever seen. **Her** *mother* was excessively fond of **her**; and **her** grandmother doted on **her** still more. This good woman had a little red riding hood made for **her**. It suited **the girl** so extremely well that everybody called **her** Little Red Riding Hood.

One day **her** *mother*, having made some cakes, said to **her**, "Go, my dear, and see how your grandmother is doing, for I hear she has been very ill. Take her a cake, and this little pot of butter." **Little Red Riding Hood** set out immediately to go to **her** grandmother, who lived in another village. As **she** was going through the wood, **she** met with *a wolf*, *who* had a very great mind to eat **her** up.

(4) Buraliste braquée

Une buraliste de Nancy a été agressée, hier à 6h, alors qu'elle ouvrait son commerce. Deux hommes encagoulés et gantés ont surgi derrière elle alors qu'elle se dirigeait vers la réserve de son bar-tabac. Chacun portait une arme de poing. Ils ont bousculé la gérante qui est alors tombée à terre, avant de demander le coffre. Terrorisée, la victime a expliqué qu'il n'y en avait pas. Ils se sont fait remettre 4 000 € en chèques et espèces. Les deux agresseurs ont ensuite pris la fuite dans une direction inconnue, laissant la buraliste sous le choc. Elle a été soignée à l'hôpital central de Nancy pour une estafilade à l'épaule gauche, due à sa chute. L'affaire est confiée à la sûreté départementale de Meurthe-et-Moselle. (*Républicain Lorrain*, 11/08/2011)

Tobacconist held up

A tobacconist in Nancy was attacked yesterday at 6 a.m., as **she** was opening **her** shop. *Two men* wearing hoods and gloves sprang out from behind **her** as **she** was heading for the storeroom of **her** bar-tobacconist's shop. Each one was holding a handgun. *They* pushed **the shop manager**, **who** fell to the floor, before asking for the safe. Terrified, **the victim** explained that there wasn't one. *They* made **her** hand over €4000 in cheques and cash. *The two attackers* went off in an unknown direction, leaving behind **the shocked tobacconist**. **She** was treated at the Central Hospital of Nancy for a cut to the left shoulder, caused by **her** fall. The affair is in the hands of the Departmental Security unit of Meurthe-et-Moselle. (*Républicain Lorrain*, 11/08/2011)

4.2 Coreference chains and genre differentiation

A study of the reference chains in our two corpora reveals five kinds of differences: formal, syntactic, thematic, lexical and functional.

4.2.1 Formal aspects

In terms of referential density (ratio of number of RE (e.g. referential NP (for example: *une petite fille de village* = 1 RE) /number of words in the text), the CF have greater RE density than the FD: 30%, that is 1 referential expression every 3 words, compared with 1 every 11 words in the FD.

Secondly, the average number of reference chains per text is different: 3 for the CF and 2 for the FD. Average length is also different: 7 RE for the CF compared with 3.42 for the FD. The maximum length of the chains varies depending on the genre: 15 RE for the CF and 27 for the FD.

In lexical terms, thirdly, the RC of the CF have a "stability coefficient", as defined by Perret (2000: 17), that is slightly more stable than that of the FD:

> on proposera le concept de *coefficient de stabilité*, obtenu **en divisant, pour un référent donné (un personnage), le nombre total de d'anaphores nominales par le nombre de désignations différentes**. (Par exemple, dans la *Mélusine* de Jean d'Arras, pour la désignation de l'héroïne on rencontre 164 anaphores nominales, et 17 désignations différentes; le coefficient de stabilité est donc 164/17 = 9,64). Plus le coefficient de stabilité est élevé, moins il y a de désignations différentes par rapport au nombre d'anaphores et donc plus la stabilité référentielle est grande. *(we will propose the concept of the* stability coefficient, *obtained **by dividing, for a given referent (a character), the total number of noun anaphors by the number of different designations.** (For example, in the* Mélusine *by Jean d'Arras, for the designation of the heroine there are 164 noun anaphors, and 17 different designations; the stability coefficient is therefore 164/17 = 9.64). The higher the stability coefficient, the fewer the different designations relative to the number of anaphors, and thus the greater the referential stability.)*

The lexical variations are more numerous in the FD, as illustrated in text (5):

(5) faits divers | blainville-sur-l'eau Immolée par le feu **la victime** décède
Trois jours après **son** admission à l'hôpital Legouest, à Metz, **le jeune homme qui s'**était immolé par le feu à Blainville-sur-l'Eau a succombé à **ses** graves blessures. **Il** est décédé dans la nuit de lundi à mardi (lire RL d'hier).

Samedi, vers 19h, **la victime âgée de 29 ans et connue pour sa fragilité**, avait pris **son** vélo pour **se** rendre sur un terrain vague de Blainville. **L'homme s'**était aspergé d'essence avant d'y mettre le feu. Malgré **ses** brûlures, **il** avait réussi à remonter sur **sa** bicyclette pour rejoindre une route proche. Là, un automobiliste s'est arrêté. Ce témoin a aussitôt prévenu les secours et la gendarmerie.
Après avoir été conditionné sur place, **le malheureux** a été dirigé par hélicoptère vers l'hôpital des Armées Legouest, à Metz.

Depuis **son** admission, les médecins réservaient leur pronostic sur les chances de survie de **leur patient**, grièvement atteint. (*Républicain Lorrain*, publié le 19/07/2011)

news briefs | blainville-sur-l'eau Severely burnt, **the victim** dies

Three days after **his** admission to Legouest Hospital in Metz, **the young man who** had set fire to **himself** at Blainville-sur-l'Eau succumbed to **his** serious injuries. **He** died during the night of Monday to Tuesday (see yesterday's RL).

On Saturday at around 7 p.m.; **the 29-year-old victim**, known to be psychologically fragile, had taken **his** bicycle to go to a wasteland in Blainville. **The man** had poured petrol over **himself** before setting fire to it. Despite **his** burns, **he** had succeeded in climbing back on to **his** bicycle to reach a nearby road. There, *a motorist* stopped. *This witness* immediately contacted the emergency services and the police.

After receiving dressings at the scene, **the unfortunate person** was taken by helicopter to the Legouest Military Hospital, in Metz.

Since his admission, the doctors had been guarded in their prognosis for the chances of survival of their **seriously injured patien**t. (*Républicain Lorrain*, published on 19/07/2011)

I propose to come back to this point later.

Lastly, the composition of the RC varies depending on the genre: that of the central characters (cf. *infra*) of the fairy tales consists predominantly of personal pronouns. That of the FD characters makes greater use of full NP. This is illustrated in examples (3) and (4–5).

Table 9 summarises this first set of observations.

4.2.2 Syntactic aspects

In both the genres, the links of the RC predominate in fulfilling the syntactic functions of the subject, depending on the more or less salient status of the referent concerned:

(6) Il était une fois une reine qui accoucha d'**un fils**$_{\text{COI}}$, si laid et si mal fait, qu'on douta longtemps s'**il** avait forme humaine. Une fée qui se trouva à **sa** naissance assura qu'**il** ne laisserait pas d'être aimable, parce qu'**il** aurait beaucoup d'esprit; elle ajouta même qu'**il** pourrait, en vertu du don qu'elle venait de **lui** faire, donner autant d'esprit qu'**il** en aurait à celle qu'**il** aimerait le mieux. Tout cela consola un peu la pauvre reine, qui était bien affligée d'avoir mis au monde **un**

Table 9: Summary of differences between RC in CF and FD.

	CF	FD
Density of RE/text	30%	9%
No. of RE/words	1 RE/3 words	1 RE/11 words
No. of RC	23	89
Average no. of RC/text	3	2
Average length of RC	7 links	3.42 links
Max. length of RC	15	27
Stability coefficient	+[a]	–
Composition of RC	+ pro	+ NP

[a] The *stability coefficient* is not a binary notion : here (+) indicates tendency to lexical changes; (–) no tendency to lexical change.

si vilain marmot$_{cod}$. Il est vrai que **cet enfant** ne commença pas plus tôt à parler qu'**il** dit mille jolies choses, et qu'**il** avait dans toutes **ses** actions je ne sais quoi de si spirituel, qu'on en était charmé. J'oubliais de dire qu'**il** vint au monde avec une petite houppe de cheveux sur la tête, ce qui fit qu'on **le$_{cod}$** nomma Riquet à la houppe, car Riquet était le nom de la famille. (*Riquet à la houppe)* / Once upon a time there was a queen who bore **a son so ugly and misshapen** that for some time it was doubtful if **he** would have human form at all. But a fairy who was present at **his** birth promised that **he** should have plenty of brains, and added that by virtue of the gift which she had just bestowed upon **him he** would be able to impart to the person whom **he** should love best the same degree of intelligence which **he** possessed **himself.** This somewhat consoled the poor queen, who was greatly disappointed at having brought into the world such **a hideous brat.** And indeed, no sooner did **the child** begin to speak than **his** sayings proved to be full of shrewdness, while all that **he** did was somehow so clever that **he** charmed everyone. I forgot to mention that when **he** was born **he** had a little tuft of hair upon **hi**s head. For this reason **he** was called Ricky of the Tuft, Ricky being his family name. *(Ricky of the Tuft).*

4.2.3 Thematic aspects

In the two genres, the title of the fairy tale or news brief usually designates the character who is to occupy a central, or salient, position or role in the narration. This is the case of Petit Chaperon rouge and Riquet à la Houppe (examples (3) and (6)) and of the NP *la buraliste braquée* and *la victime* in (4) and (5). The referent may be mentioned first (cf. 1) and usually has the longest RC whose scope encompasses the whole of the text.

The differences arise in the ways of introducing the referent, which are relatively systematic in CF and more diversified in FD.

In the CF, the formula *il était une fois* tends to lead to the use of an indefinite NP – and this happens in more than half of cases. This is illustrated by a large number of the first phrases of the *incipits* in our corpus. The NP is quite systematically accompanied by a relative pronoun, and thus by a relative clause into which a second referent is introduced:

(7) Il était une fois **une petite fille de Village, la plus jolie qu'on eût su voir;** / Once upon a time there lived in a certain village **a little country girl, the prettiest creature who was ever seen.**

Il était une fois **un gentilhomme qui** épousa, en secondes noces, ***une femme***, la plus hautaine et la plus fière qu'on eût jamais vue. / Once there was **a gentleman who** married, for ***his second wife***, ***the proudest and most haughty woman that was ever seen.***

Il était une fois **un bûcheron et une bûcheronne qui** avaient ***sept enfants***, tous garçons; / Once upon a time there lived **a woodcutter** and **his wife; they** had ***seven children***, all boys.

Il était une fois **un homme qui** avait de belles maisons à la ville et à la campagne, de la vaisselle d'or et d'argent, des meubles en broderie, et des carrosses tout dorés; / There was once **a man who** had fine houses, both in town and country, a deal of silver and gold plate, embroidered furniture, and coaches gilded all over with gold.

Il était une fois **une veuve qui** avait ***deux filles*** / Once upon a time there was **a widow who** had ***two daughters***.

Il était une fois **une reine qui accoucha *d'un fils***, si laid et si mal fait, qu'on douta longtemps s'il avait forme humaine. / Once upon a time there was **a queen who** bore ***a son*** so ugly and misshapen that for some time it was doubtful if he would have human form at all.

Accordingly, it is possible to define the model of a typical pattern for initial links:

i) Il était une fois un NP_1 qui vb NP_2

The secondary referents are often introduced by possessive NP or NP that are defined as associative anaphors, attached to NP1 (cf. ii):

(8) **sa** mère en était folle, et **s**a mère-grand plus folle encore. (*Le Petit Chaperon rouge*) / **her** mother was excessively fond of her; and **her** grandmother doted on her still more.

(9) Il était une fois une veuve qui avait deux filles: **l'aînée** lui ressemblait si fort d'humeur et de visage, que, qui la voyait, voyait la mère. (*Les fées*) / Once upon a time there was a widow who had two daughters. **The elder** was often mistaken for her mother, so like her was she both in nature and in looks.

ii) NP1 poss. det. NP2 def./poss

As for the news briefs (FD), more than half the titles (67%) refer to a human individual with relatively systematic designation modes. We have identified two "patterns" in this titles, schematically indicated below as (i) and (ii), classified according to the predominant structure: expanded NP (39% of cases; we have included cases in which the title is limited to the past participle), with a roughly equal percentage of "complete" sentences referring to the main protagonist, usually with the grammatical function of subject. In addition, one quarter of referents are introduced by means of a cataphoric NP or pronoun (iii):

i) [Det + N]+pp: 39%

(10) Accident mortel: conducteur poursuivi / Fatal accident: driver prosecuted

(11) Un homme tué par balles sur une aire de l'A31 / A man shot dead at a parking area on the A31

(12) Marseille. Garçonnet tué par un chauffard en fuite / Marseilles. Small boy killed by hit-and-run driver

(13) Uckange. Poignardé à son domicile / Uckange. Knifed at his home

ii) Complete sentence [Ind NP + vb]: 20%

(14) Corse: un détenu violent s'évade d'un hôpital / Corsica: a violent prisoner escapes from a hospital

(15) Une octogénaire périt dans un incendie / An octogenarian dies in a fire

iii) Noun or pronoun cataphor (26%)

(16) **Le faux employé** volait les personnes âgées / **Fake official** stole from the elderly

(17) **Le tireur** se rend aux gendarmes **/ Gunman** gives himself up to police

(18) **Il** tente de s'immoler dans un terrain vague / **He** tries to commit suicide in a wasteland

(19) Poignardé à **son** domicile / Knifed at **hi**s home

(20) **Il s'**immole par le feu en pleine rue / **He** burns **himsel**f to death in the street

4.2.4 Lexical aspects

There is also a difference in lexical categories: either they are almost absent, or their importance and regularity of occurrence vary.

4.2.4.1 Absence of certain grammatical categories

The proper noun is striking in its absence from the two genres, and is more completely absent from the CF (0.5%) than the FD (3.2%). The reasons for this absence are linked to the genre. In the fairy tales, the secondary characters in particular are often reduced to social or family functions; it is not therefore necessary to designate them with a proper noun. This is also true in news briefs. This has an impact on the way the characters are designated. There is also another explanation: for legal reasons, the wrongdoers are not identified and their identity must not be known.

4.2.4.2 Categories of nouns

The two genres are also different in terms of the number of categories of nouns that occur.

In the CF, there are six (cf. Table 10). Our corpus attests the strong presence of terms expressing family relationships (11.5% of RE) with a predominance of terms relating to children (20 lexical units, including 8 occurrences of *enfant* and 12 of (*fils+fille*)), which account for almost half of the family relationship terms. This lexicon is the result of the themes of the genre, which are well known: those of children left to their own devices (they are abandoned, lose their parents, etc.) who have to face a hostile world and ultimately triumph.

In the FD, the types of N are more numerous. They share with the CF the general nouns (1/4 of occurrences), professional nouns[5] (21%, which are different because of the historic period) and the relational nouns (33% of NP). The FD include other types that do not appear in the CF: spatial connection and geographic proximity N, to which are added the gentilic (demonym) nouns and lastly the N referring to those more or less directly involved in the dramatic event described: protagonists (*agresseur, vengeur*) or victims (*blessé, victime*) and more exterior figures, who are merely spectators (*témoin*) (Table 11).

Here also, the functions of tribal togetherness or cathartic identification/projection that are attributed to the FD are underlined by the lexicon, essentially used to differentiate the identity of the protagonists by means of nouns (sex, age, profession), which are common to all human beings, and to set them into an envi-

5 These are the "regulars" in FD, which refer for example to the police force, emergency services or judicial system (referred to in French as the *parquet)*.

Table 10: Typology of human nouns in Fairy Tales.

General human nouns		Human social status nouns	Professional human nouns	Relational human nouns	Family relationship human nouns	Supernatural being nouns
Adult	**Phase N**					
Femme (woman) **Homme (man)** Jeunes gens (young people)	Petite fille(little girl) **Enfant (child)**	Bûcheron (woodcutter) Bûcheronne (woodcutter's wife) Meunier (miller)	Gentilhomme (gentleman) Dame (lady) Honnête homme (civil gentleman) Reine (queen)	**Voisine (neighbour)**	**Mère (mother)** Grand-mère (grandmother) **Filles (daughters)** Mari (husband) Belle-mère (mother-in-law) Aîné (Elder/ eldest) Femme (wife) Amies (friends) Veuve (widow) Cadette (younger/ youngest) **Fils (son)**	Fée (fairy)

ronment or into relatively ordinary and banal situations. These elements tend to bring the persons involved in the FD closer to the "man in the street" and create a sense of proximity for the reader of the FD.

4.2.4.3 Expansions of referential expressions

The final point of divergence between CF and FD is related to the ways in which the NP are expanded.

Expansions occur in 11% of the NP in the CF, and are mainly of an adjectival nature[6]. They are often there to mark an appreciation, which effectively divides the characters into two sides, the "good" and the "bad". Furthermore, the qualification of the characters is extremely rudimentary as it refers to their age and physical appearance. It also enables the narrator to express his degree

6 Furthermore they are systematically postposed, as pointed out by Adam (2012, cf. in particular pp.20 *et seq*.), which in his view accentuates the stereotypisation of the characters.

Table 11: Classification of N types of NP in FD (Schnedecker & Longo, 2012).

General N (16%)	Phase N (9%)	Function/ Profession N (21%)	Relational N (32%)			Gentilics (1%)	Axiological Empathetic (1%)
			social relations	spatial connection	situational connection		
Personne (7) (person) Individu (5) (individual) Homme (49) (man) Femme (17) (woman)	Jeune / vieil homme (7) (young/ old man) Personne âgée (4) (elderly person) Enfant (12) (child) Petit garçon (3) (small boy) Fillette (3) (small girl) N en *-aire* (9) N-generian (*quinquagénaire*) (person in fifties) Jeunes (4) (young people) Retraité (4) (retired person)	Gendarmes (19) (police) Gendarmerie (8) (police service) Enquêteurs (6) (investigators) Police (24) (police) Policier (13) (police officer) Pompiers (24) (firemen) Buraliste (tobacconist) Automobiliste (8) (motorist)	Mère (5) (mother) Fille (10) (daughter) Fils (3) (son) Compagnon (6) (companion) Voisin (18) (neighbour)	Habitant (12) (inhabitant) Riverain (4) (local resident) Occupant (3) (occupant) Résident (2) (resident)	Agresseur (2) (attacker) Vengeur (avenger) Blessé (16) (injured person) Victime (35) (victim) Suspect (6) (suspect) Témoin (6) (witness) Détenu (2) (prisoner) ***Secours (23) (emergency services)*** ***Touriste (tourist)*** Rescapé (survivor) Conducteur (20) (driver) Gérante (manager) ***Auteur (4)***[ii] ***(author)***	Meusien (from the Meuse) Mosellan (from the Moselle) Maizièrois (from Maizières) Messin (3) (from Metz)	Bon Samaritain (Good Samaritan) Chauffard (3) (Hit-and-run driver) Malheureux (2) (unfortunateperson)

of empathy with a particular character, often by means of the adjective *pauvre*, which is often used for this purpose:

(21) **Ce pauvre enfant** était le souffre-douleur de la maison, et on lui donnait toujours tort. / **The poor child** bore the blame of everything that went wrong in the house. Guilty or not, he was always held to be at fault.

(22) Il fallait, entre autres choses, que **cette pauvre enfant** allât, deux fois le jour, puiser de l'eau à une grande demi-lieue du logis, et qu'elle rapportât plein une grande cruche. / One of **the poor child's** many duties was to go twice a day and draw water from a spring a good half mile away, bringing it back in a large pitcher.

(23) Tout cela consola un peu **la pauvre reine**, qui était bien affligée d'avoir mis au monde un si vilain marmot. / This somewhat consoled **the poor queen**, who was greatly disappointed at having brought into the world such a hideous brat. (Table 12)

Table 12: Recapitulation of qualifications in the CF.

	Axiological qual. +	Axiological qual. –	Empathetic qual.
1	Jolie (pretty)		
2	D'une douceur et d'une bonté sans exemple (of unrivalled sweetness and kindness) La meilleure (the best) Bonnes qualités (good qualities)	La plus hautaine, la plus fière (the proudest and most haughty) Mauvaise humeur (Ill-humour) Haïssables (hateful)	
3	Fort délicat (highly delicate)		Pauvre enfant (poor child)
4	Belles (beautiful) Fort honnête (very honest)	Si laid et si terrible (so ugly and so terrifying)	
5	Air posé & sérieux (calm & serious in demeanour)		Si pauvre lot (so little)
6	La douceur et l'honnêteté (sweetness and honesty) Un des plus belles filles (one of the most beautiful girls) Bonne femme (Good woman) Jeune fille (Young girl)	Si désagréables (so unpleasant) Si orgueilleuses (so proud)	
7		Si laid (so ugly) Si mal fait (so misshapen) Si vilain marmot (such a hideous brat)	Pauvre reine (poor queen)

Table 13: Breakdown of RE modifiers and details (Schnedecker & Longo, 2012).

Type of modifier	Noun complement	Adjective
Number	106	58
Percentage	56%	31%
Subcategories	Locator (32%)[iii]	Location (20%)
	Age (12%)	Past participle (31%)

This sharp divide between the characters is crucial for the structure and the didactic content of the tales and what they relate (cf. the abundant literature about this point, *inter alia* :Propp, 1970; Bettelheim, 1999, Paulme, 1986): a character overcomes the abandonment or weakness associated with his or her background, and tries through or despite obstacles (and thus confrontations with the "bad" characters) to achieve a favourable outcome in the end.

As for the FD, the expansions are more abundant – there are almost twice as many as in the CF – as 21% of the RE have expansions. They are more diversified in terms of categories. Adjectives are abundantly used, as are noun complements (cf. Table 13).

In almost half of cases, the NC provide information about the geographic location of referents (13) – systematically for the police force and emergency services, etc. – or their age (14). The adjectives or assimilated forms (cf. *supra* 3.3) mainly provide information about facts (cf. *supra*, 4.2.2) or location (15):

(23) un habitant **de Jezainville** / un habitant **de Folschviller** / au parquet **de Metz** / les enquêteurs de la compagnie **de Toul** et de la section de recherches **de Nancy /** a resident **of Jezainville** / a resident **of Folschviller** / at **the Metz** prosecutor's office / the investigators of **the Toul brigade** and **the Nancy investigation unit**

(24) Un automobiliste **de 28 ans** / d'une mère **de 19 ans** / son petit-fils **d'un an** / un Mosellan **de 74 ans /** a **28-year-old** motorist / a **19-year-old** mother / her **1-year-old** grandson / a **74-year-**old inhabitant of the Moselle

(25) Cette automobiliste **hollandaise** / Un quadragénaire **uckangeois** / un motard **sarrois** / un retraité **mosellan /** This **Dutch** motorist / A **Uckange** resident in his fifties / a motorcyclist **from the Sarre** / a retired man **from the Moselle**

Relatively little use is made of the modifiers to "qualify" the referents physically or morally, unlike the situation with the CF. This neutrality may seem paradoxical

given that the facts narrated can easily give rise to emotion, indignation, etc. But to judge from the content of the NP, it seems in fact that the authors of FD are, probably out of a sense of duty and professional ethics, compelled to respect a certain form of objectivity.

(26) **Le jeune Messin**, âgé de 25 ans, a été déféré hier après-midi au parquet de Metz. Un magistrat lui a remis une date de convocation pour le tribunal correctionnel de Metz. (RL 19/07/2011) / **The young man from Metz**, aged 25, was brought before the Metz prosecutor's office. A magistrate issued him with a date for his appearance before the Metz criminal court.

(27) **la victime âgée de 29 ans et connue pour sa fragilité** (RL, 19/07/2011) / **the 29-year-old victim who was known to be psychologically fragile**

(28) **Cette automobiliste hollandaise** avait alors eu la mauvaise idée d'ouvrir son coffre et les deux chiens avaient, alors, fuit… dans des directions opposées ! (RL, 09/08/2011) / **This Dutch motorist** then had the bad idea of opening her boot, and the two dogs then ran off… in opposite directions.

(29) Vendredi après-midi, alors qu'il rentrait chez lui à pied, **un habitant de Folschviller** a été surpris par l'orage de grêle qui s'est soudainement abattu sur la région de Saint-Avold (lire RL d'hier). (RL, 28/08/2011) / On Friday afternoon, as he was walking home, **a resident of Folschviller** was surprised by the hailstorm that suddenly hit the region of Saint-Avold (read yesterday's RL).

4.2.5 Functional aspects

As seen earlier (cf. *supra*), the referential configurations that distinguish CF from FD are determined by the characteristics of the genres to which they respectively belong.

The number of referents is particularly large in the initial situation of the tales, which portrays the composition (or decomposition) of the family structure, and leads to the emergence of a central character, referred to by the longest reference chain (which explains the average length of the chains). In the FD, the number of referents is just as large as in the CF. But they are more widely distributed for other reasons: the role of neighbours and witnesses, police forces, and ultimately the judicial system. This is why the links succeed each other, in the order of intervention of the various protagonists and witnesses of the dramatic event:

(30) **SCHWEYEN UN MOTARD SARROIS TUE DANS UNE COLLISION**
Alors qu'**il** venait tout juste de passer la frontière avec un groupe d'amis et se dirigeait vers Bitche, **un motard sarrois** a trouvé la mort hier, à 10h, sur la RD 35 à hauteur de Schweyen.

Le pilote menait le groupe de deux-roues. **Il** aurait tenté de dépasser un camion dans une courbe à droite, pourtant marquée par une ligne blanche continue. Un autre poids lourd arrivait en face. La collision était inévitable. **Le pilote** et **sa** moto ont été projetés sur le bas-côté. **Joachim Platt**, 52 ans, est mort sur le coup.

Des pompiers français et allemands sont intervenus sur les lieux de l'accident. ***Ces derniers*** ont pris en charge deux chauffeurs routiers d'outre-Rhin, légèrement blessés et surtout choqués.

Les gendarmes et les agents de l'UTR de Bitche ont coupé la circulation sur la RD 35, un axe fréquenté qui relie Bitche à Deux-Ponts (Allemagne).

SCHWEYEN **A MOTORCYCLIST FROM THE SARRE KILLED** IN A COLLISION

After just crossing the border with a group of friends and as **he** was heading for Bitche, **a motorcyclist from the Sarre** was killed yesterday, at 10 a.m., on the RD 35 at Schweyen.

The motorcyclist was leading the group of motorcycles. **He** reportedly tried to overtake a truck in a left-hand curve, marked however with an unbroken white line. Another truck was arriving in the opposite direction. The collision was inevitable. **The motorcyclist** and **his** motorcycle were projected on to the roadside verge. **Joachim Platt**, aged 52, died instantly.

Firemen from France and Germany intervened at the scene of the accident. *The latter* took charge of two truck drivers from Germany, slightly injured but above all suffering from shock.

The police and road service staff from Bitche cut off the traffic on the RD 35, a busy main road between Bitche and Deux-Ponts (Germany).

In contrast, the chains referring to the victim/aggressor pair are interspersed, which at least partly explains the quantitative importance of the noun anaphors, to avoid referential confusion:

(30) Deux hommes$_{\text{Sujet}}$ encagoulés et gantés ont surgi derrière elle$_{\text{CC}}$ alors qu'elle$_{\text{Sujet}}$ se dirigeait vers la réserve de son bar-tabac. Chacun portait une arme de poing. Ils$_{\text{Sujet}}$ ont bousculé la gérante$_{\text{COD}}$ qui$_{\text{Sujet}}$ est alors

tombée à terre, avant de demander le coffre. Terrorisée, la victime$_{Sujet}$ a expliqué qu'il n'y en avait pas. Il$_{sujet}$ se sont fait remettre 4 000 € en chèques et espèces. Les deux agresseurs$_{sujet}$ ont ensuite pris la fuite dans une direction inconnue, laissant la buraliste$_{COD}$ sous le choc. / **A tobacconist** in Nancy was attacked yesterday at 6 a.m., as **she** was opening **her** shop. *Two men*$_{subject}$ wearing hoods and gloves sprang out from behind **her**$_{CC}$ as **she**$_{subject}$ was heading for the storeroom of **her** bar-tobacconist's shop. Each one was holding a handgun. *They*$_{subject}$ pushed **the shop manager**$_{object}$, **who**$_{subject}$ fell to the floor, before asking for the safe. Terrified, **the victim**$_{subject}$ explained that there wasn't one. *They*$_{subject}$ made **her** hand over €4000 in cheques and cash. *The two attackers*$_{subject}$ went off in an unknown direction, leaving behind **the shocked tobacconist**$_{object}$. **She**$_{subject}$ was treated at the Central Hospital of Nancy for a cut to the left shoulder, caused by **her** fall. The affair is in the hands of the Departmental Security unit of Meurthe-et-Moselle. (*Républicain Lorrain*, 11/08/2011)

4.3 Overview

Reference chains constitute a good way – but of course not the only way – to apprehend and differentiate genres. We summarise the observation points considered below[7] (Table 14).

A list of these points can be useful for a systematic analysis of texts:

1. Number of referents in the text
2. Number of referential expressions (RE) in the text
3. Number of words in the text
4. Referential density calculation (ratio of number of RE/number of words in text)
5. Number of chains and ratio of number of RC/number of referents
6. Number of links in the different chains
7. Grammatical category of links
8. Stability coefficient
9. Length of reference chains
10. Scope of reference chains
11. Order in which 1st links are mentioned
12. Grammatical function of links
13. Suitability of reference chain of main referent to discourse topic (title of novel/text/paragraph, etc.)
14. Lexical sub-categories

7 Shaded lines indicate items that do not enable differentiation between genres.

Table 14: Overview of differences between RC in CF and FD.

	CF	FD
Density of RE/text	30%	9%
No. of RE/word	1 RE/3 words	1 RE/11 words
No. of RC	23	89
Average no. of RC/text	3	2
Average length of RC	7 links	3.42 links
Max. length of RC	15	27
Stability coefficient	+	–
Composition of RC	+ pro	+ NP
Predominant grammatical function of links	+ subject	+ subject
Coincidence RC/theme of text	+	+/–
Way of introducing main referent	*Il était une fois un* NP_1 *qui VB* NP_2	4 patterns (including use of cataphor in headings)
Rare lexical categories	Proper noun	Proper noun
Abundant lexical categories	N	N
Type of expansions	+ Adjectives	Adj. and NC
Mode of cohabitation of RC	Variable depending on characters	Interspersing/succession
Composition of RC	+ homogeneous	+ heterogeneous

15. RC cohabitation modes in the text
16. Incidence of textual division and category of links

To which the following could be added:

17. Semantic role of links
18. Thematic criterion (in the sense of identification of general content (climate change, electrical domestic appliance, etc.))
19. Nature of occurrence proposal of link (main, subordinate, etc.)[8]

These criteria are at the crossroads between paradigmatic approaches of the type advocated by Biber, which quantify the number of lexical and grammatical units and compare their representativity and respective relationships, and syntagmatic approaches, concerned with recurrences of sequences and motifs (cf. Longrée and

8 As in centring theory for example.

Mellet in this monograph ; Frontini, Boukhaled, Ganascia in this monograph). They reflect the paradigmatic approach in the counting of the referential expressions, and grammatical and lexical categories represented, and the syntagmatic approach in the linear phenomena such as the chain of referential expressions, their length and their scope. However, because they include other factors such as distance or semantic roles and syntactic functions, and textual division methods (paragraphic or semantic, particularly by framers) which are neither one nor the other, it could be said therefore that this is a "configurational approach".

Furthermore, in the sense that this approach "transcends" (in terms of the number and nature of the parameters involved) the paradigmatic and syntagmatic approaches, it could be said that it adds considerable weight, precisely because of its configurational dimension[9], to the argument – against those who doubt it (cf. Ariel 2007)[10] – that the expression of reference is strongly correlated to discourse genres and that it enables their identification and discrimination.

5 Technical and theoretical difficulties

This approach however is not without its theoretical and techn(olog)ical difficulties.

From a theoretical viewpoint, there are two major difficulties. The first is that, as Ariel (2007) suggests in her article (significantly entitled "A grammar in every register?"), the predominance of discourse genre could mean that grammar can only be described once it had been filtered in terms of the genres in which its phenomena occur, and that each genre has its own grammar. Ariel (2007) understandably sees this as implausible and above all undesirable: "One can say that language simply cannot afford to have a grammar in every register." (Ariel, art. cit.: 289). Furthermore, on this view, the description of the language – as it would be conditioned by the description of all the genres – could only be deferred, in view of the necessity of first exhaustively describing *all* the genres, an enterprise that is both unrealistic and impossible to achieve. Furthermore, – to turn to the second obstacle – the genres are made up of textual sequences of various types, which also exercise constraints on their microstructural phenomena.

9 Landragin (2014b : 29) parle de « données structurées ».

10 "There is no direct, conventional association between the specific register and the forms frequently figuring in it" (2007: 283).

From a technical viewpoint, in the perspective of a tool-based approach to these phenomena, the difficulties are no less challenging, at least when it comes to those we are able to foresee. This is firstly because, as has been clearly shown by Landragin (2011, 2014a et b), the task of annotation (whether manual or automatic) is highly complicated because it requires the indication of "units", "relations" (link between two or n units) and a "scheme", *i.e.* according to the author, a structured whole of units, relations, and recursively of schemes, in addition to the fact that it is necessary to annotate the type of anaphor, and the enunciative level (e.g. narration, dialogue, etc.). Secondly it is because it is not based solely on syntactic annotations of a categorial type (which exist) but that it also requires semantic annotations on the lexical categories used: for example, Tables (10) and (11) above, require the annotation of human nouns in operational subcategories, and the adjectives of Tables (12) and (13) according to their nature (descriptive, axiological, etc.)

6 Openings rather than conclusions

The aim of this contribution was to demonstrate that paradigmatic approaches have perhaps not sufficiently taken on board the potential offered by NP and pronouns in genre analysis. On the one hand, they consider them only from a strict morpho-syntactic viewpoint and not from a semantico-referential viewpoint. On the other hand. if they do this, the paradigmatic perspective leads to results that are too widely dispersed to be operational.

We have chosen an alternative "configurational" approach which makes it compulsory to consider "ensembles structurés d'unités" (*"structured sets of units"*), to use Landragin's words, as this is the only way to see referential expressions in relation to the text (and context) in which they occur, and thus to have a view of the phenomena of textuality and discourse genres that is not only broader but also multidimensional, as was called for by Biber & Conrad (2009: 233 *et seq*.), which takes into account the interaction between the textual organisation levels defined by Charolles (1988).

This kind of approach clearly requires new analysis tools (cf. the descriptions and applications of ANALEC software by Landragin, in the articles referred to), and corpora whose lexical units are syntactically *and* semantically annotated. In terms of efficiency, this approach has shown itself to be operational on sufficiently diversified text genres for the enterprise to be worth continuing. It raises hopes of results in terms of a "genre-dependent" reference theory and of the analysis of the genres themselves, both from a linguistic and literary or stylistic viewpoint.

References

Adam, Jean-Michel. 2001a. Entre conseil et consigne: les genres de l'incitation à l'action. *Pratiques* 111/112. 7–38.

Adam, Jean-Michel. 2001b. Types de textes ou genres de discours ? Comment classer les textes qui disent de et comment faire ?. *Langages* 141. 10–27.

Adam, Jean-Michel. 2012. Grammaire, généricité et textualité dans les contes de Perrault: l'exemple de la place de l'adjectif dans le groupe nominal. In Claire Despierres & Mustapha Krazem (eds.), *Quand les genres de discours provoquent la grammaire... et réciproquement*, 9–25. Limoges : Lambert-Lucas.

Ariel, Mira. 1990. *Accessing Noun-Phrase Antecedents*. Theoretical Linguistics Series. London & New York: Routledge.

Ariel, Mira. 2007. A Grammar in every Register? The case of Definite Descriptions. In Nancy Hedberg & Ron Zacharsky (eds.), *The Grammar-pragmatic Interface. Essays in Honor of J. K. Gundel,* 265–292. NYC/Philadelphia, J. Benjamins.

Baumer, Emmanuel. 2012. *Noms propres et anaphores nominales en anglais et en français: étude comparée des chaînes de référence*. Thèse de doctorat, Paris Diderot.

Bettelheim, Bruno. ed. 1999. *Psychanalyse des contes de fées*. Paris : Pocket.

Biber, Douglas & Conrad, Susan. 2009. *Register, Gender and Style*, Cambridge : Cambridge U. P.

Charolles, Michel. 1988. Les plans d'organisation textuelle: périodes, chaînes, portées et séquences. *Pratiques* 57. 3–15.

Charolles, Michel. 1995. Cohésion, cohérence et pertinence du discours. *Travaux de Linguistique* 29. 125–151.

Charolles, Michel. & Pery- Woodley, Marie-Paule. 2005. Les adverbiaux cadratifs: introduction. In Michel Charolles & Marie-Paule Péry-Woodley (eds.), *Les adverbiaux cadratifs. Langue Française* 148. 3–8.

Chastain, Charles. 1975. Reference and Context. In Keith Gunderson (ed.), *Language Mind and Knowledge*, 194–269, Minneapolis: University of Minnesota Press.

Condamines, Anne. 2005. Anaphore nominale infidèle et hyperonymie: le rôle du genre textuel. *Revue de sémantique et pragmatique* 18. 33–52.

Cornish, Francis. (ed.). 2000 *Référence discursive et accessibilité cognitive. Verbum* XXII/1.

Dupont, Vincent & Bestgen, Yves. 2006. Learning From Technical Documents : The Role of Intermodal Referring Expressions. *Human Factors, The Journal of The Human Factors and Erogonomics Society Summer* 48/2. 257–64.

Fox, Barbara. 1987. *Discourse structure and anaphora*. Cambridge : Cambridge U. P.

Frontini, Francesca, Boukhaled, Mohamed-Amine, Ganascia, Jean-Gabriel 2018. Syntactic Characterisation of French Theatrical Characters: a Study with Motifs and Correspondence Analysis. In Dominique Legallois, Thierry Charnois and Meri Larjavaara. *Grammar of Genres and Styles*.

Jenkins, Christina. 2002. Les procédés référentiels dans les portraits journalistiques. *XV Skandinaviske romanistkongress*. 507–516.

Kirsner, Robert. S. & Van Heuven, Vincent. 1988. The Significance of Demonstrative Position in Modern Dutch. *Lingua*, 76. 209–248.

Kronrod, Ann & Engel, Orit. 2001. Accessibility Theory and Referring Expressions in Newspaper. *Journal of Pragmatics* 33. 683–699.

Landragin, Frédéric & Schnedecker, Catherine (eds.). 2014. Les chaînes de référence. *Langages* 195.

Longo, Laurence. 2013. *Vers des moteurs de recherche* « intelligents »*: un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaînes de référence*, Thèse de doctorat, Université de Strasbourg.

Longo, Laurence & Todirascu, Amalia. 2009. Une étude de corpus pour la détection automatique des thèmes, *actes des 6èmes journées de linguistique de corpus,* 10–12 septembre 2009, Lorient.

Lord, Carol & Dahlgren, Karen. 1997. Participant and Event Anaphora in Newspaper Articles. In John Haiman & Sandra A. Thompson (eds.), *Essays on Language Function & Language Type Dedicated to T. Givon*, 323–356. Amsterdam: J. Benjamins.

Maes, Alfons, Arts, Anja, Noordman, Leo. 2004. Reference Management in Instructive Discourse. *Discourse Processes* 37/2. 117–144.

Mélanie-Becquet, Frédérique & Landragin, Frédéric. 2014. Linguistique outillée pour l'étude des chaînes de référence. *Langages* 195. 117–137.

Paulme, Denise. (ed.) 1986. *La mère dévorante. Essai sur la morphologie des contes africains*. Paris: Gallimard.

Perret, Michèle. 2000. Quelques remarques sur l'anaphore nominale aux 14° et 15° siècle. *L'information grammaticale* 87. 17–23.

Propp, Vladimir. 1970. *Morphologie du conte*. Paris : Seuil.

Schnedecker, Catherine. 1997. *Nom propre et chaînes de référence*. Paris : Klincksieck.

Schnedecker, Catherine. 2005. Les chaînes de référence dans les portraits journalistiques: éléments de description. *Travaux de linguistique* 51. 2005/2. 85–133.

Schnedecker, Catherine. 2011. La notion de « saillance »: problèmes définitoires et avatars. In Olga Inkova (ed.) *Saillance, Aspects linguistiques et communicatifs de la mise en évidence dans un texte*. 23–43. Besançon, PUFC.

Schnedecker, Catherine. 2014. Chaînes de référence et variations selon le genre. *Langages* 195. 23–42.

Schnedecker, Catherine & Landragin, Frédéric. 2014. Les chaînes de référence. Présentation. *Langages* 195. 3–22

Schnedecker, Catherine & Longo, Laurence. 2012. Impact des genres sur la composition des chaînes de référence: le cas des faits divers. In Franck Neveu *et al.* (eds.), *3ème Congrès Mondial de Linguistique Française*, Lyon, 1957-1972, http://www.shs-conferences.org/articles/shsconf/abs/2012/01/shsconf_cmlf12_000061/shsconf_cmlf12_000061.html

Swales, John. 1990. *Genre analysis: English for academic and research settings*. Cambridge: Cambridge U.P.

Swanson, Wency. 2003. *Modes of Co-reference as Indicator of Genre*. Bern: P. Lang.

Tutin, Agnès. 2002. A corpus-based study of pronominal anaphoric expressions in French, *Proceedings of DAARC 2002 (Discourse Anaphora and Anaphora Resolution),* Lisbon.

Walker, Marilyn, Joshi, Aravind & Prince, Ellen. 1998. Centering in Naturally Occurring Discourse: An Overview. in Marilyn Walke M. *et al.* (eds.), *Centering Theory in Discourse*. 1–28, Oxford, Clarendon Press.

Walker, Marilyn, Joshi, Aravind & Prince, Ellen. (eds.) 1998. *Centering Theory in Discourse*. Oxford: Clarendon Press.

Olivier Méric

Taking into account coherence relations to describe a textual genre: methodology and application to the discourse of tourist attraction guides

Abstract: This study aims to present an empirical process of corpus-based text analysis grounded on an optimal non-discrete unit segmentation called 'micro-contribution', and on the tagging of unit relations. The theoretical framework of this method remains pragma-semantic even if the micro-contribution definition is essentially based on cognitive concepts. Indeed, the constraint of relevance (Sperber and Wilson 1986) and the constraint of completeness (Português 2011; Borderieux 2016) are enough to define the concept of micro-contribution, while the relational tagging used in the Rhetoric Structure Theory concept developed by Mann and Thompson (1988) sets the coherence relationship features of these text units. One of the distinctive features of this suggested method is its *bottom-up* approach: from the first step, the method relies mainly on the texts which are representative of a specific communicative situation. Then, step by step, the researcher discovers the structural organisation of the different semantic levels. Thus, at the end of the analysis, he can suggest a specific feature collection describing the representative prototype of the studied text genre. In this study, the various stages of this analytical technique are grounded on a corpus which consists of texts produced during visits to tourist attractions.

1 Introduction

The development of linguistic theory in the last decades shows different paradigm shifts. One of the most noticeable developments is the perspective change from a sentence focused analysis to a more textual or discourse based approach. This evolution leads researchers to leave a closed system bound by a conventional written code, which was considered the unit of linguistic analysis by Chomsky (1965, 1966) and his followers, to an open system, where the structure of a text,

Olivier Méric, laboratoire TIL (Texte – Image – Langage) EA, 4182 de l'ED 491 de L'Université de Bourgogne Franche-Comté, France

https://doi.org/10.1515/9783110595864-004

its communicative setting and the context of both author and addressee are taken into account in order to analyse how meaningful interactions are built (Adam 2008, 2014; De Beaugrande 1995; Halliday 2006). In a semiotic framework, Saussure ([1916] 1995) states that language is a system grounded on a unit defined on a dyadic relation between the form of a sign (the signifier) and its meaning (the signified). In a postulational framework based on a similar dyadic association between form and meaning, Bloomfield provides various definitions of linguistic units: the word becomes the smallest free form[1] which "may be uttered alone (with meaning) but cannot be analysed into parts that may (all of them) be uttered alone (with meaning)" (Bloomfield 1926: 197), the phrase becomes "a non-minimum free form" (Bloomfield 1926: 197), and the sentence becomes "a maximum construction in any utterance" (Bloomfield 1926: 198). In his approach, Bloomfield builds a system grounded on axioms and limited to a sentence. Whether it be semiotic, postulational or structuralist, in such approaches the analyst is constrained to understand the language as a closed system where he regularly searches for the best basic building block by which relationships could describe how meaning is built and interpreted. From this regulated and normed system, new approaches have been developed grounded on new ideas of linguistic units which "attempt to study the organization of language above the sentence or above the clause, and therefore to study large linguistic units" (Stubbs 1983: 1). This first step of moving beyond the sentence provides a place where the context and extra-linguistic features can be taken into account in order to fully understand the meaning of a discourse: a place where language is considered in its use.[2] The concept of discourse evolves, e.g. Benveniste (1966: 266) considers that discourse is interactive, Fairclough (1992: 28) argues that language is "more than just language in use: it is language use, whether speech or writing, seen as a type of social practice". As the purpose of this article is not to present a historical list of the concepts defining what discourse is, the discussion will remain oriented around the objective of discourse analysis: "Do we look to describe the general characteristics of functioning discourse, or the proper characteristics of a specific discourse, in others words of a text?" (Charaudeau 1995: 103). The first choice comes within the competency of an immanent approach which constrains discourse interpretation to the description of a universal normative internal structure that, until now, has failed to clarify the discursive processes applied in a communicative interaction.

1 Bloomfield states that "a form which may be an utterance is free", giving the example of *book* or *the man*.

2 Brown and Yule (1983: 1) state that "the analysis of discourse is, necessarily, the analysis of language in use".

Reboul and Moeschler (2005: 36) ascribe this failure to the closed feature of an immanentist system. According to these authors, and considering the enunciative situation of the studied discourses compiled in the aforementioned corpus, text analysis research is deliberately oriented toward a specific description of discourse through a bottom-up corpus-driven analysis based on the optimal units and on the relationships they maintain. Even if these optimal units are further defined, it is relevant to highlight here that they are non-discrete combinations of linguistic units whose relationships are part of the process of text coherence construction. Therefore, an empirically opened approach is kept which allows for the progressive discovery of discourse structures without necessarily coining them in a more universal framework in order to characterise a specific textual genre. The determination of the optimal units relies on two cognitive parameters: relevance, introduced by Sperber and Wilson (1986) and related to Grice's maxim of relation; and completeness, introduced by Roulet (1986) and related to Grice's maxim of quantity. The relations between the optimal units follow the Rhetorical Structure Theory model developed by Mann and Thompson (1988), while the annotating process shows different levels of text organisation.

This contribution aims to demonstrate how, applying a bottom-up approach to a representative corpus, the coherent relationships of optimal units may reveal the different internal levels of specific discourse organisation produced in a particular communicative situation; hence, how coherence relationships may characterise a specific discourse genre. To achieve this objective, the analytic process is illustrated through a corpus compiled of tourist visits conducted in French and Spanish, assisted by a socio-technical device, and French and Spanish visits guided by a visitor service officer. Before detailing the corpus characteristics, it is necessary, in a second section, to introduce and define the optimal units and their grounded theory. Then, in the third section, in addition to the corpus description, the steps followed in the analytic process are specified in detail. The fourth section presents the experimental results stemming from the use of the previously outlined approach. Finally, a discussion on further investigations and a conclusion close the reflexion.

2 Optimal units and their theoretical background

As Fairclough states, discourse cannot be dissociated from a social practice represented by a communicative situation where speakers and addressees produce and interpret a discourse in a specific context. Figure 1 illustrates the theoretical model upon which this analysis is based to describe the discursive context conception of the discourse production when a subject is immersed in a specific communicative situation. The locutor has his own perception of the subject

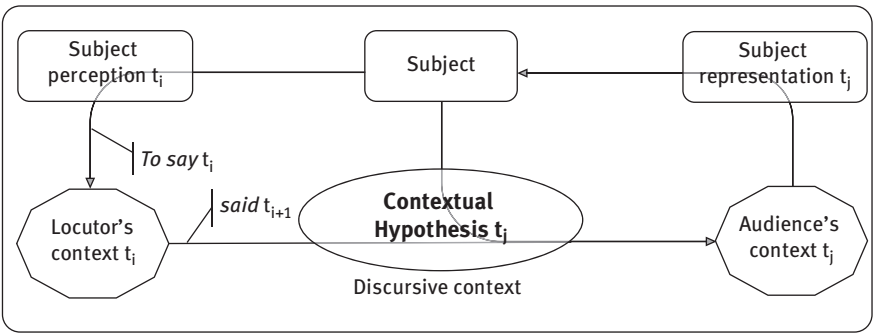


Figure 1: Discursive theoretical model of a micro-contribution production

about which he has something to say; the *to say* content depends on intrinsic parameters such as his own knowledge, intentions or objectives, which differs from the *said* content, as the extrinsic parameters of the discursive context lead the locutor to constantly adapt his production to his direct environment, taking into account external events and addresses' behaviour. The addressees infer the meaning of what is *said* from the contextual hypotheses which are built thanks to the locutor's discourse, the addressees' perception of the subject, and the unpredicted events occurring in the discursive context. Then, the addressees process the received information to draw their own representations of the subject relying on their own knowledge, intentions and objectives. Thus, we can conclude that when t_{i+1} and t_j occur at the same moment, the locutor and the addressees are in a synchronic communication; however, when they occur at different moments, the locutor and the addressees are in an a-synchronic communication.

Even if the pragmatic dimension is as present in the production as in the interpretation, the conducted study follows the locutor's point of view, thus the text analysis is situated at its time and context production (t_{i+1}). Therefore, the outcomes describe the contribution called *said*, which, according to Portugués, has to be complete:

> La contribution est une contrainte propre à l'énonciateur et qui a pour but de transmettre avec une efficacité maximale l'information à un interprétant. La communication entre les deux interlocuteurs repose sur le postulat selon lequel l'énonciateur doit contribuer, c'est-à-dire former une contribution qui aura tous les éléments permettant à l'énonciateur de saisir quelle est l'information que l'énonciateur souhaite transmettre. [The contribution is an enunciator's constraint and has the purpose of efficiently transmitting information to an interpretant. The communication between both interlocutors relies on the postulate that the enunciator must contribute, that is to say, he must form a contribution which will present all necessary information to allow the interpretant to understand the enunciator's message]. (2011:89)

Therefore, the two cognitive parameters – relevance and completeness – provide essential information to distinguish a text from its parts which may also be considered as contributions. Indeed a text can only be considered as such if it is complete and relevant regarding the discursive context and may include parts which have their own relevance and completeness regarding the text itself. Taking these considerations into account, the smallest relevant and complete segment, called 'micro-contribution' in this study, is the smallest optimal unit of a text. Depending on the length of the text, between the macro-contribution and the micro-contribution levels, intermediate relevance and completeness may exist at the meso-contribution level substantiated by the schemata in this analysis (Figure 2). On the one hand, the text is segmented into different levels of contributions and its coherence arises from the relevance of each contribution; that is to say each contribution is relevant if the link that ties it with another contribution ensures the pragmatic and semantic continuity of the text. On the other hand, an empirical contribution, produced within a social praxis, can be considered as a text and be divided into three semantic levels: a macro-semantic, a meso-semantic, and a micro-semantic level. Each level presents its optimal non-discrete unit (cf. Figure 2).

- At the macro-semantic level, the optimal unit is the text itself. The researcher analyses the structure and the organisation of the text, the nature of its parts and their relationships, whether it be at the macro-, meso- or micro-contribution level, in order to suggest a rhetorical structure of the studied text genre.
- At the meso-semantic level, the optimal unit is the schema. The researcher analyses the intermediate parts of the text, which present their own relevance and completeness, and their relationships at the meso and micro semantic level in order to suggest a rhetorical structure of the schema.
- At the micro-semantic level, the optimal unit is the micro-contribution. The researcher analyses the smallest relevant and complete parts of the text which are composed of praxemes[3] (Lafont 1978). Each praxeme conveys its own meaning but this meaning does not necessarily respect the completeness constraint, and if it does, the praxeme is also a micro-contribution.

This bottom-up pragma-semantic approach sets the theoretical background for the non-discrete optimal units of a text and the questions the researcher must answer to complete the text analysis.

3 The praxeme comes from the association of *praxis* and *semeion*, thus it combines the meaning production with the locutor's social practice. I choose this entity to guarantee the pragma-semantic continuity from the text to the smallest free form.

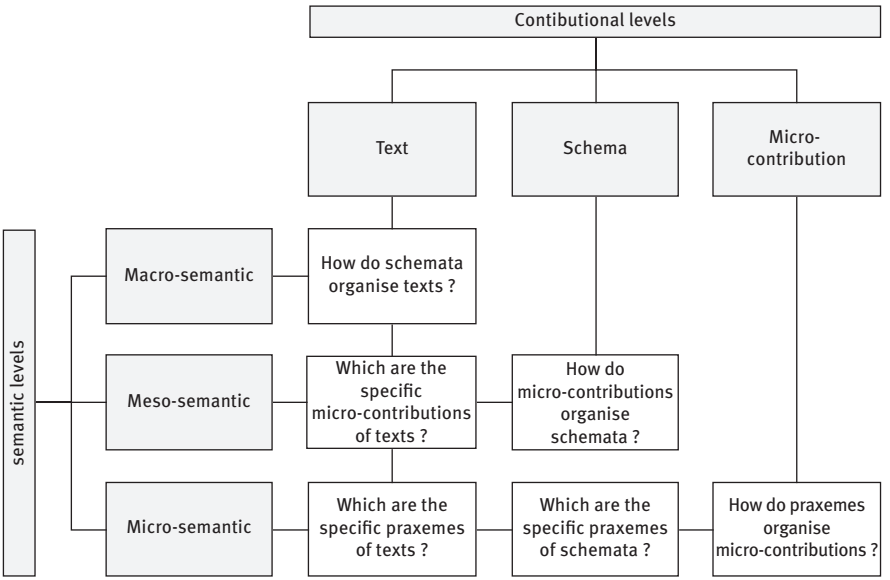


Figure 2: Cross level analysis of texts

All the texts compiled in the corpus are authentic, produced in a specific socio-professional praxis. They are recorded or written, respect the completeness constraint and they deal with relevant subjects; hence, they can all be considered as macro-contributions. In the next section, the methodology followed is detailed which defines the different optimal units that shape the analysis of each pragma-semantic level of the studied text genre.

3 Corpus and analytic process

Biber, Connor, and Upton argue that the corpus based bottom-up approach "should provide a comprehensive linguistic description of discourse units and the flows of discourse", but should also "describe generalizable patterns of discourse organization that hold across all texts of the target corpus" (2007: 156). The corpus appears to be the object of the research and it is organised to be representative of the linguistic phenomena associated with a specific social praxis. A French ($_f$) and a Spanish ($_s$) corpus were compiled to represent the discourse produced in an enunciative situation during a tourist attraction visit. Both corpora are divided into two socio-professional sub-corpora related to the visit modality: visits assisted by a socio-technical device (Assisted Visit Conference), and guided visits by an education and visitor service officer (Guided Visit Conference). Table 1

Table 1: Criteria to select the discursive contexts for the studied corpora.

Criteria	Guided Visit Conference	Assisted Visit Conference
Canal of production	Oral	Written
Canal of enunciation	Oral	Oral
Format	Ephemeral	Perennial
Background	Institutional	Institutional
interpretant	Native adult speakers	Native adult speakers
number	A group	A person
on site	Present	Absent while discourse production
Interaction	Direct and synchronous	socio-technical device and asynchronous
Author	Native officer	Native professional
Enunciator	Native officer	Native actor
Factuality	Informative – factual	Informative – factual
Functions	Describing, informing, entertaining	Describing, informing, entertaining
Theme	Cultural visit	Cultural visit
Geographic situation	France and Spain	France and Spain
Period	from 2010 to 2015	from 2010 to 2015

Figure 3: Axes of the corpus analysis

presents the criteria which, according to Biber (1993: 245), describe the selected discursive professional situations to be included in each corpus context.

The bilingual feature of the presented specialized socio-professional corpora is adapted to the Sinclair (1996: 12) definition of a comparable corpus: "a comparable corpus is one which selects similar texts in more than one language or variety" sharing common characteristics such as genre, domain, theme, etc., without being a translation. Therefore, there are two different comparative axes determining the *tertium comparationis* (cf. Figure 3).

The axis of the abscissa represents the language variety of the different modalities while the axis of the ordinate represents the French and Spanish language differences. All the Guided Visit Conference texts (146055 tokens in French and 21178 tokens in Spanish) come from my transcriptions of the professional activity which I have personally recorded. All the Assisted Visit Conference texts

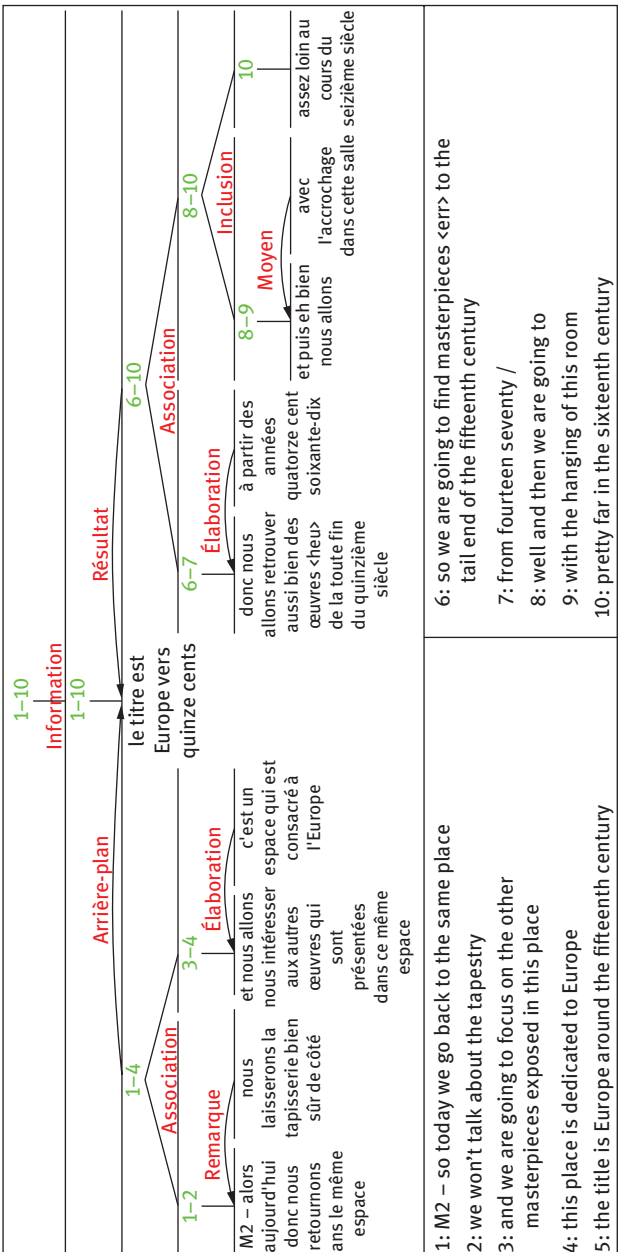
(94767 tokens in French and 79621 tokens in Spanish) have been provided by private companies specialized in audio-guide discourse production[4] or by the tourist institution itself.[5] Therefore, we can consider the texts representative of the professional practices of the domain in the selected period.

In the bottom-up pragma-semantic approach, the first step is to segment the text following the constraints of relevance and completeness so that texts are divided into micro-contribution independently of their specific oral or written features such as punctuation marks. Thus, the discourse appears to be a divisible set of contributions whose elements are linked to ensure coherence. In order to evidence this pragmatic and semantic textual continuity, the framework of Rhetorical Structure Theory (Mann and Thomson 1987, 1988; Mann, Matthiessen, and Thompson 1992; Taboada and Mann 2006a, 2006b) is applied, which provides a systemic method to identify and describe the relations that exist between the different micro-contributions. Sperber and Wilson's relevance theory ensures text coherence: not only does each micro-contribution have to be relevant, but the link between the two micro-contributions has to be relevant in itself as well. The meaning of a contribution is not the sum of the meaning of its micro-contributions, their relationships are also an important element of the process of meaning construction. Therefore, each new micro-contribution is constrained by its own relevance and the relevance of its relationships step by step building the coherence of the text. Once the text is segmented, it is necessary to tag the relationships that each micro-contribution may have with another micro-contributions. These two stages – segmenting and tagging – of the bottom-up pragma-semantic approach are mainly manual in nature, but tools like *RSTTool* software help to systematise the process and provide meaningful outcomes (cf. Figure 4).

Rhetorical Structure Theory does not have the ability to explain the theory of discourse structure, and thus appears to be a neutral tool describing relations at the micro-contribution level independently of linguistic markers (Bateman and Rondhuis 1997: 26) in agreement with the bottom-up pragma-semantic approach, since the structure of the discourse stems from the relationships between the smallest relevant and complete segments of the text. In order to make the process more objective, Mann and Thompson suggested various generic relations divided into nucleus-satellite and multinuclear relations (2005-2015), leaving enough flexibility in their taxonomy to allow researchers to define their own relations according to their communicative situations. At the meso-semantic level, according to Mann and Thompson, "schemas define the structural constituency arrangements

4 Histoire de son, Sycomore, audio-guías Bluehertz.

5 Bibracte Museum.



1: M2 – so today we go back to the same place
2: we won't talk about the tapestry
3: and we are going to focus on the other masterpieces exposed in this place
4: this place is dedicated to Europe
5: the title is Europe around the fifteenth century
6: so we are going to find masterpieces <err> to the tail end of the fifteenth century
7: from fourteen seventy /
8: well and then we are going to
9: with the hanging of this room
10: pretty far in the sixteenth century

Figure 4: Example of micro-contribution relation tag tree

of text. They are abstract patterns consisting of a small number of constituent text spans, a specification of the relations between them, and a specification of how certain spans (nuclei) are related to the whole collection" (1987: 8). The schemata organise the text "like multinuclear groups, but each element of the group has a distinct functional label, e.g., the Paper schema may be defined to have the elements *Title Author Abstract Section References*. Several elements may serve the same function in a schema, e.g., we may have several Section nodes in a Paper analysis" (O'Donnell 2003). Once the text is segmented and tagged, it is re-organised according to the contribution level:

- At the macro-contribution level, the first outcome is the schema organisation of the text (macro-semantic information). The second outcome gathers all the statistics regarding the micro-contribution relations employed to produce the texts of the studied discursive genre (meso-semantic information). The third outcome refers to the "classic" Corpus Query Language including frequencies, specificity scores, collocations and concordances of the specific praxemes used in the text (micro-semantic information).[6]
- At the meso-contribution level, the first outcome is the micro-contribution organisation of the different schemata (meso-semantic information). The second outcome refers to the "classic" Corpus Query Language system applied to each schema. To do so, the researcher compiles all the text elements that compose a schema and makes the queries to this "artificial" sub-corpus in order to obtain the frequencies, the specificity scores, the collocations and the concordances of the specific praxeme used in each schema (micro-semantic information).
- At the micro-contribution level, the researcher creates an artificial sub-corpus for each relationship in order to obtain particular frequencies, specificity scores, collocations and concordances of the praxeme used in each micro-contribution (micro-semantic information).

These three analytic steps answer the question raised in Figure 2, providing information according to the different contributional and semantic levels. Thus, the process of text organisation is to describe how its parts are organized, as well as providing information on how coherence is built on the basis of the statistics on the uses of relationships. Table 2 outlines the detailed steps.

The last required stage of the bottom-up pragma-semantic approach is the description of the particular patterns of the prototype which is representative of the specific studied discursive genre.

6 Information collected thanks to TXM available at: http://textometrie.ens-lyon.fr/?lang=en

Table 2: Bottom-up pragma-semantic approach of corpus-based discourse organisation analysis.

Required step of the analysis	Researcher's actions
1. Text segmentation	Respecting the relevance and completeness constraint, the researcher determines the smallest optimal units of texts called 'micro-contribution'.
2. Micro-contribution relations tagging	While tagging the existing relations between the different micro-contributions and defining the structural constituency arrangements of the text (schemata), the researcher has to determine the different relationships and schemata applying Rhetorical Structure Theory.
3. Contributional sub-corpora	In order to prepare the multi-contributional level analysis, the researcher creates meso-contributional sub-corpora by gathering the micro-contributions of each schema, and the micro-contributional sub-corpora gathering the micro-contributions of each relationships.
4. Macro contribution analysis	At this level, the researcher gets information on the text organisation, on the relationships, and on the specific praxemes used in the text
5. Meso -contribution analysis	At this level, the researcher gets information on the schema organisation, and on the specific praxemes used in the different schemata.
6. Micro-contribution analysis	At this level, the researcher gets information on the specific praxemes used in the different micro-contributions relations.
7. Prototype patterns	Then, the researcher suggests the particular patterns of a prototype.

4 Experimental illustration

Applying the process to the previously introduced corpus, this section illustrates the different outcomes described at each stage. Table 3 shows the segmentation examples of French and Spanish Assisted Visit Conference and Guided Visited Conference corpora with their respective translations.

Once the text is segmented, the researcher selects a set of relationships and schemata that have been previously defined. They may evolve while he is tagging the existing relationships between the different micro-contributions as illustrated in Figure 5.

Table 3: Segmentation examples of the studied corpus.

Examples from original text	Translation[a]
Guided Visit Conference $_{f}$	
alors aujourd'hui donc nous retournons dans le même espace /	so today we go back to the same place /
nous laisserons la tapisserie bien sûr de côté /	we won't talk about the tapestry /
et nous allons nous intéresser aux autres œuvres qui sont présentées dans ce même espace /	and we are going to focus on the other masterpieces exposed in this place /
c'est un espace qui est consacré à l' Europe /	this place is dedicated to Europe /
le titre est Europe vers quinze cents /	the title is Europe around the fifteenth century /
donc nous allons retrouver aussi bien des œuvres <heu> de la toute fin du quinzième siècle /	so we are going to find masterpieces <err> to the tail end of the fifteenth century /
à partir des années quatorze cent soixante-dix /	from fourteen seventy /
Assisted Visit Conference $_{f}$	
les ducs mènent une existence rythmée par de somptueuses festivités /	the dukes live an existence cadenced by sumptuous festivities /
en témoignent nombre d'œuvres /	lots of masterpieces show it /
portraits en tenue raffinée /	portrait in sophisticated clothes /
bijoux /	jewels /
et accessoires de table /	and table accessories /
délicatement ciselés /	carefully chiselled /
ou ornés de matériaux précieux /	or ornamented with finery /
qui dénotent la recherche constante du luxe et du confort quotidien /	which evidence the constant research of luxury and daily comfort /
il existe de nombreux portraits des ducs de Bourgogne /	it exists numerous portrait of dukes of Burgundy /
ils attestent du besoin des ducs de se faire connaître /	they show the necessity the dukes had to be famous and present /
et d'assurer leur présence /	
Guided Visit Conference $_{e}$	
son tres pórticos que se ven neogóticos /	they are three neo-gothic porches /
miren sus formas /	look at their shapes /
¿ verdad ? /	you see ? /
es el principio de la construcción de Gaudí /	It's the beginning of the Gaudí's construction /
ahora vamos a concentrarnos muy bien en la parte central /	now we are going to be well focused on the central part /
es la más importante /	it's the most important /
es el nacimiento de Jesús /	it's the birth of Jesus /
ahí tenemos el nacimiento de Jesús /	there is the birth of Jesus /
los reyes magos /	the wise men /
los pastores /	the shepherds /

Table 3: (continued)

Examples from original text	Translation[a]
allí los tenemos /	here they are /
¿ hay algunas personas catalanes en este grupo por casualidad ? /	is there any Catalan in this group by sheer chance ? /
¿ No ? /	No ? /
bueno pues estos pastores van vestidos como catalanes /	well these shepherds are dressed like Catalan /
Guided Visit Conference $_{e}$	
cuando entramos en la basílica /	When we come in the basilica /
después de haber contemplado el mensaje bíblico de sus fachadas /	after contemplating the biblical message of its facades /
y nos hayamos en la nave central /	and reaching the central nave /
es cuando disfrutamos de un gran espacio /	we enjoy a large space /
donde la luz y el color lo invaden todo /	where light and colour invade everything /
es como una invitación a dejar fuera nuestras preocupaciones /	it's like an invitation to leave behind our preoccupations /

[a] The translations were done for the article and did not represent what a native English visitor service officer could have pronounced in such professional situations.

In the conducted research, the relationships and schemata are defined according to a specific format. For instance, the relationships and schemata of the previous example (Figure 5) are specified in the following manner:

Arrière-plan: (*Background*)	Constraints on the nucleus N: N is the described micro-contribution. The audience (A) cannot contextualise N before knowing S. Constraints on the satellite S: S contextualises N. Constraints on N + S: S increases the A's capability to understand and interpret N. Locutor's intention : L wants that the contextualisation of N increases the A's capability to understand and interpret N.
Démonstration: (*Evidence*)	Constraints on the nucleus N: N is a fact or a situation. Constraints on the satellite S: S is an explanation or an argument. Constraints on N + S: S evidences N. Locutor's intention: L produces S in order to convince A that N is true or real.
Élaboration: (*Elaboration*)	Constraints on the nucleus N: N is a fact or a situation. Constraints on the satellite S: S is an objective and complements information dealing with N. Constraints on N + S: S introduces a detail on N and S is linked to N according to one of these relationships: set / element, process / step, object / attribute, generalisation / detail. Locutor's intention: L wishes that A gets a better understanding of N thanks to the detailed information provided in S.

(continued)

Association: (*joint*)	Constraints on the nuclei (Ns): The Ns are independent elements gathered around a common concept. Locutor's intention: L wishes that A identifies an associative link between the micro-contributions dealing with a common concept.
Choix: (*Choice*)	Constraints on the nuclei (Ns): The Ns are independent options which can be selected by A. Locutor's intention: L wishes that A identifies the different suggested options without asking A to make a choice.
Inclusion: (*Embedded*)	Constraints on the nuclei (Ns): The Ns has to be parts of the same micro-contribution which is segmented so that one or various micro-contributions can be embedded as satellites to be linked with one of the Ns. Locutor's intention: L wishes that A identifies the extra embedded information linked to the segmented micro-contribution by a nucleus-satellite relationships.
Information: (*information*)	Constraints on the nuclei (Ns): The nuclei are linked by nucleus-satellite or multinuclear relations in order to communicate relevant information according to the developed topics.

In the analysis, the links with 19 nucleus-satellite relationships were tagged: *Background*, *Otherwise*, *Purpose*, *Cause*, *Circumstance*, *Concession*, *Condition*, *Evidence*, *Elaboration*, *Inquiry*, *Motivation*, *Means*, *Preparation*, *Restatement*, *Comment*, *Result*, *Summary*, and *Answer*; 7 multinuclear relationships: *Joint*, *Choice*, *Contrast*, *Embedded*, *List*, *Sequence*, *Simultaneity*, and 5 schemata: *Contact*, *Instruction*, *Information*, *Commentary*, *Reaction*. At the end of the tagging stage, the two first outcomes are the Rhetoric Structure Theory trees and the statistics concerning the relationships; however, for the multi-contributional level analysis it is necessary to build the contributional sub-corpora according the micro-contribution links. Here is an example of the sub-corpora elements according to Figure 5:

Arrière-plan: (*Background*)	les ducs mènent une existence rythmée par de somptueuses festivités
Démonstration: (*Evidence*)	en témoignent nombre d'œuvres qui dénotent la recherche constante du luxe et du confort quotidien
Élaboration: (*Elaboration*)	portraits en tenue raffinée bijoux et accessoires de table délicatement ciselés ou ornés de matériaux précieux

(continued)

Association: (*joint*)	portraits en tenue raffinée bijoux et accessoires de table
Choix: (*Choice*)	délicatement ciselés ou ornés de matériaux précieux
Inclusion: (*Embedded*)	en témoignent nombre d'œuvres qui dénotent la recherche constante du luxe et du confort quotidien
Information: (information)	les ducs mènent une existence rythmée par de somptueuses festivités en témoignent nombre d'œuvres portraits en tenue raffinée bijoux et accessoires de table délicatement ciselés ou ornés de matériaux précieux qui dénotent la recherche constante du luxe et du confort quotidien il existe de nombreux portraits des ducs de Bourgogne

As previously detailed, the analysis of the first outcomes and the obtained sub-corpora help us to describe the multilevel organisation of the studied discourse, and to suggest the final prototype patterns.

For instance, in the studied corpus at the macro-semantic level of the Assisted Visit Conference texts, the schema organisation follows this kind of structure: *Contact*, *Instruction*, series of *Information* / *Instruction*, *Contact*; the series of *Information* / *Instruction* may include the schema *Commentary*. At the meso-semantic level, the Assisted Visit Conference texts are descriptive and explicative due to a prevalence of relationships such as *Elaboration*, *Evidence*, *Comment*. Regarding the micro-semantic level, it is possible to compare the diversity of the employed praxemes, which was higher in the Assisted Visit Conference texts than in the Guided Visit Conference texts. At the same time, it is possible to apply a specific pattern analysis considering, e.g., the specificity score[7] (S) of the verb distribution in the different sub-corpora texts (cf. Table 4).

From such results, similarities and differences can be noted between the French and Spanish discourses, but also between assisted and guided visits: the

7 Score of a word being present f times in a sub-corpus of t tokens given that it appears a total of F times in a whole corpus of T tokens.

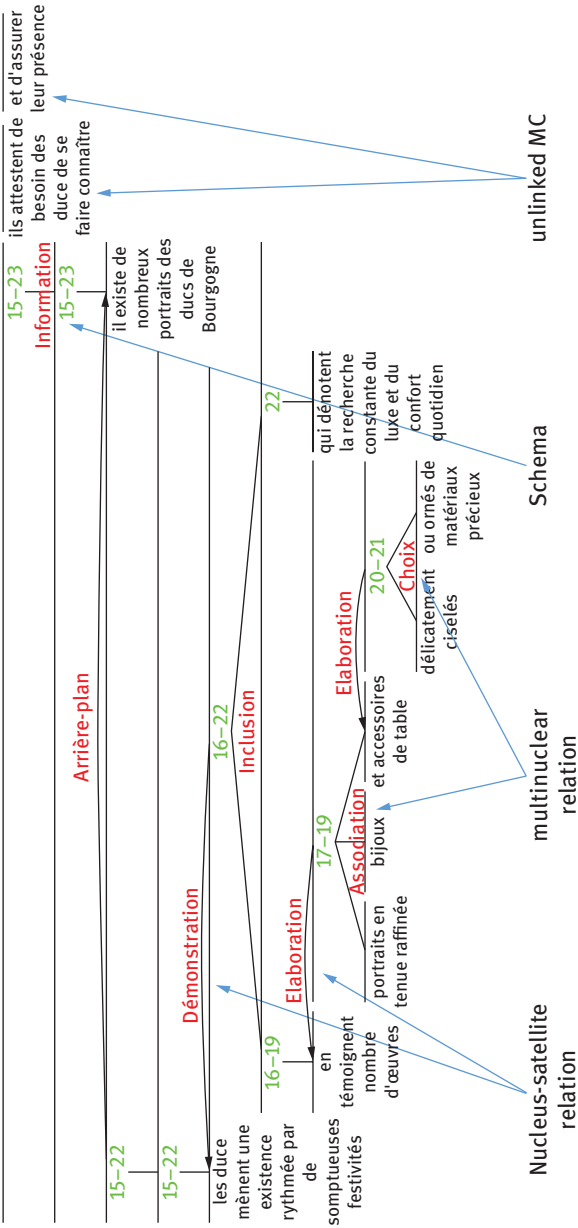


Figure 5: Process of micro-contribution relation tagging
Note: All the French micro-contributions come from the French Assisted Visit Conference of Table 3, where their translations can be read.

Table 4: French and Spanish sub-corpora specific verbs.

Guided Visit Conference $_f$		Assisted Visit Conference $_f$		Guided Visit Conference $_e$		Assisted Visit Conference $_e$	
Verbes (Verbs)	**S**	**Verbes (Verbs)**	**S**	**Verbes (Verbs)**	**S**	**Verbes (Verbs)**	**S**
aller (go)	71.1	peindre (paint)	31.6	ir (go)	36.5	encontrar (find)	15.6
avoir (have)	68.4	témoigner (prove)	14.9	decir (say)	20.2	destacar (highlight)	14.4
être (be)	48.3	attribuer (ascribe)	12.1	estar (be)	15.6	recorrer (go through)	11.7
dire (say)	43.7	former (form)	12.0	haber (hay) (there is/are)	14.0	realizar (make)	9.0
voir (see)	41.7	sembler (seem)	10.1	venir (come)	12.9	disfrutar (enjoy)	6.8
faire (do)	11.4	apparaître (appear)	8.3	ver (see)	12.6	levantar (raise / … up)	6.3
vouloir (want)	7.8	destiner (reserve)	8.2	tener (have)	10.8	conservar (preserve)	6.0
parler (talk)	6.8	découvrir (discover)	7.6	leer (read)	10.6	construir (build)	6.0
regarder (look)	6.1	jouer (play)	7.4	ocurrir (happen)	10.2	observar (observe)	5.6
souvenir (remind)	5.8	rendre (make / go)	7.3	mirar (look)	9.6	tratar (try)	5.3
falloir (must)	5.4	photographier (photography)	7.0	hablar (talk)	8.6	destinar (assign)	5.0
marier (marry)	5.0	manifester (occur)	6.8	saber (know)	7.9	comenzar (start)	4.7
amener (lead)	5.0	entourer (round)	6.7	ser (be)	7.2	situar (locate)	4.3
savoir (know)	4.4	évoquer (evoke)	6.4	hacer (do)	7.0	proceder (proceed)	4.1
penser (think)	3.4	dessiner (drawn)	6.3	preguntar (ask)	6.2	producir (produce)	3.9
adorer (love)	3.1	accentuer (focus)	6.2	oír (listen to)	5.9	disponer (dispose)	3.9
déguster (taste)	2.9	identifier (identify)	6.1	querer (want)	5.8	acceder (access)	3.6
concerner (concern)	2.8	montrer (show)	5.9	pasar (pass)	5.4	representar (represent)	3.6
commencer (start)	2.7	conserver (conserve)	5.7	dormir (sleep)	5.0	convertir (convert)	3.5
appeler (call)	2.7	orner (adorn)	5.6	gritar (shout)	4.9	cubrir (cover)	3.5

Assisted Visit Conference texts do not present any specificity score for *have*, *go* and *be*, while in the Guided Visit Conference texts these verbs appear to be the most specific, whether they are used as auxiliaries or not. However, the first Assisted Visit Conference specific verb in Spanish (*encontrarse*) refers to the concept of localization, which does not seem to be a concern in the French Assisted Visit Conference discourse. At this micro-semantic level, semantic orientations are noticeable: in the Guided Visit Conference, several verbs deal with motion (*go*, *come*, *lead*, *start*), senses and interactions (*see*, *look*, *taste*, *say*, *talk*, *ask*, *listen to*, *pass*), while in the Assisted Visit Conference, they deal with the masterpiece (*paint*, *form*, *photography*, *drawn*, *preserve*, *adorn*, *build*) and cognitive meanings (*ascribe*, *seem*, *reserve*, *discover*, *evoke*, *focus*, *identify*, *show*, *highlight*, *observe*, *assign*, *proceed*, *represent*). The Assisted Visit Conference discourse may be interpreted as more descriptive and explicative than the Guided Visit Conference discourse, which seems to be more experiential and emotional.

At the meso-contributional level, considering the schema *Information*, for instance, it presents a ternary organisation:

- Contextualisation: The locutor introduces the theme either describing the "scenery", or relying on shared knowledge in his description, or catching his audience's attention with a question.
- Theme utterance: Once the background is set, the locutor provides the main information.
- Details: Then, after the theme utterance, the locutor can develop the theme or start a reflexion to make the audience's interpretation more interactive.

This ternary structure appears in all corpora considered, whether Spanish or French; however the micro-semantic analysis shows a difference between the two corpora concerning the enunciator implication (Rabatel 2004). In the Assisted Visit Conference texts the enunciator is left behind in favour of the institutional discourse, while in the Guided Visit Conference texts the enunciator's discourse is as present as the institutional discourse. This result was inferred from the distribution of pronouns within the schemata.

Another interesting outcome at the meso-contributional level stems from the comparative analysis between the different schemata using, for instance, an exploratory technique such as a Factorial Correspondence Analysis based on the praxeme Frequencies[8] (cf. Figure 6). In this example, the relative inertia[9] of

8 For further details consult: http://documents.software.dell.com/Statistics/Textbook/correspondence-analysis

9 It represents the proportion of the total inertia accounted for by the respective part of speech.

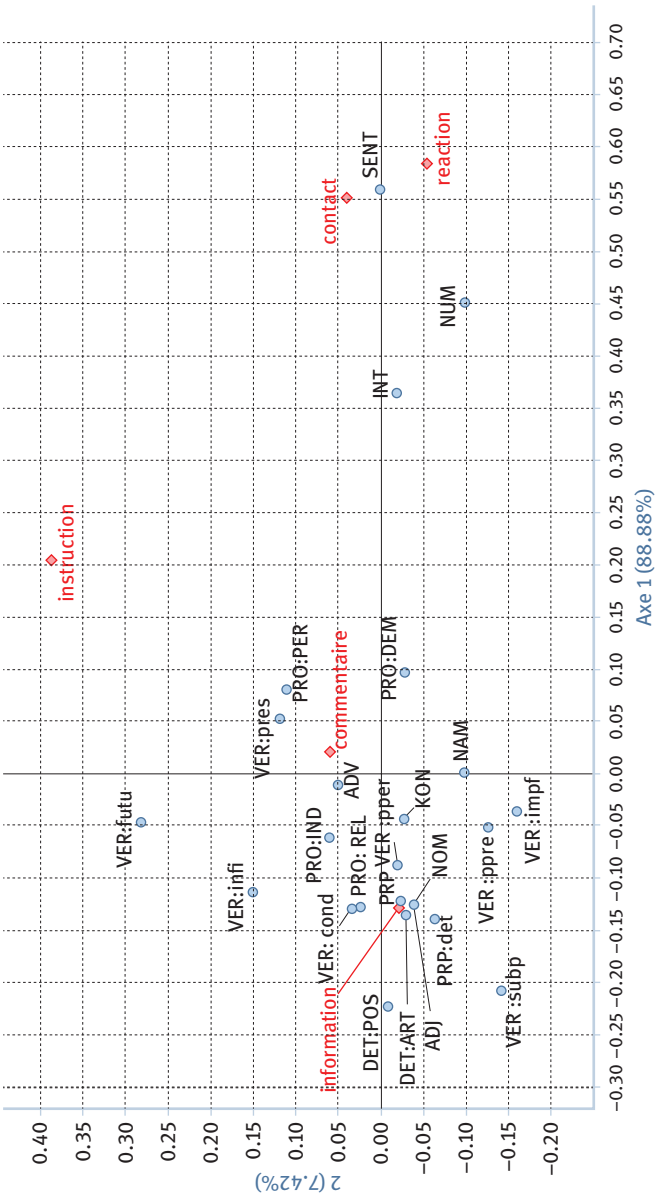


Figure 6: Factorial Correspondence Analysis of schemata based on part of speech

the part of speech in each schema shows that 88.88%[10] of the differences[11] exist between two groups, *Information*/*Commentary* and *Contact*/*Reaction* thanks to the relative inertia of the punctuation (SENT), the numeral adjective (NUM), and the interjections (INT). The second axis shows that 7.42% of the differences exist between the schema *Instruction* and the others mainly because of the relative inertia of the future verbs (VER:FUTU). This statistical technique helps us to evidence the specific feature of schemata in its context of use, since it provides the researcher with a compared micro-semantic analysis of all the schemata implied in the text organisation.

Finally, at the micro-contributional level, it is possible, for instance, to compare the different strategies used according to the discourse. Instead of comparing the sub-corpora regarding their modality or language, all the built-in sub-corpora of the different relations can be analysed using the same exploratory techniques. For instance, in the Assisted Visit Conference discourse, the locutors employ *afin de* (in order to), while in the Guided Visit Conference discourse, they mainly use *pour* (to) when they want to express the purpose. Thus, the micro-semantic features of each micro-contribution can be described according to the relationships they maintain with other micro-contributions.

Although the main purpose of this article is the presentation and the description of the bottom-up pragma-semantic approach, the principle features of a prototype can be outlined as illustrate in the seventh step of the method, as shown in Table 5.

As all the specific features are highlighted thanks to their frequencies of use, the different statistical tools, and the empirical bottom-up pragma-semantic approach, the final compilation of the different outcomes can only describe a prototype which represents the specific discourse organisation produced in a particular communicative situation. Thus, it embodies the specific analysed discursive genre in an empirical bottom-up approach.

5 Discussion

Two other analytical approaches present methodological similarities reaching comparable results. They also aim to describe the typical structural patterns of discourse in order to draw a specific prototype of a specifically studied commu-

10 Percent of inertia: inertia is defined as the total Chi-square divided by the total sum.

11 Differences are defined based on the deviations from the expected values.

Table 5: Main patterns of French Assisted Visit Conference text prototype.

	Text	Schema	Micro-contribution
Macro-semantic	The discourse presents a discrete organisation where each topic is confined between two schemata instructions following this sequence pattern: *Contact, Instruction,* series of *Information* / *Instruction, Contact*. The series of *Information* / *Instruction* may include the schema *Commentary*.		
Meso-semantic	At this level, the micro-contribution analysis shows the genre is descriptive and explicative due to a prevalence of relationships such as *Elaboration*, *Evidence*, and *Comment*. The audience receives a high amount of encyclopaedic data in accurate language without any hesitations or reformulations.	For example, the schema *Information* presents a ternary sequence including contextualisation, theme utterance, and details. *Information* and *Commentary* present comparable features different from *Contact* and *Reaction* which appear to be monological simulated interactions.	
Micro-semantic	In the AVC, the verbs show that the discourse deals with the concept of a masterpiece (*paint*, *form*, *photography*, *drawn*, *preserve*, *adorn*, *build*) and cognitive meaning (*ascribe*, *seem*, *reserve*, *discover*, *evoke*, *focus*, *identify*, *show*, *highlight*, *observe*, *assign*, *proceed*, *represent*)	In the schema *Information*, the enunciator is left behind in favour of the institutional discourse. The spatial deictic centre is permanently focused on the point from where the audience is listening to the message. The temporal deictic centre is regularly updated to move from virtual to real world and vice-versa.	Use of the third person with impersonal structure strengthening the discourse objectivity even if an institutional subjectivity is observed in the lexical choice.

nicative situation. On the one hand, the Biber/Connor/Upton Approach (Biber, Connor, and Upton 2007) defines a top-down corpus-based analysis of discourse organization and is grounded on the *Move analysis* previously introduced by Swales (1981, 1990). The *Move analysis* approach suggests that text segmentation is a necessary stage in the analysis in order to grasp the general text organisation through the particular communicative functions (Upton and Cohen 2009: 4) that each move represents. Thus, the typical *Move* structure pattern (Upton and Cohen 2009: 15) helps the researcher to suggest the specific prototype. However, what the bottom-up pragma-semantic approach presents another angle for analysis: in the "Move" approach, the set of the major communicative functions of each discourse unit (Biber, Connor, and Upton 2007) is determined *a priori* by the researcher (top-down approach); while in the bottom-up pragma-semantic approach the schemata, which could be considered equivalents of Biber's set, stems from the empirical segmentation and the Rhetorical Structure Theory tagging stages (bottom-up approach). Therefore, the text segmentation is independent of the definitions of schemata and relations; it only depends on the completeness and relevant cognitive constraints. Furthermore, the Rhetorical Structure Theory tagging process guarantees a bottom-up text organization discovery at the different levels of contribution, while the *Move* analysis approach identifies the functional type of each discourse applying the analytical framework of Biber's communicative functions, which have been previously introduced to develop the analytical framework.

Yet another discourse analysis approach, the Contributional Approach, is based on the well-known contribution concept employed by Grice since the publication of his conversational maxims, although he did not provide a detailed description of what he considered a contribution. Português (2011), and Borderieux (2016), from a post-Gricean point of view, define the contribution as a set of complete and autonomous utterances, and they consider that the Contributional Approach helps one to analyse the text at various levels whether it be a sentence, a paragraph, or a chapter, in order to describe a text as a sequence of argumentative contributions (Borderieux 2016: 6). They also highlight the importance of logical links which mutually tie the contributions together to build textual coherence. They aim to expand on their contributional theory by applying their approach to other types of text in order to suggest rules that could explain the contributional organisation of a text (Borderieux 2016: 11). Borderieux himself argues that is easier to evidence these rules in highly normed texts such as patent texts, and later apply them to more complex texts. However, with a definition of contribution based on utterance, which is itself based on sentences, it might be difficult to generalise the contributional approach outcomes to more complex texts, especially if they are produced orally. In the bottom-up pragma-semantic approach, the

contribution is a complete and autonomous set of micro-contributions which are independent of all oral or scriptural conventions, which allows for the segmentation of all types of text and may help to expand this approach to more complex texts. Furthermore, revealing the text organisation from the micro-contribution level also reveals the text coherence construction by naming and analysing the relationships thanks to the Rhetorical Structure Theory. Therefore, taking apart the text in micro-contributions is the first necessary stage to enable the researcher to discover, step by step, the structural organisation of the text.

6 Conclusion

The main characteristics of a bottom-up textual analysis grounded in the cognitive constraints of relevance and completeness, which ensure the pragma-semantic dimension, have been introduced. The bottom-up pragma-semantic approach process is divided into seven stages, which provides researchers with an empirical method of text analysis where three semantic levels are considered in order to account for the different levels of the text interpretation. The coherence of each semantic level (macro-semantic, meso-semantic, and micro-semantic) is built upon the relevance and the completeness of its contributional units, as well as the relationships they maintain with one another. Once the researcher has gone through the seven stages, he may suggest a prototype representing the specific features of the considered communicative situation.

The bottom-up pragma-semantic approach has been drawn from theoretical background according to the optimal text unit, named micro-contribution, and the discursive theoretical model of its production. A methodological framework has also been suggested based on semantic and contributional levels, where the researcher can systematically make an analysis of each contribution according to its different semantic levels.

Finally, the different steps and exploratory techniques of the bottom-up pragma-semantic approach have been illustrated by presenting some outcomes of the analysis I have conducted on a corpus compiled from socio-professional discourses (Méric 2016).[12] The selected examples demonstrated how, applying a bottom-up approach, the coherence micro-contribution relationships reveal the structures of the different levels of coherence implied in a specific discourse

12 This study does not aim to provide the whole specific features of the prototype representing the considered communicative situation.

organisation. In the future, it would be beneficial if other particular communicative situations were analysed. It is my hope that even if this contribution does not launch a new theory, it will at the very least provide the necessary basis for the further development of a new text analysis approach grounded on a replicable empirical bottom-up process, in order to evidence the architecture of various text genres.

References

Adam, Jean-Michel. 2011. *Genres de récits. Narrativité et généricité des textes*. Paris: L'Harmattan.

Adam, Jean-Michel. [2008] 2014. *La linguistique textuelle*. Cursus Lettres. Paris: Armand Colin. 3rd ed.

Bateman, John A. & Rondhuis, Klaas Jan. 1997. Coherence relations: Towards a general specification. *Discourse Processes* 24(1), 3–49.

Benveniste, Emile. 1966. *Problèmes de linguistique générale*. Paris: Gallimard.

Biber, Douglas, Connor, Ulla, & Upton, Thomas Albin. 2007. *Discourse on the move: Using corpus analysis to describe discourse structure*. John Benjamins Publishing.

Bloomfield, Leonard. 1949. A set of postulates for the Science of Language. *International Journal of American Linguistics* 15(4), 195-202.

Borderieux, Julien. 2016. L'approche contributionnelle. *SHS Web of Conferences* 27, 06002. EDP Sciences.

Charaudeau, Patrick. 1995. Une analyse sémiolinguistique du discours. *Langages* 29, 96–111.

Chomsky, Noam. 1965. *Aspect of the Theory of Syntax*. Cambridge: Massachusetts Institute of Tech Cambridge Research Lab of Electronics.

Chomsky, Noam. 1966. *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. New York: Harper & Row.

De Beaugrande, Robert A. 1995. Text linguistics. In Verschueren, J. e.a. (eds.), *Handbook of Pragmatics, Manual*, 536–544. Amsterdam & Philadelphia: John Benjamins.

Fairclough, Norman. 1992. *Language and power*. London: Longman.

Halliday, Michael Alexander Kirkwood. 2006. *Linguistic studies of text and discourse* (Vol. 2). London & New York: Continuum.

Mann, William C. & Thompson, Sandra A. 1987. Rhetorical structure theory: A theory of text organization. In *ISI/RS*, 87–190. Los Angeles: University of Southern California, Information Sciences Institute.

Mann, William C. & Thompson, Sandra A. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse 8*(3), 243–281.

Mann, William C., Matthiessen, Christian M. I. M. & Thompson, Sandra A. 1992. Rhetorical structure theory and text analysis. *Discourse description: Diverse linguistic analyses of a fund-raising text*, 39–78.

Mann, William C. & Taboada, M. 2005–2015. *Relations definitions*. http://www.sfu.ca/rst/01intro/definitions.html (accessed 27 May 2017)

Méric, Olivier. 2016. *Organisation discursive de la visite médiée de sites touristiques : théorisation contributionnelle et valorisation d'une praxis professionnelle*. Dijon: University of Burgundy PhD Thesis.

O'Donnell, Michael. 2003. *Dealing with relations*. http://www.wagsoft.com/RSTTool/section6.html (accessed 27 May 2017).
Portuguès, Yann. 2011. *Contribution à une théorie linguistique du texte : la complétude textuelle comme heuristique*. Orléans: University of Orléans PhD Tesis.
Rabatel, Alain. 2004. L'effacement énonciatif dans les discours rapportés et ses effets pragmatiques. *Langages* 4, 3–17.
Reboul, Anne & Moeschler, Jacques. 2005. *Pragmatique du discours: De l'interprétation de l'énoncé à l'interprétation du discours*. Paris: Armand Colin.
Roulet, Eddy. 1986. Complétude interactive et mouvements discursifs. *Cahiers de linguistique française* 7, 193–210.
Saussure, Ferdinand. [1916] 1995. *Cours de linguistique générale*. Paris: Editions Payot & Rivages.
Sinclair, John McHardy. 1996. Preliminary recommendations on corpus typology. *EAGLES Document TCWG-CTYP/P*, http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html (accessed 27 May 2017).
Sperber, Dan & Wilson, Deirdre. 1986. *Relevance: Communication and cognition*. Oxford: Blackwell.
Stubbs, Michael. 1983. *Discourse Analysis: The Sociolinguistic Analysis of Natural Language*. Chicago: University of Chicago Press.
Swales, John. M. 1981. *Aspects of article introductions*. Language Studies Unit, Birmingham: University of Aston.
Swales, John M. 1990. *Genre analysis: English in academic and research settings*, Cambridge: Cambridge University Press.
Taboada, Maite & Mann, William C. 2006a. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies* 8(3), 423–459.
Taboada, Maite & Mann, William C. 2006b. Applications of rhetorical structure theory. *Discourse studies* 8(4), 567–588.
Upton, Thomas Albin & Cohen, Mary Ann. 2009. An approach to corpus-based discourse analysis: The move analysis as example, *Discourse Studies* 11(5), 585–605.
Werlich, Egon. 1976. *A Text Grammar of English*, Heidelberg: Quelle & Meyer.

Ekaterina Lapshinova-Koltunski and Marcos Zampieri

Linguistic features of genre and method variation in translation: a computational perspective

Abstract: In this contribution we describe the use of text classification methods to investigate genre and method variation in an English – German translation corpus. For this purpose we use linguistically motivated features representing texts using a combination of part-of-speech tags arranged in bigrams, trigrams, and 4-grams. The classification method used in this study is a Bayesian classifier with Laplace smoothing. We use the output of the classifiers to carry out an extensive feature analysis on the main difference between genres and methods of translation.

1 Introduction

In the present study, we use text classification techniques to explore variation in translation. We analyse the interplay between two dimensions influencing this variation: translation methods (human and machine translation) and text registers or genres (e.g. fiction, political speeches, etc.). Our starting assumption is that the interplay between these dimensions is reflected in the lexico-grammar of translated texts, i.e. in their linguistic features. Our assumption here is that genres and methods represent two dimensions influencing linguistic properties of translations, and can thus be confounding in a specific task. For instance, if we want to automatically distinguish between human and machine translation, we need to exclude features which are rather genre-specific as they can compromise the results of the classification.

In our previous work, see e.g. Lapshinova-Koltunski (2017), we used a set of features derived from theoretical frameworks, such as genre / register theory, e.g. Halliday & Hasan (1989), Biber (1995), Neumann (2013), or translationese studies, e.g. Baker (1993), Baroni & Bernardini (2006), Volansky et al. (2011). In the present analysis, we use a data-driven approach, which will help us to dis-

Ekaterina Lapshinova-Koltunski, Saarland University
Marcos Zampieri, University of Cologne

https://doi.org/10.1515/9783110595864-005

cover new language structures reflecting variation in translation. Classification techniques will help us to identify discriminative features of the two variation dimensions under analysis. For this, we train classifiers to distinguish translated texts according to either their register or method of translation, using the VARTRA corpus (Lapshinova-Koltunski, 2013), a collection of English to German translations. Our assumption is that text classification methods can level out discriminative features of different translation varieties that intuition alone cannot grasp; thus enabling us to investigate in more detail the properties of each of them. More than the classification results *per se*, we use level out interesting linguistic features that can be further used in linguistic analysis and NLP applications.

Text classification methods have been applied in a wide range of tasks such as spam detection (Medlock, 2008), native language identification (NLI) (Gebre et al., 2013) temporal text classification (Niculae et al., 2014), and the identification of lexical complexity in text (Malmasi et al., 2016). In the aforementioned studies, researchers are interested in how well classification methods can perform or, in other words, how reliably these methods are able to attribute correct labels to a set of texts. Therefore, most researchers in text classification are concerned in exploring features and algorithms that deliver the best performance for each task. In recent works by Diwersy et al. (2014) and Zampieri et al. (2013), however, text classification methods were proposed to investigate language variation across corpora (e.g. diatopic and dialectal variation) using linguistically motivated features.

In this contribution, we propose an approach to automatically classify translated texts regarding genre and method of translation. We are interested not only in obtaining state-of-the-art classification performance, but also in leveling out interesting linguistic features from the data. The features used here constitute combinations of part-of-speech (POS) tags. These POS combinations represent, however, language patterns, e.g. a finite auxiliary verb followed by a participle represents a verbal phrase in a passive voice. Analyzing sets of the POS combinations that result from the classification experiments, we try to identify those that are specific for the classes under analysis.

This study is structured as follows: the following section presents the theoretical background, as well as the related work. Here, we also describe our previous experiments on text classification for the analysis of translation variation. In Section 3, we introduce the dataset, as well as the methodology applied for the analysis. Section 4 presents the results of the classification. We further investigate the results in Section 5, in which we concentrate on the analysis of features specific for translation methods and genres. Section 6 summarizes the findings and presents a discussion on the related issues.

2 Theoretical background and related work

2.1 Theoretical background

Translation is influenced by several factors, including the source and the target language, registers or genres a text belongs to, as well as the translation method involved. Since the present study focuses on genre and method variation, we will also base our research on the studies related to this type of variation.

Genre-specific variation of translation is related to studies within register and genre theory, e.g. Halliday & Hasan (1989), Biber (1995), which analyse contextual variation of languages. In the present contribution, we use the term **genre** and not **register**, although they represent two different points of view covering the same ground, see e.g. Lee (2001), and we use the latter in our previous studies, see e.g. Lapshinova-Koltunski (2017) and Lapshinova-Koltunski & Vela (2015). Mostly, we refer to genre when speaking about a text as a member of a cultural category, about a register when we view a text as language. However, in this study we consider both as lexico-grammatical characterisations, conventionalisation and functional configuration determined by a context use. The differences between genres can be identified through a corpus-based analysis of phonological, lexico-grammatical and textual (cohesion) features in these genres; see the studies on linguistic variation by Biber (1995) or Biber et al. (1999), and linguistic variation among genres can be traced in the distribution of these features. Multilingual studies concern linguistic variation across languages, comparing genre and register settings specific for the languages under analysis, e.g. Biber (1995) on English, Nukulaelae Tuvaluan, Korean and Somali, and Hansen-Schirra et al. (2012) and Neumann (2013) on English and German. The latter two also consider this type of linguistic variation in translations. Other translation scholars e.g. Steiner (2004) and House (2014), also pay attention to genre and register variation when analysing language in a multilingual context of translation. However, they either do not account for the distributions of the corresponding features, or analyse individual texts only. In the works by De Sutter et al. (2012) and Delaere & De Sutter (2013), register-related differences are also described for translated texts. Yet, these differences are identified on the level of lexical features only.

The features that are most frequently used in studies on variation in corpus-based approaches are of shallow character and include lexical density (LD), type-token-ratio (TTR), and part-of-speech (POS) proportionality. Steiner (2012) uses these features to characterise profiles of various subcorpora distinguished by language (English and German), text production type (translation and original) and eight different registers. The author defines a number of contrast types

including register controlled ones which implies (1) contrasts within one register between English and German, and (2) contrasts between registers within each of the languages, see Steiner (2012, p. 72). In our analysis, we consider genre variation only within translations.

Applying a quantitative approach, Neumann (2013) analyses an extensive set of linguistic patterns reflecting register variation and shows the differences between the two languages under analysis. The author also demonstrates to what degree translations are adapted to the requirements of different registers, showing how both register and language typology are at work. Kunz et al. (2017) show that register variation is also relevant for a number of textual phenomena. They analyse structural and functional subtypes of coreference, substitution, discourse connectives and ellipsis on a dataset of several registers in English and German. They are able to identify contrasts and commonalities across the two languages and registers with respect to the subtypes of all textual phenomena under analysis. The authors show that these languages differ as to the degree of variation between individual registers in the realisation of the phenomena under analysis, i.e. there is more variation in German than English. They attest the main differences in terms of preferred meaning relations: a preference for explicitly realising logico-semantic relations by discourse markers and a tendency to realise relations of identity by coreference. Interestingly, similar meaning relations are realised by different subtypes of discourse phenomena in different languages and registers.

Whereas attention is paid to genre settings in human translation analysis, they have not yet been considered much in machine translation. There exist some studies in the area of statistical machine translation (SMT) evaluation, e.g. errors in translation of new domains (Irvine et al., 2013). However, the error types concern the lexical level only, as the authors operate solely with the notion of domain and not genre. Domains represent only one of the genre parameters and reflect what a text is about, i.e. its topic, and further settings are thus ignored. Although some NLP studies, e.g. those employing web resources, do argue for the importance of genre conventions, see e.g. Santini et al. (2010), genre remains out of the focus of machine translation. In the studies on adding in-domain bilingual data to the training material of SMT systems (Wu et al., 2008) or on application of in-domain comparable corpora (Irivine & Callison-Burch, 2014), again, only the notion of domain is taken into consideration.

Variation in terms of translation method has not received much attention so far. There are numerous studies in the context of NLP that address both human and machine translations (Papineni et al., 2002; Babych et al., 2004). Yet they all serve the task of automatic MT system evaluation and focus solely on translation error analysis, using human translation as a reference in the evaluation of

machine translation outputs. Evaluations serves the task to prove to what extent automatically translated texts (hypothesis translations) comply with the manually translated ones (reference translations). The ranking of machine-translated texts is based on scores produced with various metrics. The metrics applied in the state-of-the-art MT evaluation are automatic and language-independent: BLEU and NIST (Doddington, 2002). However, since they do not incorporate any linguistic features, BLEU scores need to be treated carefully, which was demonstrated by Callison-Burch et al. (2006). This fact has been advancing the development of new automatic metrics, such as METEOR (Denkowski & Lavie, 2014), Asiya (González et al., 2014) and VERTa (Comelles and Atserias, 2014). They incorporate lexical, syntactic and semantic information into their scores. The accuracy of the evaluation methods is usually proven through human evaluation. More specifically, the automatically provided scores are correlated with the human judgements which are realised by ranking MT outputs (Bojar et al., 2014; Vela and van Genabith, 2015 and others). Some of the existing metrics incorporate linguistic knowledge.

There are even more works on MT evaluation that operate with linguistically-motivated categories, e.g. Popovic and Ney (2011) or Fishel et al. (2012). However, none of them provides a comprehensive analysis of the differences between human and machine translation in terms of specific linguistically motivated features. In fact, the knowledge on the discriminative features of human and machine translation can be derived from the studies operating with machine learning procedures for MT evaluation, such as Stanojević and Simaán (2014) or Gupta et al. (2015). Corston-Oliver et al. (2001) use classifiers that learn to distinguish human translations from machine ones. These classifiers are trained with various features including lexicalised trigram perplexity, part of speech trigram perplexity and linguistic features such as branching properties of the parse, function word density, constituent length, and others. Their best results are achieved if perplexity calculations were combined with finer-grained linguistic features. Their most discriminatory features that differentiate between human and machine translations are not just word n-grams. They include the distance between pronouns, the number of second person pronoun, the number of function words, and the distance between prepositions.

Volansky et al. (2011) operate with translationese-inspired features, and are able to distinguish between manual and automatic translations in their dataset with 100% accuracy. However, the manual and automatic translations they are using have different source texts. We believe that the distinction they are able to achieve is not the distinction between translation methods, but rather between different underlying texts, since their most discriminatory features are the ones that show good performance in any text classification task (token n-grams). El-Haj et al. (2014) make use of readability as a proxy for style and analyse consistency

in translation style considering how readability varies both within and between translations. They compare Arabic and English human and machine translations of the originally French novel "The Stranger" (French: *L'Étranger*). The results show that translations by humans (both male and female) are closer to each other than to automatic translations. The authors also measure closeness of translations to the original in terms of the selected measures, which should serve as an indicator of translation quality.

To the best of our knowledge, there have not been many studies published about the interplay between the two dimensions influencing translation that are in focus of our study. Kruger and van Rooy (2012) try to answer the question on the relationship between register and the features of translated language. Their hypothesis was that the translation-related features would not be strongly linked to register variation suggesting that in translated text reveal less register variation, or sensitivity to register, which is a consequence of translation-specific effects. However, their findings provide limited support for this hypothesis. They state that the distribution and prevalence of linguistic realisations of the features of translated language may vary according to register. Therefore, the concept of translated language should be more carefully analysed and defined in terms of registers (Kruger and van Rooy, 2012, p. 61–62). Jenset and McGillivray (2012) analyse the interaction between registers, source language and translators' background on the basis of morphological features. The interaction between the dimensions of register, author and translator was also analysed by Jenset and Hareide (2013) who use patterns of sentence alignment as features. Thus, there is no comprehensive description of the linguistic features that represent the dimensions of translation variation. We analyse the interplay between the dimensions of genre (register) and method trying to detect specific features that reflect this interplay.

2.2 Previous experiments

In our previous analyses (Zampieri & Lapshinova, 2015), we applied text classification methods on a set of English-German translations. We used two different sets of features: n-grams taking 1) all word forms into account and 2) semi-delexicalized text representations – all the nouns were replaced with placeholders, which represented the novelty of that approach. Our task was two-fold: (a) to discriminate between different genres (fiction, political essays, etc., a total of seven classes); and (b) to discriminate between translation methods (human professional, human student, rule-based machine and two statistical systems – five classes).

We performed several classification tasks: (1) We use word n-grams to train five translation method and seven genre classes; (2) We use delexicalised n-grams to classify four and five translation method classes and seven register classes; (3) We use delexicalised n-grams to classify between human and machine translation.

The results of the first experiment show that the classifier performs better for the distinction of genres than of translation methods (F-measure of 57.30% and 35.30% respectively). This result is not surprising, as content words (including proper nouns) are domain specific and, therefore, the classifier can better differentiate between genres that vary in their domains. The results of this experiment shows that it is important to use (semi-)delexicalised features in a dataset that represents both dimensions of variation in translation – genre and method.

The results of the second experiment show that delexicalised features reduce the performance of the genre classifier (from 57.30% to 45.40%) but increase the performance of the translation method classifier (from 35.30% to 43.10%), especially if we reduce the dataset to four classes instead of five (we concatenate both statistical machine translation outputs). The results of this experiment confirm the importance of (semi-)delexicalised features, as we achieve similar scores for the analysis of both dimensions of variation in our data.

In the last experiment, we reduce the number of translation method classes to two – human and machine. This classification is less fine-grained and represents manual and automatic procedures of translation. As expected, this experiment delivers better classifier results: up to 60.5% F-measure in distinguishing between manually and automatically translated texts.

In the last step, we performed qualitative analysis of the output features paying attention to those which turned to be most informative for the corresponding classification task. In this way, we were able to identify a set of semi-delexicalized n-grams that are discriminative for either certain genres or translation methods in our data. This step was manual and included evaluation of trigrams only, as the performance of trigram models achieved the best results in the classification task. We generated two lists of features specific either to human or machine translation, and fourteen lists of features discriminating genre pairs. The features informative for translation method included full nominal phrases that differentiated in the type of determiners (articles in human and possessives in machine translations), personal pronouns expressing coreference that differentiated in the grammatical number (singular in human and plural in machine translations), event anaphors that differentiated in the type of pronouns (demonstrative in human and personal in machine), etc. These features differed from those specific for genre identification. For instance, for the discrimination between political essays ad fictional texts, discourse markers expressing different

relations turned to be informative. Moreover, the lists included also features related to verbal phrases, e.g. passive vs. active voice, infinitives and modal verbs differing in their meaning.

In this study we base upon the results of this analysis: we use a two-member classification of translation methods (manual and automatic) and seven-member classification for genres. Moreover, we decide to fully delexicalise the features and run our experiments on delexicalised features instead of semi-delexicalised ones.

3 Methods

3.1 Data

For the purpose of our study we looked for suitable translation corpora containing different genres and methods of translation. The only corpus known to us that possesses these characteristics is VARTRA (Lapshinova-Koltunski, 2013). VARTRA comprises multiple translations from English into German. These translations were produced with five different translation methods as follows: (1) human professionals (PT1), (2) human student translators (PT2), (3) a rule-based MT system (RBMT), (4) a statistical MT system trained with a large quantity of unknown data (SMT1) and (5) a statistical MT system trained with a small amount of data (SMT2).

VARTRA contains texts from different genres, namely: political essays (ESS), fictional texts (FIC), instruction manuals (INS), popular-scientific articles (POP), letters of share-holders (SHA), prepared political speeches (SPE), and touristic leaflets (TOU). Each sub-corpus represents a translation variety, a translation setting which differs from all others in both method and genre (e.g. PT1-ESS or PT2-FIC, etc.). The corpus is tokenised, lemmatised, tagged with part-of-speech information, segmented into syntactic chunks and sentences. The annotations were obtained with Tree Tagger (Schmid, 1994).

Before classification was carried out, we split the corpus into sentences.[1] The length of each sentence varies between 12 and 24 tokens. This results in a dataset containing 6,200 instances.

1 The decision to split the corpus into sentences was motivated by the amount of texts available in the VARTRA corpus. Splitting the corpus into sentences generated enough data points for text classification and made the task more challenging.

The features used in the experiments we report in this study were based on the combinations of POS tags arranged in form of bag-of-words (BoW), bigrams, trigrams, and 4-grams.[2] In Example (1), we illustrate the representation of the sentences in the corpus. (1-a) represents a sentence from the corpus, (1-b) shows the representation, where all nouns are substituted with the placeholder *PLH* resulting in what we call a semi-delexicalized text representation.

(1) a. *Die weltweiten Herausforderungen im Bereich der Energiesicherheit erfordern* über *einen Zeitraum von vielen Jahrzehnten nachhaltige Anstrengungen auf der ganzen Welt*.
b. *Die weltweiten PLH im PLH der PLH erfordern* über *einen PLH von vielen PLH nachhaltige PLH auf der ganzen PLH*.
c. ART ADJA NN APPRART NN ART NN VVFIN APPR ART NN APPR PIAT ADJA ADJA NN APPR ART ADJA NN.

This type of representation lies between fully delexicalized representations, such as the one proposed by Diwersy et al. (2014) for the study of variation in translation and diatopic variation of French texts, and the fully lexicalized representation, common in most text classification experiments, which uses all words in text without any substitution. This representation minimizes topic variation. Previous studies have shown that named entities significantly influence the performance of text classification systems (Zampieri et al., 2013; Goutte et al., 2016). We used this representation in our previous experiments that we describe in Section 2.2 above.

In (1-c), we use a fully delexicalized structure representing texts only using the POS annotation available at the VARTRA corpus. Zampieri et al. (2013) show that classification experiments using POS and morphological information as features can not only be linguistically informative, but also achieve good performance in discriminating between texts written in different Spanish varieties. Therefore, we use this representation to test whether this is also true for translated texts. The underlying tagset used in TreeTagger is "Stuttgart/Tübinger Tagsets" (STTS)[3], one of the commonly used tagsets for German. In Table 1, we illustrate a segment from the corpus with an explanation of selected tags.

2 Note that in this study we make a clear distinction between BoW and unigrams. The BoW models used in this study do not comprise any smoothing method, whereas the *n*-gram models are calculated using Laplace smoothing.

3 http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html

Table 1: Illustration of the STTS-tagged corpus segment.

word	POS	category
Die	ART	article
weltweiten	ADJA	adjective
Herausforderungen	NN	common noun
im	APPRART	preposition+article
Bereich	NN	common noun
der	ART	article
Energiesicherheit	NN	common noun
erfordern	VVFIN	full finte verb
über	APPR	preposition
einen	ART	article
Zeitraum	NN	common noun

The decision to use these features was motivated by our goal of investigating translation variation influenced by both genre and method, and our aim to obtain a classification method that could perform well on different corpora by capturing structural differences between these translation varieties.

3.2 Algorithm

In our experiments we use a Bayesian learning algorithm similar to Naive Bayes entitled Likelihood Estimation (LE) and previously used for used for language identification by Zampieri & Gebre (2012,2014). Just like Naive Bayes classifiers, LE works based on an independence assumption that the presence of a particular feature of a class is not related to the presence of any other feature. The independence assumption makes the algorithm extremely fast and a good fit for text classification tasks. Bayesian classifiers are inspired by Bayes theorem represented by the following equation.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \tag{1}$$

Where $P(A|B)$ is a conditional probability of A given B. Using the notation by Kibriya et al. (2004), a Naive Bayes classifier computes class probabilities for a given document and a set of classes C. It assigns each document to the class with the highest probability P (c|ti).

$$P(c|t_i) = \frac{P(t_i|c)\,P(c)}{P(t_i)} \tag{2}$$

LE calculates a likelihood function on smoothed n-gram language models. Smoothing is carried out using the Laplace smoothing calculated as follows:

$$P_{lap}(w_1...w_n) = \frac{C(w_1...w_n) + 1}{N + B} \tag{3}$$

The language models can contain characters and words (e.g. bigrams and trigrams), linguistically motivated features such as parts-of-speech (POS) or morphological categories such as the one used in Zampieri et al. (2013) for the study of diatopic variation. In this contribution LE is used with POS tags as features.

Models are first calculated for each particular class in the dataset. Subsequently LE calculates the probability of a document belonging to a given class. In our case classes are represented either by genres or method of translation. The function that calculates the probability of a document given a class, represented by L (language model) is the following:

$$P\,(L|text) = \arg\max_{L} \sum_{i=1}^{N} \log P(n_i|L) + \log P(L) \tag{4}$$

Where N is the number of n-grams in the test text. The language model L with the highest probability determines the predicted class of each document.

4 Classification results

In this section, we present the results obtained in various classification experiments using a Bayesian classifier. To evaluate the performance of the classifiers we used standard metrics in text classification such as precision, recall, f-measure, and accuracy. The linguistic analysis and discussion of the most important differences between both method and genre variation will be presented later in Section 5.

4.1 Translation methods: human vs. machine

In this first experiment we investigate differences between translation methods. The VARTRA corpus divides translation methods into five categories, three representing automatic methods and two containing translations produced by humans.

Table 2: N-grams: Human x Machine.

Features	Precision	Recall	F-Measure
bigrams	60.70%	60.51%	60.61%
trigrams	62.50%	62.50%	62.50%
4-grams	57.84%	57.25%	57.54%

In Zampieri and Lapshinova-Koltunski (2015) we trained a classifier to discriminate between these five methods and we observed that variation was more prominent when comparing human vs. machine translations. For this reason we unify PT1 and PT2 into one class and RBMT, SMT1, and SMT2 into the other.

We represent texts using the POS tags as features as presented earlier in this chapter. We use a total of 600 texts for each class split in 400 documents for training and 200 for testing. Results are presented in terms of precision, recall, and f-measure in Table 2. The baseline is 50% accuracy.

In all three settings, the model performs above the expected baseline of 50.0% f-measure. The best performance is obtained using a POS trigram model (62.5% f-measure and precision).

4.2 Genres

In this section, we train a model to automatically distinguish between seven different genres represented in our dataset. For the sake of clarity we list here the genres contained in the VARTRA corpus: political essays (ESS), fictional texts (FIC), instruction manuals (INS), popular-scientific articles (POP), letters of share-holders (SHA), prepared political speeches (SPE), and touristic leaflets (TOU). All experiments in this section are binary classification settings in which the classifier is trained to discriminate between two genres at a time. The baseline four each setting is therefore 50% accuracy.

We again use POS tags as features as described in 4.1 arranged in bigrams, trigrams, and 4-grams. We use a total of 500 texts for each class split in 300 documents for training and 200 for testing. We evaluate the performance of our method in terms of accuracy and present results in Tables 3, 4, and 5.

In Table 3 using POS bigrams we observed that the best results were obtained when discriminating instruction manuals from fictional texts, 81.25% accuracy. The worst results were obtained between speech and essays, 61.25% accuracy.

Corroborating with the findings of the previous section, we observed that for genres the overall best results are obtained when using POS trigrams. The model

Table 3: Genres Classification in Translation in Binary Settings: POS Bigrams.

Classes	ESS	FIC	INS	POP	SHA	SPE	TOU
ESS	–	78.00%	75.75%	65.00%	66.25%	61.25%	71.75%
FIC	–	–	81.25%	79.50%	77.75%	74.50%	80.50%
INS	–	–	–	74.50%	75.50%	79.00%	74.25%
POP	–	–	–	–	68.25%	67.50%	69.00%
SHA	–	–	–	–	–	66.00%	69.25%
SPE	–	–	–	–	–	–	72.75%

Table 4: Genres Classification in Translation in Binary Settings: POS Trigrams.

Classes	ESS	FIC	INS	POP	SHA	SPE	TOU
ESS	–	76.25%	74.00%	66.25%	71.00%	61.25%	72.50%
FIC	–	–	76.50%	76.50%	77.75%	76.25%	84.00%
INS	–	–	–	74.25%	76.75%	76.75%	74.75%
POP	–	–	–	–	71.00%	67.75%	71.75%
SHA	–	–	–	–	–	65.00%	68.00%
SPE	–	–	–	–	–	–	71.25%

Table 5: Genres Classification in Translation in Binary Settings: POS 4-grams.

Classes	ESS	FIC	INS	POP	SHA	SPE	TOU
ESS	–	73.75%	75.25%	69.50%	67.00%	65.25%	72.50%
FIC	–	–	70.20%	75.00%	78.25%	76.00%	79.25%
INS	–	–	–	68.50%	70.50%	74.50%	74.50%
POP	–	–	–	–	71.25%	68.75%	70.75%
SHA	–	–	–	–	–	67.00%	66.25%
SPE	–	–	–	–	–	–	73.00%

is able to discriminate between tourism leaflets and fictional texts with impressive results of 84% accuracy.

We observe that in the vast majority of settings presented in Table 4, results obtained using POS trigrams were higher than those using POS bigrams.

Finally, in our last setting using POS 4-grams, we observed that this set of features do not achieve the best results in distinguishing between genres. For this reason, we preset feature analysis on POS trigrams which were the features that obtained the best results in this section.

5 Feature analysis

Text classification allows us to not only measure how well certain subcorpora (e.g. human and machine translations) are distinguished from each other, but also which individual features contribute to this distinction. Therefore, we analyse the output features resulting from the classification in this section. The main aim here is to identify the most informative features from the delexicalized n-grams in our experiments and to interpret them in terms of linguistic categories. This step is manual and carried out by looking through the most informative features and thus discriminative for certain genres and translation methods in our translation data.

Delexicalized trigrams consist of a sequence of words and placeholders, e.g. (1) *ART NN VMFIN* (2) *KON PPER*, etc. Intuitively, we try to recognise more categories on a more abstract level of linguistic description, i.e. category of modality expressed through modal verbs, discourse-building devices, such as discourse markers and coreference and others for the given trigrams. Thus, example (1) represents a finite clause containing a full nominal phrase, and example (2) represents a pattern related to the level of discourse: a Connector at sentence start followed by a personal pronoun that likely refers to something previously mentioned in the text. We decide for the evaluation of trigrams, as the performance of trigram models achieved the best results in both classification tasks.

5.1 Translation methods

The classification results for the distinction of translation methods outputs two lists of features: (1) the list of the features specific for human translations and (2) the list of features specific for machine translation. We analyse up to the first 20 features per translation method, summarising our observations in Table 6.

As seen from the lists, both translation methods have similar types of features that differentiate them from each other: they can be classified in terms of more abstract linguistic categories, such as discourse and modality. And some of them concern the preferred typology of phrases which can be related to the style of writing: nominal vs. verbal. The differences between the features discriminating between human and machine translations are visible on a more fine-grained level. For example, if we take into account morpho-syntactic preferences of discourse phenomena, we observe differences in the position of cohesive triggers: in machine translations, several patterns contain a punctuation mark in the first position, which means that the observed pattern often

Table 6: Features discriminating between human and machine translations.

	human	machine
Discourse	conjunct at sent.start followed by a personal referring expression	conjunct at sent.start followed by a full NP adverbial at clause start conjunct at clause start
	coreference at sentence start (demonstrative) coreference and negation	full and pronominal referring expressions at clause start
Modality	–	+
Phrases	NP connected to other phrases conj linking NP NP with named entities	V2 phrases followed by a definite NP conj linking VP NP describing location predicative adjectives locative prepositional phrases
Verbs	verb subcategorisation patterns	apposition V2 structure imperative constructions

represents sentence and clause start. Human translations rather show preferences for more sentence-starting devices. Example (2) reveals the possible reasons for this observation: human translators tend to split longer sentences into several ones (2-b), whereas a machine translation system keeps the structure (2-c) as it was in the source (2-a).

(2) a. *He used it to modernise the castle but he must have skimped on the kitchen, since 1639 it fell into the sea and carried away the cooks and all their pots.*
b. *Er erbeutete das vom Schiff mitgeführte Gold und benutzte es zur Modernisierung des Schlosses. Allerdings scheint er beim Ausbau der Küche etwas geknausert zu haben, denn dieser Teil des Schlosses rutschte im Jahre 1639 ins Meer ab. Die Köche wurden mitsamt Töpfen weggespült.*
c. *Er benutzte es, um das Schloss zu modernisieren, aber er muss auf dem Küchentisch gespart haben, seit 1639 ins Meer fiel und trugen die Köche und alle ihre Töpfe.*

Another difference that is clearly seen from the examples in the corpus data is the preference of human translations for conjuncts, whereas for machine translations, subjuncts and adverbials seem to be more typical. This is seen in the illustration in example (3).

(3) a. Human: *Und er wandte sich von der goldenen Dame ab und hätte gar zu gern die silberne genommen…*
b. Machine: *Darüber hinaus muss der Bohrvorgang verfeinert, so dass die tieferen Aquiferen nicht durch Arsen-Lager Wasser rann von den flachen Grundwasserleiter durch die Bohrungen selbst vergiftet werden.*

The multiword conjunction *so dass* is very frequent in machine translations in our data: the total of 108 occurrences as compared to 6 occurrences in human translations. A closer look at the data reveals that all the occurrences of *so dass* are found in machine translations produced with a statistical system trained on a large amount of data.

Human translations have also a number of features related to modality expressed via modal verbs. Our additional quantitative analysis of the modal verb distributions shows that, in general, human translations contain slightly more modal verbs than the machine ones: 16.49% vs. 15.72% out of all sentences in our data. This means that although modal verbs are more frequent in human translations, linguistic patterns with modals are more distinctive for machine-translated texts.

There is a difference in adjectival constructions. Predicative adjectives that turn to be discriminative for machine translations are also more frequent in this translation variety than in the human one (24.86% vs. 23.72%).

(4) a. *The roads are excellent, with miles of motorway and dual carriageway…*
b. *Es gibt ausgezeichnete Straßen, davon ungefähr 112 km Autobahn und noch weit mehr Kilometer mit zweispurigen Fahrbahnen.*
c. *Die Straßen sind sehr gut, mit Meilen von Autobahn und Schnellstraße…*

As seen from example (4), the machine-translated sentence (4-c) is closer to the source one (4-a) in terms of the predicative vs. attributive usage of the adjective (*are excellent – sind sehr gut*). At the same time, human translation (4-b) is closer to the source in terms of lexical choice (*excellent – ausgezeichnet*).

Another interesting difference is the prevalence of nominal structures in human translations as opposed to machine ones (46% vs. 24%) in the analysed trigram patterns. At the same time, machine-translated texts in our corpus contain more verbal phrases under their discriminative features (50% vs. 34%). We believe that this tendency is observed due to the shining though effect (cf. Teich, 2003): German-English contrastive analyses, e.g. the one by Steiner (2012), show that German has a preference for nominal structures, whereas English is more verbal. So, if a similar preference is observed in English-to-German translations, this could be interpreted as a phenomenon of shining though.

5.2 Genres

Using the same strategy, we generate a list of features discriminating genre pairs. For the sake of space, we will concentrate on the analysis of two genres only: fictional texts and political speech. For the first one, we achieve the best results in the trigram classification, whereas the classification results seem to be the worst for the second.

5.2.1 Fictional texts

We analyse six lists of patterns that turn out to be discriminative for fiction in the six classification tasks involving fictional texts: (1) fiction vs. political essays, (2) fiction vs. instruction manuals, (3) fiction vs. popular-scientific texts, (4) fiction vs. letters-to-shareholders, (5) fiction vs. political speeches and (6) fiction vs. tourism leaflets. First, we sort the patterns that occur more than once in the lists. The data contains several types of patterns: (a) informative in five classification tasks, i.e. member of five lists; (b) member of four lists; (c) member of three lists; (d) member of two lists; (e) member of one list – informative on one particular classification task, e.g. fictional texts vs. instruction manuals. The most frequent patterns are considered to be the most specific ones for fictional texts, as they were informative in several classification tasks, i.e. in discriminating fictional texts from several other genres. For instance, the pattern *, ADV KOUS*, e.g. *, so dass / , noch bevor / , auch wenn* (a discourse marker that links a subordinate clause), is informative in the first five classification tasks (fiction against essays, instructions, letters-to-shareholders, political speeches and tourism texts).

Table 7 illustrates the distribution of fiction-discriminative patters. The final list comprises 170 patterns.

Table 7: Feature lists discriminating between fiction and other genres

	list membership	types
(a)	5 lists	5
(b)	4 lists	5
(c)	3 lists	23
(d)	2 lists	49
(e)	1 list	88
total		**170**

Table 8: Features informative for fiction in most classification tasks for fiction.

pattern	example	excluding
$, ADV KOUS	*, so dass/ , noch bevor/ , auch wenn*	POP
ADV KOUS PPER	*so dass er/ auch wenn sie*	POP
ADV VVFIN $.	*gern wiederholen ./ nebeneinander stellen*	POP
$(PPER VMFIN	*(sie können/ (wir wollen/ (es dürfte*	INS
PPER VMFIN PPER	*sie können ihn/ wir wollen uns/ ich möchte ihm*	SHA

In the following, we analyse those that occur in most tasks (5 lists) and a subset of those that occur in one classification task only (1 list). Table 8 illustrates the five language patterns that turned to be the most informative in most classification tasks for the discrimination of fictional texts. The last column of the table provides the information on the genre, for which the result is not valid. The first three patterns are not discriminative for fictional texts, when classified against popular-scientific articles, the fourth pattern is not informative when instructional manuals are involved. And the last one is not discriminative for fictional text, when they are classified vs. letters-to-shareholders.

Most language patterns in Table 8 are discourse-related devices, i.e. discourse markers expressing conjunctive relations (*so dass*, *auch wenn*) or pronouns triggering cohesive reference (*er, sie, es*). The last two patterns contain also modal verbs which can be interpreted in terms of sentence or text modality.

The discriminative power of features does not necessarily imply a high frequency of a particular pattern in fictional texts. Nevertheless, the distribution of the first pattern across genres in our data shows that this trigram is more frequent in fiction than in the other genres, see Table 9 (the number are normalised per 1000 per total number of trigrams). However, the numbers in the table do not reveal the reasons for this pattern not being discriminative in the classification task for fiction vs. popular-scientific texts.

In the last step, we analyse the list of language patterns discriminating fiction in one particular task that includes 88 trigrams. In Table 10, we present a summary of these patterns describing them in terms of more general linguistic categories, e.g. specific phrases or functions.

Table 9: Distribution of *$, ADV KOUS* across genres.

genre	TOU	SHA	SPE	POP	ESS	INS	FIC
freq	0.10	0.11	0.14	0.23	0.29	0.35	0.41

Table 10: Features informative for fiction in one classification task only.

feature	example pattern	language example
phrases with adjectives	ADJA KON ADJA	*heller und dunkler / ernste und vielschichtige*
	CARD ADJA NN	*zwei junge Männer / drei wunderschöne Damen*
	PPOSAT NN ADJD	*ihre Hörner steil / ihre Lieder schwer*
phrases with adverbs	ADV VAFIN PPER	*so ist es / dann hast du*
	ADV VVFIN ,	*anders war , / unterwegs ist ,*
	ADV VVFIN PPER	*jetzt kriegt er / dann grunzt er*
coreference via pronouns	PPER ADV VVINF	*sie weiter sprechen / dir nur sagen*
	PPER VAFIN PPER	*sie hatte sie / ich habe sie*
	$. PPER VMFIN	*. Sie macht / Sie beginnen*
discourse markers	KON PPER VMFIN	*Aber sie möchte / und er konnte*
	KOUS PPOSAT NN	*dass meine Mutter / ob ihr Brief*
	VAFIN $. KON	*würde. Aber / hatte . Und*

As seen from the table, the patterns specific for fiction include adjective and adverb modification and elements that contribute to structuring discourse in a text. The latter are especially specific for narrative texts, which our fictional texts belong to. These observations coincide with the results of other empirical analyses on genres, e.g. those obtained by Neumann (2013). The author also points to personal pronouns and predicative adjectives as indicators of narration and casual style which are specific for fictional texts.

5.2.2 Political speeches

We proceed with the analysis of political speeches, which turned to be the hardest genre to identify in the classification with trigrams. The same analysis steps as we used for fictional texts are applied here.

First, we summarise the patterns that occur more than once in the lists. The political speeches contain less types of patterns than the fictional texts. An informative pattern can be a member of maximum four classification tasks only (for fictional texts, we also had five). So, we have fours lists of patterns: (a) informative in four classification tasks, i.e. member of our lists; (b) member of three lists; (c) member of two lists; (d) member of one list – informative on one particular classification task, e.g. political speeches vs. fictional texts. As in the previous case with fictional texts, we consider the most frequent patterns to be the most specific

ones for political speeches, as they contribute to the distinction of speeches from several other genres. For instance, the pattern *PTKVZ $. ART*, e.g. *vor . Das / bei . Die / weiter . Das* (sentence end followed by a sentence starting with a nominal phrase with an article), is informative in the following four classification tasks: political speeches vs. fiction, instructions, letters-to-shareholders, and tourism texts. Table 11 illustrates the distribution of speech-discriminative patters.

The total number of patterns is smaller than that of fictional texts (156 vs. 170 pattern types). We believe that the more distinctive features a genre has, the more distinctive it is from other genres, and thus can be easily identified with automatic classification techniques.

Now we will have a closer look at the patterns that are informative in most tasks and some of those that are discriminative for political speeches in one classification task only (1 list).

Table 12 illustrates the three language patterns that turned to be the most informative for the discrimination of political speeches. The last column of the table provides the information on the two genres, for which the result is not valid. All the three patterns are not discriminative for political speeches, when classified against popular-scientific articles. The first and the last patterns are not informative, when political speeches are distinguished from political essays. And the second pattern is not discriminative in the classification against tourism texts.

Table 11: Features discriminating between political speeches and other genres.

list membership	types	
(a)	4 lists	3
(b)	3 lists	18
(c)	2 lists	49
(d)	1 list	86
total		**156**

Table 12: Features informative for political speeches in most classification tasks for political speeches.

pattern	example	excluding
PTKVZ $. ART	*vor . Das / bei . Die / weiter . Das*	**ESS, POP**
VVFIN PPER ADV	*arbeiten wir zurzeit / auch wenn sie*	**POP, TOU**
VVPP VAINF $.	*gern wiederholen ./ nebeneinander stellen*	**ESS, POP**

Table 13: Features informative for political speeches in one classification task only.

feature	example pattern	language example
phrases with adjectives	ADJA NN KOKOM	*politische Themen wie / weltweite Probleme wie*
	ADJD APPR ART	*wichtig für die / möglich für die*
	ADV ADJA NN	*sehr geehrte Mitglieder / ebenfalls sprunghafte Fortschritte*
infinitive phrases	$, PTKZU VVINF	*, zu sprechen / , zu bekämpfen / , zu besprechen*
	ADJD PTKZU VVINF	*schwer zu entscheiden / richtig zu verteidigen*
	PTKZU VVINF KON	*zu* übernehmen *und / zu unterstützen und*
coreference via pronouns	PPER PPOSAT NN	*wir unsere Ziele / wir unseren Feinden / ich Ihre Fragen*
	PPER VVINF $,	*wir prüfen , / Sie antworten , / wir erreichen ,*
	$, VMFIN PPER	*, müssen wir / , möchte ich / , können wir*
discourse markers	$. ADV VAFIN	*. Bisher haben / . Natürlich ist*
	$. ADV VVFIN	*. Möglicherweise brauchen / : Erstens versucht*
	VAFIN $. KON	*würde. Aber / hatte . Und*

In the last step, we analyse the list of language patterns discriminating political speeches in one particular task that includes 86 trigrams. In Table 13, we present a summary of these patterns describing them in terms of more general linguistic categories, e.g. specific phrases or functions.

As seen from the table, the patterns specific for political speeches include phrases with adjective, infinitive phrases and the elements that contribute to structuring discourse in a text. From the first sight, there are types of patterns that are similar to those analysed for the fictional texts. However, our qualitative analysis reveals substantial differences. The main differences is caused by the difference in the register orientation: in a narration (most of the fictional texts in the data), there is more orientation towards the content, whereas in political speeches, we observe a clear orientation of the author towards the audience. It is especially prominent in coreference-related features. Fictional texts utilise a great number of third person pronouns, whereas political speeches have much more first and second person pronouns, see example (5).

(5) *Was passierte mit den Kindern? Wollen Sie sagen, dass Sie eine Milliarde Dollar ausgegeben haben und nicht wissen...* [*What happened to the children? Do you mean that you spent a billion dollars and you don't know...*]

This again, coincides to what was previously observed in register/genre-related analysis (Biber et al., 1999; Neumann, 2013).

6 Conclusion and discussion

This study is, to our knowledge, the first attempt to use text classification techniques to discriminate methods and genres in translations using fully delexicalized text representations and to identify their specific features and relevant systemic differences in a single study. We report results of up to 62.50% f-measure in distinguishing between human and machine translations using POS trigrams and 81.25% accuracy in discriminating between speech and essays.

The results obtained using POS tags as features was surprisingly higher than those obtained using (semi-)delexicalized representations presented in Zampieri & Lapshinova-Koltunski (2015). This seems to indicate relevant systemic differences across genres and methods of translation that algorithms relying on (morpho)-syntactic features are able to recognize.

At the same time, the results show that it is much harder to differentiate between translation methods than between different genres, even if fully delexicalized features are used. This confirms the results by Lapshinova-Koltunski (2017) which shows that if we compare the influence of the genre dimensions in translation variation is much stronger than that of translation method.

The results of our analysis can find application in both human and machine translation. In the first case, they deliver valuable knowledge on the translation product, which is influenced by the methods used in the process and the context of text production expressed by the genre. In case of machine translation, the results will provide a method to automatically identify genres in translation data thus helping to separate out-of-genre data from a training corpus.

The resulting lists of features can also be beneficial for automatic genre classification or human vs. machine distinction tasks. The knowledge on the differences between genres that these features deliver can also help to understand main differences between texts translated by humans and with machine translation systems. This information is especially valuable for translator training. Nowadays, translator training includes courses on post-editing technologies, since the application of such technologies has increased in translation industry recently. Translators need to know where the main problems (not necessarily errors) of machine-translated texts lie and what differs them from the texts by professional translators. This knowledge increases productivity in translation process.

In our future work, we want to perform a classification task for translation method within each genres. We assume that the differences between texts that differ in translation methods can be identified better, if classification is carried out within on genre only. Moreover, this will provide us with the information on how human and machine translations differ, if one particular genre is involved.

References

Babych, B., Hartley, A., and Sharoff, S. (2004). Modelling legitimate translation variation for automatic evaluation of mt quality. In *Proceedings of LREC-2004*, Vol. 3.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. *Text and technology: In honour of John Sinclair*, 233:250.

Baroni, M. and Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Biber, D. (1995). *Dimensions of Register Variation. A Cross Linguistic Comparison*. Cambridge University Press, Cambridge.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman, Harlow.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., and Specia, L. (eds.) (2014). *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-Evaluation the Role of Bleu in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 249–256.

Comelles, E. and Atserias, J. (2014). VERTa Participation in the WMT14 Metrics Task. In *Proceedings of the 9th Workshop on Statistical Machine Translation (WMT)*, 368–375, Baltimore, Maryland, USA. Association for Computational Linguistics.

Corston-Oliver, S., Gamon, M., and Brockett, C. (2001). A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of the 3th Annual Meeting on Association for Computational Linguistics*, 148–155.

De Sutter, G., Delaere, I., and Plevoets, K. (2012). Lexical lectometry in corpus-based translation studies: Combining profile-based correspondence analysis and logistic regression modeling. In *Quantitative Methods in Corpus-based Translation Studies: a Practical Guide to Descriptive Translation Research*, volume 51, 325–345. John Benjamins Publishing Company, Amsterdam, The Netherlands.

Delaere, I. and De Sutter, G. (2013). Applying a multidimensional, register-sensitive approach to visualize normalization in translated and non-translated Dutch. *Belgian Journal of Linguistics*, 27:43–60.

Denkowski, M. and Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Diwersy, S., Evert, S., and Neumann, S. (2014). A semi-supervised multivariate approach to the study of language variation. *Linguistic Variation in Text and Speech, within and across Languages*.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd International Conference on Human Language Technologies (HLT)*, 138–145.

El-Haj, M., Rayson, P., and Hall, D. (2014). Language independent evaluation of translation style and consistency: Comparing human and machine translations of Camus' novel "The Stranger". In Sojka, P., Horák, A., Kopeček, I., and Pala, K. (eds.), *Proceedings of the 17th International Conference TSD 2014*, volume 8655 of *Lecture Notes in Computer Science*, Brno, Czech Republic. Springer.

Fishel, M., Sennrich, R., Popovic, M., and Bojar, O. (2012). Terrorcat: a translation error categorization-based mt quality metric. In *7th Workshop on Statistical Machine Translation*.

Gebre, B. G., Zampieri, M., Wittenburg, P., and Heskens, T. (2013). Improving native language identification with tf-idf weighting. In *Proceedings of the 8th NAACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA8)*, Atlanta, USA.

Gonzàlez, M., Barrón-Cedeño, A., and Màrquez, L. (2014). IPA and STOUT: Leveraging Linguistic and Source-based Features for Machine Translation Evaluation. In *Proceedings of the 9th Workshop on Statistical Machine Translation (WMT)*, 394–401, Baltimore, Maryland, USA. Association for Computational Linguistics.

Goutte, C., Léger, S., Malmasi, S., and Zampieri, M. (2016). Discriminating Similar Languages: Evaluations and Explorations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 1800–1807, Portoroz, Slovenia.

Gupta, R., Orăsan, C., and van Genabith, J. (2015). ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal.

Halliday, M. and Hasan, R. (1989). *Language, context and text: Aspects of language in a social-semiotic perspective*. Oxford University Press, Oxford.

Hansen-Schirra, S., Neumann, S., and Steiner, E. (2012). *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. De Gruyter, Berlin, New York.

House, J. (2014). *Translation Quality Assessment. Past and Present*. Routledge.

Irvine, A. and Callison-Burch, C. (2014). Using comparable corpora to adapt mt models to new domains. In *Proceedings of the ACL Workshop on Statistical Machine Translation (WMT)*.

Irvine, A., Morgan, J., Carpuat, M., III, H. D., and Munteanu, D. S. (2013). Measuring machine translation errors in new domains. *TACL*, 1:429–440.

Kibriya, A., Frank, E., Pfahringer, B., and Holmes, G. (2004). Multinomial naive bayes for text categorization revisited. In *Proceedings of the Australian Conference on Artificial Intelligence*, 488–499.

Kunz, K., Degaetano-Ortlieb, S., Lapshinova-Koltunski, E., Menzel, K., and Steiner, E. (*2017*). Gecco – an empirically-based comparison of english-german cohesion. In De Sutter, G., Delaere, I., and Lefer, M.-A., editors, *New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies*. Mouton de Gruyter. TILSM series.

Lapshinova-Koltunski, E. (2013). VARTRA: A comparable corpus for analysis of translation variation. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, 77–86, Sofia, Bulgaria. Association for Computational Linguistics.

Lapshinova-Koltunski, E. (2017). Exploratory analysis of dimensions influencing variation in translation: The case of text register and translation method. In De Sutter, G., Delaere, I., and Lefer, M.-A., (eds.), *New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies*. Mouton de Gruyter. TILSM series.

Lapshinova-Koltunski, E. and Vela, M. (2015). Measuring 'registerness' in human and machine translation: A text classification approach. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, 122–131, Lisbon, Portugal. Association for Computational Linguistics.

Lee, D. Y. (2001). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the bnc jungle. *Technology*, 5:37–72.

Malmasi, S., Dras, M., and Zampieri, M. (2016). LTG at SemEval-2016 Task 11: Complex Word Identification with Classifier Ensembles. In *Proceedings of SemEval*.

Medlock, B. (2008). Investigating classification for natural language processing tasks. Technical report, University of Cambridge – Computer Laboratory.

Neumann, S. (2013). *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. De Gruyter Mouton, Berlin, Boston.

Niculae, V., Zampieri, M., Dinu, L. P., and Ciobanu, A. M. (2014). Temporal text ranking and automatic dating of texts. In *14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*. Association for Computational Linguistics.

Papineni, K., Roukus, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318.

Popovic, M. and Ney, H. (2011). Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

Santini, M., Mehler, A., and Sharoff, S. (2010). Riding the rough waves of genre on the web. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*, 3–30. Springer.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Stanojević, M. and Simaán, K. (2014). BEER: BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 414–419, Baltimore, Maryland, USA. Association for Computational Linguistics.

Steiner, E. (2004). *Translated Texts. Properties, Variants, Evaluations*. Peter Lang Verlag, Frankfurt/M.

Steiner, E. (2012). A characterization of the resource based on shallow statistics. In Hansen-Schirra, S., Neumann, S., and Steiner, E., editors, *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. Mouton de Gruyter, Berlin, New York.

Teich, E. (2003). *Cross-Linguistic Variation in System und Text. A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.

Vela, M. and van Genabith, J. (2015). Re-assessing the WMT2013 Human Evaluation with Professional Translators Trainees. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*.

Volansky, V., Ordan, N., and Wintner, S. (2011). More human or more translated? original texts vs. human and machine translations. In *Proceedings of the 11th Bar-Ilan Symposium on the Foundations of AI With ISCOL*.

Wu, H., Wang, H., and Zong, C. (2008). Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In Scott, D. and Uszkoreit, H., editors, *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*, 993–1000, Manchester, UK.

Zampieri, M. and Gebre, B. G. (2012). Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS2012*, 233–237, Vienna, Austria.

Zampieri, M. and Gebre, B. G. (2014). Varclass: An open source language identification tool for language varieties. In *Proceedings of Language Resources and Evaluation (LREC)*, Reykjavik, Iceland.

Zampieri, M., Gebre, B. G., and Diwersy, S. (2013). N-gram language models and POS distribution for the identification of Spanish varieties. In *Proceedings of TALN2013*, 580–587, Sable d'Olonne, France.

Zampieri, M. and Lapshinova-Koltunski, E. (2015). Investigating genre and method variation in translation using text classification. In Sojka, P., Horák, A., Kopecek, I., and Pala, K., editors, *Text, Speech and Dialogue – 18th International Conference, TSD 2015, Plzen, Czech Republic, Proceedings*, volume 9302 of *Lecture Notes in Computer Science*, 41–50. Springer.

Francesca Frontini, Mohamed Amine Boukhaled
and Jean-Gabriel Ganascia

Approaching French theatrical characters by syntactical analysis: a study with motifs and correspondence analysis

Abstract: This work uses computational approaches to study literature. It addresses the question of characterisation in theatrical plays, concentrating on the work of Molière, and trying to identify distinctive traits in the "voices" of some of the famous characters created by the French playwright.

The used technique adopts a syntagmatic approach, targeting differences in the distribution of sequences of grammatical categories, derived from texts in a bottom up way. First of all morpho-syntactic motifs of length 3 to 5 are extracted from the lines of each character with the help of a sequential pattern mining algorithm; then the differences in the distribution of such motifs across the characters are investigated by means of Correspondence Analysis.

In the first experiment four well known Molière protagonists (Harpagon, Dom Juan, Scapin, Sganarelle) are compared; in the second one a set of minor characters (the so called "Reasoners") are investigated.

The combined conclusions of the two proposed experiments highlight the complex intersection between voice and register in determining the stylistic choices of the author. In order for his characters to be perceived as realistic, Molière not only provides them with a distinctive voice, but also models their dialogue according to the communicative function they fulfil in the play.

1 Introduction

The use of computational approaches in the study of literature has a long-standing tradition. If we consider the word *computational* in its etymological sense of counting, we can date back such approaches prior to the era of computers (Lutosławski 1890; Mosteller and Wallace 1964). Again in the field of corpus linguistics the application of corpus methods to the analysis of style and genre dates

Francesca Frontini, Univ. Paul Valéry Montpellier 3, CNRS, PRAXILING UMP 5267
Mohamed Amine Boukhaled, Jean-Gabriel Ganascia, LIP6 (Laboratoire d'Informatique de Paris 6), Sorbonne Université and CNRS/OBVIL

https://doi.org/10.1515/9783110595864-006

back to the origins and continues today (Leech and Short 1981, Semino and Short 2004; Biber 2011; Mahlberg 2013). It is nevertheless undeniable that in recent years quantitative methods have moved out of the margins and into the forefront of literary studies, thanks to the availability of large quantities of digitised texts and to the success that "big data" methods have had in the identification of historical trends in literature (Moretti 2005; Jockers 2013) that would have been hard to spot to a naked eye. While the advantages of computational methods are evident when treating huge corpora, they exist also for smaller ones; single books or even parts of them, as we shall see, may disclose interesting and new insights when analysed from a different and new perspective.

The present work locates itself in the long tradition of stylometry, namely the application of statistical methods to the study of literary style (Holmes 1998[1]). Stylometric methods have often been applied to tackle issues of authorship attribution[2], but more recently a different discipline has evolved out of this field, namely *computational stylistics*. In these studies, techniques that were originally developed to identify the most likely author of a text of unknown attribution are applied as an analytic tool for the investigation of significant stylistic traits characterising a literary work, an author, a genre, a period, etc. As highlighted by Craig (2004) authorship attribution and computational stylistics, while sharing similar methods are nevertheless two different disciplines. While authorship attribution questions can be framed as a classification problem (the problem of who is the most likely author of a text given a bunch of candidates) – and indeed authorship questions do not only concern works of literature but also arise in forensic contexts – computational stylistics poses a series of open ended questions concerning the identification of those traits that are most distinctive of a particular set of texts in contrast to some other set of texts. From a computational point of view, computational stylistic methods may be framed as algorithms that rank linguistic features in a given text based on measures of *interestingness*. Moreover authorship attribution aims to identify unconscious traits in the work of a given author, that give him/her away and that are for this reasons normally defined as "fingerprints". On the other hand literary style is something that an author masters in a more conscious way. For this reason is possible that different works by the same author may show different stylistic traits, although others may be found in all of his/her works.

Research in computational stylistics is typically associated with the study of the authorial signal, namely the identification of a given author's typical traits

1 See also Grzybek (2014) for a history of the concept of *stylometry*.

2 See Stamatatos (2009) for an overview of authorship attribution methods.

through a comparison of his or her work to that of others. Studies have often addressed style in novels, and privileged the analysis of discrete units, typically of words. In this chapter we describe a novel stylometric method that combines the bottom up extraction of morpho-syntactic sequences, or patterns, with a type of statistical analysis called Correspondence Analysis, and we apply this method to the study of characterisation, namely to automatically finding characterising traits in the discourse of different theatrical characters by the same playwright.

2 The study of characterisation

Successful writers of literary fiction and theatre plays are generally renowned for their ability to create memorable characters that take on a life of their own and become almost as real as living people for their readers/audiences. The study of characterisation, namely the investigation into how these effects are achieved, is not a new topic in computational stylistics or in corpus studies, dating back to Burrows' seminal work on Jane Austen's characters (1987). In more recent works, authorship attribution methods have often been applied to the different characters of a novel or a play to identify whether the author has managed to provide each of them with a "distinct" voice. For instance Vogel and Lynch (2008) compare the dialogue of individual Shakespearean protagonists against the whole text of a play or even against all plays from the same author. Style, voice, and other issues may also combine; Agramon et al. (2009) mine the differences in the language use by authors and their characters in the Black Drama; in an interesting case study, Karina van Dalen-Oskam (2014) investigates the voices of two women writers who collaboratively published epistolary novels, as well as that of their fictional letter writers. Corpus based methods combining quantitative and qualitative analysis have also been applied in characterisation studies; most notably Michaela Mahlberg (2012) finds typical lexical patterns for memorable Dickens' characters by extracting those lexical bundles that stand out (namely those that are over-represented) in comparison with those found in a more general corpus, and later uses close reading to identify their function.

Stylometric techniques are often used in the study of characterisation. Having chosen a set of texts, some measurable properties, or *features*, are identified, and the texts are processed to extract counts of such properties or features. Features can be very diverse, but in most of the aforementioned works lexical or even sub-lexical elements (character n-grams) are used as features in the analysis.

As mentioned above, the analysis of the voice of characters may borrow methods from authorship attribution, such as the well known Burrows' Delta

(2002; see also Evert et al. 2015 for a discussion on these methods), which measures the differences in the distribution of a set of individual, high-frequency words in texts to identify the most likely author for a text. Other experiments make use of clustering techniques, such as Principal Component Analysis, to analyse both authorial differences and character idiole cts (Burrows and Craig 2012). Clearly, when authorship attribution methods are applied to the study of literary style the researcher already has a priori knowledge about what the results should look like; for instance when comparing several novels by different authors, we expect novels by the same author to cluster together. When comparing different characters, we may predict which characters have the most distinctive voice. At the same time, the fact that the algorithm can, or cannot, correctly group texts by the same author or lines of the same character is seen as an indication of its distinctiveness in terms of style or, in this case, of voice.

In most cases such techniques are types of multivariate analysis; an important advantage of their use is the fact that they require very little a priori feature selection, thus being suitable for use in exploratory scenarios. Texts may thus be described using a large set of linguistic features. For instances all words above a given frequency threshold may be counted and used as features. Most such words will have an approximately similar distribution in all texts, and only a group will significantly differ. Crucially, multivariate analysis may be used not only to find similarities or dissimilarities in groups of texts, but also to find out which are the linguistic features that are most distinctive for a group of texts. In our case, the distinguishing features in the voice of each character. This provides the scholars with a very useful analytical tool. As noted by Klaussner et al. (2015:2), "To determine an author's characteristic features, we first seek elements that he or she uses *consistently*, which we therefore regard as representative, but we likewise seek elements which the author uses *distinctively* in comparison to an opposing author". The same may be said at a smaller scale about the voice of characters.

The investigation of stylistic traits other than individual lexical frequencies is not so common in current research, but some interesting work has been carried out in the attempt to apply multivariate analysis to syntactic sequences such as Part of Speech Ngrams – sequences of N words in a text, identified only by their grammatical category (Nerbonne and Wiersma 2006; Wiersma et al. 2011).

In what follows we apply a novel methodology to the study of characterisation in French plays from a syntactic point of view. The work we present is intended to support textual analysis in two ways, namely by:

1. Verifying the degree of characterisation of each character with respect to others, and
2. Automatically inducing a list of linguistic features that are significant and representative for that character.

The methodology we propose relies on sequential data mining for the extraction of morpho-syntactic patterns and on correspondence analysis for the comparison of pattern frequencies for each character and for the visual representation of such differences. We will outline our methodology and test it on the work of the French playwright Molière, cross-comparing characters from different plays.

Our proposed analysis will show that multivariate methods, when removed from the realm of authorship attribution and projected into that of literary criticism, stop being an quantitative method in a scientific experiment, and they become a hermeneutic instrument that, while not providing clear-cut, yes-no answers, may help the expert in gaining new and interesting insights into the texts and the author.

3 Identifying distinctive syntactic patterns

In our study, we consider a *syntagmatic approach*, focussing on the combinatorial properties of language, and based on a similar configuration to the one proposed by Quiniou et al. (2012). In fact we target differences in the distribution of sequences of grammatical categories, derived from texts in a bottom up way.

The analysis is carried out from a comparative perspective, since we do not use a reference corpus to determine under- or overuse of structures but we compare a group of texts with each other. Clearly several dimensions of variation may exist in a group of texts, authoriality, genre, style, voice; given one and the same text, our method may derive different characterising features depending on the other texts it is compared to. Ideally one should compare texts that differ on only one a priori dimension. In the present experiment, the texts share the same author (Molière) and the same genre (drama) but differ in terms of *voice*, since we consider the lines of different characters from different plays separately. The purpose is therefore to identify the distinguishing traits of each character, in order to identify which devices the author uses to create them.

Thanks to pre-annotated digital Text Encoding Initiative editions, it is easy to extract all the lines of each character and group them in separate files. The text is first segmented into sentences, and then syntactical categories are annotated. For example the sentence "J'aime ma maison où j'ai grandi" is first mapped onto a sequence of PoS tags (1).

(1)	*J'*	*aime*	*ma*	*maison*	*où*	*j'*	*ai*
	PRO:PER	VER:pres	DET:POS	NOM	PRO:REL	PRO:PER	VER:pres
	grandi.						
	VER:pper	SENT					

Subsequently sequential patterns of a pre-determined length are extracted, with possible gaps. Our ad hoc tool, EREMOS (Extraction et REcherche de MOtifs Syntaxiques)[3], allows for several configurations, and allows us to set the length of patterns and the possibility of gaps. For instance, assuming that we chose to extract PoS sequences of length 3 to 5 with one possible gap, we will obtain the following sequences (among many others):

[PRO:PER][VER:pres][DET:POS][NOM]
(pronoun, verb in the present form, determiner, noun)

[PRO:PER][VER:pres][*][NOM]
(pronoun, verb in the present form, *anything*, noun)

[DET:POS][NOM][PRO:REL][PRO:PER]
(determiner, noun, relative pronoun, personal pronoun)

[PRO:REL][*][VER:pres][VER:pper][SENT]
(relative pronoun, *any part of speech*, verb in the present form, present participial, end of sentence punctuation)

Such sequences may not be in a one to one correspondence with grammatical constructions as defined by linguists *per se*, but may be nevertheless be considered as the superficial manifestations of underlying syntactic structures. For instance if a text overuses the sequence [PRO:PER][VER:pres][*][NOM] with respect to the other ones it is compared to, this tells us that the character uses more active sentences with pronominal subject and a direct object. Clearly the analysis needs to go beyond this, in order to identify which functions such forms absolve in the text, as we shall see later in the analysis. For this reason EREMOS is also equipped with an instance retrieving method that allows researchers to see all instances in the text corresponding to any given pattern. This latter feature is very important as humans can verify the evidence in texts and map the automatically induced patterns to the actual structures that such patterns are capturing.

Such morpho-syntactic patterns are extracted for each text with their absolute counts, which are then transformed into relative frequencies. A minimal level of filtering is applied, removing patterns that are present in less than 5% of the sentences of a text; some patterns will have some instances in one text, but not in the others. It is well known that (depending on the window and gap size) sequential pattern mining produces a large quantity of patterns even on relatively small

3 http://eremos.lip6.fr/

samples of texts. A large majority of such patterns will be uninteresting, since their distribution will be more or less the same in all texts; the ones we are interested in are those that are relatively overused in one text or in a subset of texts with respect to the others.

In order to identify the most relevant patterns for each text/character we choose to adopt correspondence analysis (CA), which is a multivariate statistical technique (Benzécri 1977; Greenacre 2007) often adopted for data analysis (Lebart et al. 1998) and supported by many statistical analysis and visualization tools[4]. CA allows us to represent both the characters and the patterns present in a text in a bi-dimensional space (see Figure 1) and thus to make visually clear not only which characters are more similar to each other but also which patterns are over/under-represented – that is, more distinctive – for each character or group of characters.

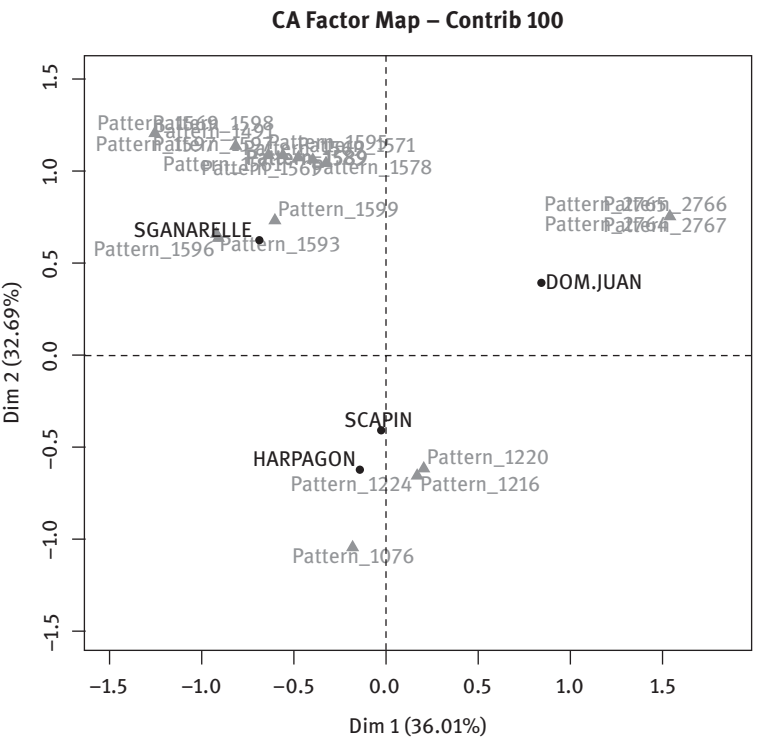


Figure 1: Correspondence analysis result plot, with first 100 patterns by contribution.

4 The results presented in this study are generated using an R module for Correspondence analysis and pattern ranking based FactoMiner (Husson et al. 2013).

Correspondence analysis produces several tables. Two of these contain respectively the coordinates to project both the texts and the patterns into the plot. These allow for a selective printing of a subset of patterns on the plot; moreover the proximity of a pattern to any of the texts can be easily calculated by Euclidean distance, thus allowing for the automatic filtering of patterns that are more strongly associated with one text than to the others. A third and the most important result table for our methodology contains the *contribution* of each pattern on the two axes; contribution is defined as the actual contribution of each pattern to the overall displacement of the position of texts in the resulting plot. If a pattern is strongly overrepresented in a text with respect to the others, it will contribute greatly to the displacement of that text in the bi-dimensional space, by "pulling" this text away from the others. Thus, the average contribution of such a pattern on the two axes of this pattern will be higher than the one of other patterns that have more or less the same frequencies in all texts.

The possibility of ranking patterns according to their contribution constitute a clear advantage of CA over similar techniques (such as Principal Component Analysis) when using a large amount of patterns. Patterns can be ranked according to their combined contribution on both axes; in other words it is possible to find out which morpho-syntactic patterns are most responsible for the horizontal and vertical displacement of characters in the bi-dimensional space; using this measure, the highest contribution patterns can be retained, thus enabling the researcher to filter out less interesting patterns. Subsequently contribution can be used as an *interestingness* measure.

To sum up our method runs as follows:

1. patterns are extracted for each text with EREMOS
2. pattern counts from all texts under analysis are imported into one big matrix
3. patterns that are not present in one text are assigned zero by default (smoothing is also possible)
4. the matrix may or may not be normalised, transforming absolute frequencies into relative ones
5. correspondence analysis is performed
6. morpho-syntactic patterns are ranked by contribution
7. a plot is printed, containing only the top most contributive patterns
8. thanks to Euclidean distance, tables are produced, with the top most contributive patterns for each character
9. pattern instances are retrieved for each text and analysed

Although similar clustering techniques have been used for classification and authorship attribution in the past, we propose the use of such measures as exploratory tools to provide researchers with corroborating evidence for any

pre-existing hypotheses that they may have, as well as to support them in the formulation of new hypotheses. Thus we consider this tool as a possible computational aid to the hermeneutical analysis of literature.

In what follows we shall better explain our method while reporting the results of first experiments some of Molière's most memorable characters. Clearly our tool is aimed at experts in literature, who can use this method for exploratory research. Some experiments with literary scholars have already been carried out, with promising results (Frontini & Benard 2015). The analysis in the present contribution does not intend to be a thorough critical analysis of Molière's characters, but mostly to show that the system is able to retrieve, in an unsupervised way, facts about these characters that are well established by literary scholars using traditional methods of analysis.

4 A case study on four of Molière's protagonists

It is only natural to begin an analysis on characterisation in Molière with four of his most memorable protagonists: **Harpagon** – *Avare* ("The Miser"); **Dom Juan** – *Dom Juan*; **Scapin** – *Les fourberies de Scapin* ("Scapin's Deceits"); **Sganarelle** – *Le médecin malgré lui* ("The doctor in spite of himself"). Numerous other characters obviously also come to mind, but previous experience has shown that the syntactic differences between prose and poetry are great enough to obfuscate characterisation differences between characters[5]. Thus we limit ourselves here to prose characters. The dialogues were extracted from an already pre-annotated digital edition[6] and automatically annotated with parts of speech using TreeTagger (Schmid 1994, 1995)[7].

In the present experiment, extracted patterns are 3-4-5grams of PoS tags, with at most one gap. We perform correspondence analysis using the pattern counts extracted for each character, and in particular we proceed as follows:

– First the 100 most contributive patterns are filtered and projected onto the plot together with the characters themselves; this will give us an idea of which characters are mostly marked and which patters they are associated to;

5 Analysis produced with prose and poetry together generate plots that show each group clearly clustering together.

6 We use the edition of the *Molière Project*, supervised by Georges Forestier at the Labex OBVIL (http://obvil.paris-sorbonne.fr/projets/projet-moliere)

7 The list of French parameters for TreeTagger can be found in Stein (2003).

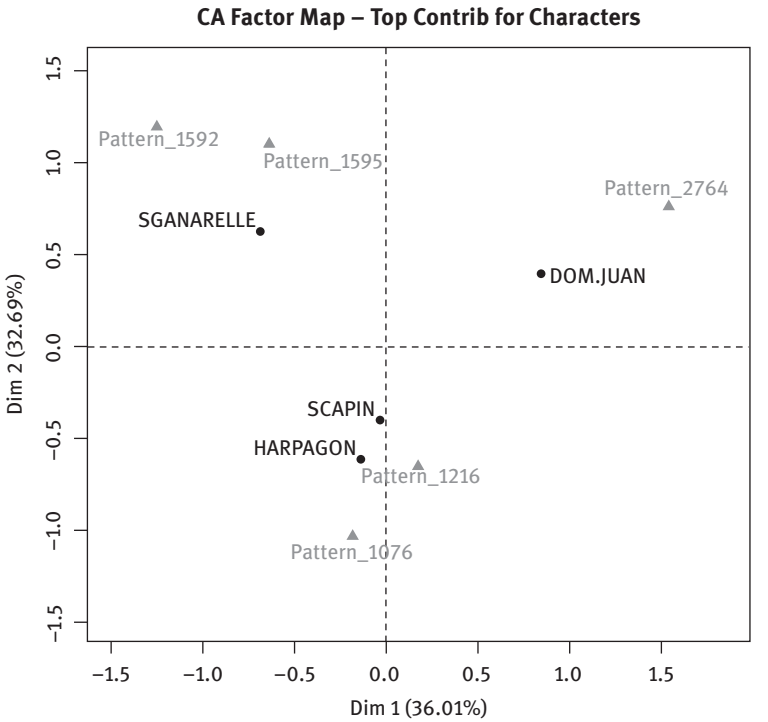


Figure 2: Some of the most contributive patterns for each character.

- then the 5 most contributive patterns for each character are extracted using Euclidean distance;
- finally the instances of selected patterns are extracted from the texts and analysed.

The plots (see Figure 1 and 2) show the relative distances between the four characters according to CA; they moreover shows the identifiers of – respectively – the first 100 most contributive patterns among those extracted altogether, and the most contributive ones for each character. As you can see, contributive patterns tend to position themselves at the outskirts of the plot, and are close to one (or two) characters. This graphically represents the fact that such patterns are *overused* by Molière to "give life" to one character with respect to the others.

As you can see, printing out 100 patterns (out of the 2768 ones extracted from the lines of the four characters) is useful to gain an overall picture and to determine which characters are most distinctive, but it makes the plot quite hard to read; moreover, the full pattern would be to long to print on the plot, thus

Table 1: Some of the top contributive patterns and the characters they are mostly associated to.

Pattern ID	Pattern	Character
Pattern_1592	[PRO:PER] [VER:pres] [KON] [*] [NOM]	Sganarelle
Pattern_1595	[PRO:PER] [PRO:PER] [*] [KON]	Sganarelle
Pattern_2764	[KON] [PRO:PER] [*] [VER:pres] [*] [PRP]	Dom Juan
Pattern_1216	[PRO:PER] [PRO:PER] [VER:pres] [VER:pper]	Harpagon / Scapin
Pattern_1076	[KON][PRO:PER][*][KON]	Harpagon / Scapin

identifiers are used. This is why the use of results in tabular form is also necessary. Table 1 provides some of the top contributive patterns (together with their identifiers), and the indication of the closest character.

From the combined analysis of plot and table, it becomes clear that the most isolated character seems to be *Sganarelle*, the protagonist of a piece in which a simple man is forced by circumstances to pretend to be a great doctor; most of the top contributive patterns are distributed around this character; that is to say, they are overused by Molière to give Sganarelle a peculiar voice. And indeed his language is quite different, from a syntactic point of view, from the other protagonists.

Among his most significant patterns we find syntactic structures that are typically used to express diagnosis (example 2).

(2) Instances of Pattern_1592 [PRO:PER] [VER:pres] [KON] [*] [NOM][8] for Sganarelle:
- … **il arrive que ces vapeurs** … Ossabandus, nequeys, nequer, potarinum, quipsa milus [it happens that such vapours…]
- **je tiens que cet empêchement** de l' action de sa langue est causé par de certaines humeurs … [I hold that this impediment in her tongue is caused by certain humours …]
- **il se trouve que le poumon**, que nous appelons en latin armyan, … [it happens that the lung, which we call in latin armyan …]
- **on voit que l' inégalité** de leurs opinions dépend du mouvement oblique du cercle de la lune … [we can see that the inequality of their opinions depends on the oblique movement of the moon orbit …]

8 '*' stands for any PoS tag.

In other cases (example 3) the pattern groups assertions having a performative and/or or a stancetaking[9] function and that are used initially to try and clear up misunderstandings (vainly it turns out), then to assure people of his assertions, and finally, once discovered, to confess.

(3) Instances of Pattern_1595 [PRO:PER] [PRO:PER] [*] [KON] for Sganarelle:
- je te dis que ... [I tell you that...]
- Je vous promets que ... [I promise you that...]
- Je vous jure que ... [I swear that ...]
- je vous dis que ... [I tell you that ...]
- Je vous assure que ... [I assure you that ...]
- Je vous assure que ... [I assure you that ...]
- je vous apprends que ... [I inform you that ...]
- Je vous apprendrai que ... [I shall assure you that ...]
- je vous avoue que ... [I confess you that ...]

Dom Juan, a nobleman and a complex character, is instead isolated by under-representation, in that he has less distinctive patterns, which may mean that his language is less repetitive and, possibly more elaborate. This is also evident from one of the few patterns that are strongly associated with him (example 4), which captures the over-use of subordinate clauses.

(4) Instances of Pattern_2764 [KON] [PRO:PER] [*] [VER:pres] [*] [PRP] for Dom Juan:
- sachez que je n' ai point d'autre dessein que de vous épouser ... [know that I have no other design than to marry you]
- elle va vous dire que je lui ai promis de l'épouser [... she is going to tell you that I promised her to marry her]
- Vous soutenez également toutes deux que je vous ai promis de vous prendre pour femmes [You both claim that I have promised you to marry you]
- ... et que je sais me servir de mon épée quand il le faut [and that I know how to use my sword when needed]
- ...

9 "I define stance as a person's expression of their relationship to their talk (their epistemic stance—e.g., how certain they are about their assertions), and a person's expression of their relationship to their interlocutors (their interpersonal stance—e.g., friendly or dominating", Kiesling (2009)).

One should also take into consideration that the play *Dom Juan* was written by Molière in "**prose rythmée**[10]" (rhythmic prose), which is not the case with the other plays in the current sample. This may also explain the isolation of this character given the higher degree of syntactic variability that metric constraints impose.

Finally the two comical characters *Scapin* and *Harpagon* are both characterised by patterns of lower syntactic complexity. This is especially the case with Harpagon (examples 5,6) whose patterns convey the image of a self-centered person, who wants to have things his way, and who is subject to violent disappointments (especially when money is concerned).

(5) Instances of Pattern_1216 [PRO:PER] [PRO:PER] [VER:pres] [VER:pper] for Harpagon:
- on m' a privé ... [they have deprived me ...]
- on m' a dérobé ... [they have robbed me ...]
- on m' a volée ... [they have stolen my ...]
- on m' a pris ... [they have taken my ...]
- ...

(6) Instances of Pattern_1076 [KON][PRO:PER][*][KON] for Harpagon:
- que je veux que ... [that I want that ...]
- et il faut que ... [and it is necessary that ...]
- et vous verrez qu' ... [and you will see that ...]
- ...

Manual inspection shows how these morpho-syntactic patterns have a slightly different function in Scapin, the clever servant who interacts with several characters in order to try to carry out his plan. In example 7 we see the same pattern as in example 4, but used mostly to report events.

(7) Instances of Pattern_1216 [PRO:PER] [PRO:PER] [VER:pres] [VER:pper] for Scapin:
- Je l' ai trouvé tantôt tout triste ... [I have found him so sad...]
- nous nous[11] sommes allés promener sur le port ... [we have gone to walk in the harbour...]
- ...

10 Following Georges Forestier (Forestier et al. 2010, v. I, p. 1623), we define "prose rythmée" as a particular type of prose alternating irregular, non-rhyming verses.

11 Notice how the reflexive pronoun is labelled as a personal pronoun. See the French TreeTagger tagset (Stein 2003) for details.

It is worth noticing how such structures in the past tense are under-represented in the character of Sganarelle, whose discourse is prevalently in the present tense; while Dom Juan, Sganarelle and Scapin are all actively lying in their respective plots, the use of past tense in the last of the three may be more reflective of conscious scheming.

This first analysis shows us that some well-known traits of the analysed protagonists can be automatically retrieved among the great mass of syntactic traits automatically extracted by sequential pattern mining. At the same time, one important issue seems to emerge, concerning the relationship between communicative function and characterisation when analysing syntactic features. Indeed, it is clear that the two aspects cannot be fully disentangled. As has been demonstrated by discourse analysis studies (Biber & Conrad 2009) the use of given syntactic structures highly differ according to register and to communicative situation; this is not only true for discrete features such as the proportions of verbs to nouns or the use of pronouns, but also for complex structures such as the ones targeted by our syntagmatic approach. This explains the fact that the extracted patterns seem to be highly representative of the kind of situations in which the character finds him/herself in the plot, as well as of his/her station in life. So for instance servants and middle class characters (Scapin, Sganarelle) use less complex structures than aristocrats (Dom Juan); characters who are forced to justify themselves, and to lie (Sganarelle) require specific syntactic structures with respect to the other ones. Some distinguishing psychological traits (such as the desire to control others in Harpagon) emerge, but are not as predominant as one might expect. This also tells us something particular about classical French plays where characterisation was often left to the actor. Molière wrote most of his protagonists to be played by himself. So a lot of the characterisation needs to be inferred by modern day performers rather than being explicitly given in the stage directions, and some room for freedom is left.

Nevertheless the emergence of such more "context motivated" patterns shows how such an analysis may provides useful insights into the way in which Molière constructed his characters and managed to give them each a voice that was plausible both from the social and the contextual point of view. Our second experiment will show how syntactic patterns can provide an interesting perspective in investigating the role that certain characters have in the plays.

5 A case study on *raisonneurs*

In our second experiment we focus on the figure of the *raisonneurs*, characters who take part in discussions with comical protagonists providing a counterpart to

Table 2: Plays and characters.

Play	Raisonneur	Counterpart
Ecole des femmes	Chrysalde	Arnolphe
Ecole des maris	Ariste	Sganarelle
Tartuffe	Cléante	Orgon
Mysantrope	Phylinte	Alceste
Malade imaginaire	Béralde	Argan

their follies. Such characters were interpreted at times as spokesmen for Molière himself, and the voice of reason, at other times as comical characters themselves and no less foolish than their opponents. Table 2 lists the plays we are going to analyse as well as the characters. Hawcroft's essay *Reasoning with fools* (2007) highlights the differences between five of these characters based on their role in the plot. Using this analysis as guidance, we compare significant linguistic patterns in order to see how these differences are marked by the stylistic choices of the author. Given the results of the previous experiment, we focus on the analysis of the discourse traits and on how they match to the communicative function each character needs to fulfill (Biber and Conrad 2009).

Figure 3 shows the result of the correspondence analysis, with the five *raisonneurs* and the 10 patterns with the highest contribution labeled with their identifiers.

The relative distances between the characters seem to match what is already known from literary criticism; first of all Béralde, who is the only character to express himself in prose, is isolated on the right of the X axis. As already remarked, it is not advisable to compare characters in prose and verse, but we have retained the example of Béralde to show how the proposed technique can easily identify differences in genre. As for the other characters, Hawcroft stresses the difference in the roles of Ariste, Philinte and Chysalde on the one hand and of Cléante on the other. The latter is a more pro-active character, more crucial to the plot; he is also less accommodating than the other three, who are depicted mostly as loyal friends and brothers, trying to help the hero to avoid the consequences of his foolish actions and beliefs. Instead, Cléante has also to worry about his sister's wellbeing: having to face not only the besotted brother in law, Orgon, but also the man who has duped him, Tartuffe.

In order to confirm the first impression of eccentricity of the character of Cléante with respect to Phylinte and Chrisalde, it is necessary to turn our attention to what it is that exactly causes the spatial distribution, namely the high contribution patterns, we find above. Our technique allows us not only to find the

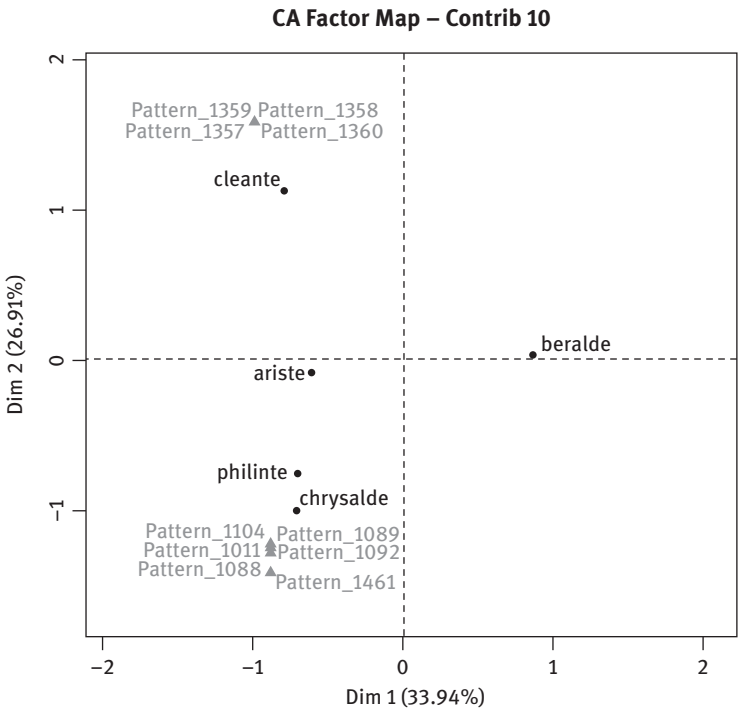


Figure 3: Top contributive patterns for the raisonneurs experiment.

corresponding pattern for each identifier on the plot, but also to extract all underlying instances in the texts. Due to space constraints, only a brief demonstrative analysis will be performed.

Phylinte and Chrysalde are strongly associated with patterns containing prepositional phrases separated by commas. Such patterns are used in contexts where the characters give advice in a very cautious, indirect way. The overuse of punctuation itself, in these two characters, seems to be an indication that the character should be played as a soft-spoken person, who is fond of his friend and careful not to offend, e.g.:

(8) Pattern 1011 [,] [*] [PRP] [any word] [NOM]

Instances from **Chrysalde**:

– Entre ces deux partis il en est un honnête**, Où dans l' occasion** l' homme prudent s' arrête … [In between these two extremes there is a right way, that at times the prudent man will choose …]

– Il faut jouer d' adresse**, et d' une âme** réduite, Corriger le hasard par la bonne conduite ... [One has to play a dextrous game with a prudent spirit, and compensate for hazard with good conduct...]

Instances from **Phylinte**:

– **, Et pour l' amour** de vous, je voudrais, de bon cœur, Avoir trouvé tantôt votre sonnet meilleur. [, and for the love of you, I would have gladly wished to have liked your sonnet better.]

On the other hand, the patterns most associated with Cléante contain modal constructions, and are indicative of a more direct way of advising, and of stronger arguments, e.g.

(9) Pattern 1360

[PRO:PER] [any word] [VER:infi] [PRP]

– Les bons et vrais dévots, qu' **on doit suivre à** la trace, Ne sont pas ceux aussi qui font tant de grimace. [True and real believers, whom one has to follow, are also those that do not make too much of a show.]
– Et s' il **vous faut tomber dans** une extrémité, Péchez plutôt encore de cet autre côté. [And if you have to fall into one extremity, choose rather the other one.]

Finally, the patterns extracted for Béralde are indicative of the greater simplicity and repetitiveness of his prose, and of the stereotypical role he has in the play, which is that of a man concerned with his brother, as in:

(10) Pattern 865 [,][DET:POS][any word][PUN]

– Oui**, mon frère,** puisqu' il faut parler à cœur ouvert, ... [Yes, my brother, since one has to speak with an open heart...]

Just as for the experiment with the protagonists, this brief analysis is clearly meant not to provide new insights on the issue of *raisonneurs*[12] but rather to show that the system behaves in a consistent way with respect to basic and known assumptions on the plays. At the same time, it is possible to see how such an instrument can be used to investigate "old" issues from a new perspective, providing the researcher with new useful insights on Molière's use of language, and

12 Hawcroft's analysis of the *raisonneurs* provided us with an interesting testing ground, but it should be compared to and read in the light of the influential interpretations of Georges Forestier on this topic.

how syntactic structures and their underlying communicative functions can contribute to shape the linguistic profile of theatrical characters.

6 Voice and register

The combined conclusions of the two proposed experiments seem to confirm the first intuitions about the complex intersection between voice and communicative function in determining the stylistic choices of the author in modelling the discourse of his characters. On the one hand, Molière tries to give a unique voice to his characters by use of distinctive traits, in particular, for some characters such as Harpagon and Sganarelle repetitions or the use of peculiar constructions that are especially enhancing the comical effect; on the other hand, he is also trying to create believable human specimens, that act and speak in accordance to the situation and the role they fulfil, as is evident from Dom Juan's high register and complex syntax, but also from the differences found in the discourse of various *raisonneurs*. In this sense, the author is also guided, if not constrained, by the syntactic and pragmatic rules of language, that require his using some structures over other ones to convey certain meanings. In a certain sense, authors are less free in their syntactic choices then in their lexical ones, and that is why syntax is less responsive to the authorial signal and more to genre or register as the studies of Douglas Bibler show.

At the same time the analysis of syntactic choices, as emerging from the combination of contiguous syntactic categories, can provide us with a different and interesting insight in texts even of a relatively short length[13] such as the ones analysed here. In particular morpho-syntactic pattern could also be used to compare the features of theatrical dialogue to those of genuine spoken dialogues, in order to investigate how far theatrical prose is able to mimic speech and real oral interaction[14]. Another interesting area of research that may benefit of the proposed approach beyond the study of characters is the investigation of Molière's dialogues on a more typological level, comparing for instance different types of scenes (long monologues, the comic exchanges, etc.) as done by Gabriel Conesa (1983) in his well known study. Here too,

13 The overall number of tokens for the lines of each character is in average 10,000 for the protagonists, and 2,500 for non protagonists such as the *raisonneurs*.

14 To remain on a simplistic and paradigmatic level, it is interesting to notice how the percentage of verbs is higher than that of nouns in almost all characters, a clear feature of oral language; the rate of pronouns is also higher than that of most written genres.

already known distinctive features could be provided with additional support corpus evidence, by the bottom up extraction and filtering of distinctive syntactic sequences.

7 Conclusion

The work presented in this contribution takes a syntagmatic perspective to the study of characterisation by analysing the voice of Molière's characters in terms of distinguishing syntactic patterns. In particular the extraction of all possible syntactic sequences of a given length (or window) is proposed as a particularly useful one to extract interesting features in an exploratory scenario. Clearly the proliferation of patterns and the difficulty for humans to make sense of the huge amount of resulting dimensions of variation between texts is a major obstacle to this approach. Multivariate analysis and clustering techniques are commonly used in such scenarios to treat and possibly reduce such large quantities of dimensions. They are well known in the field of computer aided literary criticism; for instance Principal Component analysis is implemented in some of the widely used tools for stylometric analysis (Eder, Kestermont and Rybicki 2013), though more often applied to unitary lexical elements.

The strength of correspondence analysis lies in the fact that it allows users to easily identify the reasons for certain texts to group together or to diverge. This helps to overcome the lack of transparency in the presentation of results, something that often disappoints experts when faced with experiments using similar techniques, and is therefore well suited to syntagmatic approaches, that are per definition combinatorial and thus high-dimensional. Thus the proposed methodology offers a useful instrument to facilitate literary analysis and criticism; not only does it calculate and represent the distances between characters (which may be possible using other clustering techniques) but it also provides a way to motivate and explain this difference based on the extraction of significant and distinctive sets of patterns for each character, which is a strong requirement for all computational stylistics methods. Although the present study concentrates on theatre and characterisation, the presented methodology can be extended to any language and literary genre, provided a reliable automatic PoS-Tagging for that language/genre is available. We have successfully applied it to the study of stylistics in general, such as novels (Frontini et al. to appear).

An on-going debate is currently raging as to whether computational stylistic methods provide a way to radically change the methodology of literary criticism and to make it more "scientific". Ramsay in his influential book "Reading

Machines. Toward an algorithmic criticism" (2011) argues that although these methods carry the potential to have this kind of effect, it is not inevitable. Similar claims are made by Matthew Jockers, in *Macroanalysis* (2013). The authors of the present contribution developed the tool described above through close contact with academic experts on Molière, who were able to evaluate our results and return a positive response. The main goal of our work is to provide experts with a tool that can be used to find a positive confirmation of well-known facts using techniques grounded in statistical data analysis, but that can eventually lead to "surprise" results that guide the expert towards further investigations.

References

Argamon, Shlomo, Charles Cooney, Russell Horton, Mark Olsen, Sterling Stein & Robert Voyer. 2009. Gender, Race, and Nationality in Black Drama, 1950-2006: Mining Differences in Language Use in Authors and their Characters. *Digital Humanities Quarterly* 3(2).

Benzécri, Jean-Paul. 1977. Histoire et préhistoire de l'analyse des données. Partie V : l'analyse des correspondances. *Cahiers de l'analyse des données* 2(1). 9–40.

Biber, Douglas. 2011. Corpus linguistics and the study of literature: Back to the future? *Scientific Study of Literature* 1(1). 15–23. doi:10.1075/ssol.1.1.02bib

Biber, Douglas & Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press.

Burrows, John F. 2002. "Delta": a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing* 17(3). 267–287. doi:10.1093/llc/17.3.267

Burrows, John F. 1987. *Computation into criticism: A study of Jane Austen's novels and an experiment in method*. Clarendon Pr.

Combe, Dominique. 2002. La stylistique des genres. *Langue française* 135(1). 33–49. doi:10.3406/lfr.2002.6461.

Conesa, Gabriel. 1983. *Le dialogue moliéresque*. Presses Universitaires de France.

Craig, Hugh. 2004. Stylistic analysis and authorship studies. In Susan Schreibman, Siemens Ray & John Unsworth (eds.), *A companion to digital humanities*, 273–288. Oxford: Blackwell.

Dalen-Oskam, Karina van. 2014. Epistolary voices. The case of Elisabeth Wolff and Agatha Deken. *Literary and Linguistic Computing* 29(3). 443–451. doi:10.1093/llc/fqu023

Dell'Orletta, Felice, Simonetta Montemagni & Giulia Venturi. 2011. READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, 73–83. (SLPAT '11). Stroudsburg, PA, USA: Association for Computational Linguistics.

Eder, Maciej, Mike Kestemont & Jan Rybicki. 2013. Stylometry with R: a suite of tools. *Digital Humanities 2013: Conference Abstracts*, 487–89. Lincoln: University of Nebraska-Lincoln.

Egbert, Jesse. 2012. Style in nineteenth century fiction: A Multi-Dimensional analysis. *Scientific Study of Literature* 2(2). 167–198. doi:10.1075/ssol.2.2.01egb

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Steffen Pielström, Christof Schöch & Thorsten Vitt. 2015. Towards a better understanding of Burrows's Delta in literary authorship attribution. *Proceedings of NAACL-HLT 2015*. Denver, Colorado, USA.

Forestier, Georges, Claude Bourqui, C.E.J. Caldicott, Alain Riffaud, Anne Piéjus & David Chataignier (eds.). 2010. *Molière. Oeuvres complètes*. Vol. 1 & 2. (Bibliothèque de La Pléiade). Paris: Gallimard.

Frontini Francesca & Elodie Benard. 2015. The Syntax of Stage. Studying Linguistic Patterns in Molière. Paper presented at the Göttinger philologisches Forum, Göttingen. https://www.uni-goettingen.de/de/empfehlung-»the-syntax-of-stage«-vortrag-von-francesca-frontini–elodie-bénard-am-3-dezember-2015-crc-texstrukturen/525494.html (7 February, 2016).

Frontini, Francesca, Mohamed-Amine Boukhaled & Jean-Gabriel Ganascia. To appear. Mining for characterising patterns in literature using correspondence analysis: an experiment on French novels. *Digital Humanities Quarterly* Proceedings of the Göttingen Dialogue for Digital Humanities.

Greenacre, Michael. 2007. *Correspondence analysis in practice*. CRC press.

Grzybek, Peter. 2014. The Emergence of Stylometry: Prolegomena to the History of Term and Concept. *Text within Text – Culture within Culture*, 58–75. Budapest, Tartu: L'Harmattan.

Hawcroft, Michael. 2007. *Molière: Reasoning with Fools*. Oxford England, New York: Oxford University Press.

Holmes, David I. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing* 13(3). 111–117. doi:10.1093/llc/13.3.111.

Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History*. (Topics in the Digital Humanities). University of Illinois Press.

Kiesling, Scott F. 2009. Style as Stance. In Alexandra Jaffe (ed.), *Stance: Sociolinguistic Perspectives*. Oxford University Press, USA.

Klaussner, Carmen, John Nerbonne & Çağrı Çöltekin. 2015. Finding characteristic features in stylometric analysis. *Digital Scholarship in the Humanities* 30(suppl 1). i114–i129.

Lebart, Ludovic, André Salem & Lisette Berry. 1998. *Exploring Textual Data*. Vol. 4. (Text, Speech and Language Technology). Dordrecht: Springer Netherlands.

Leech, Geoffrey N. & Mick Short. 2007. *Style in fiction: A linguistic introduction to English fictional prose*. (13). Pearson Education.

Lê, Sébastien, Julie Josse & François Husson. 2008. FactoMineR: an R package for multivariate analysis. *Journal of statistical software* 25(1). 1–18.

Lutosławski, Wincenty. 1898. *Principes de stylométrie appliqués à la chonologie des oeuvres de Platon*. Paris: Ernets Leroux.

Mahlberg, Michaela. 2012. *Corpus Stylistics and Dickens's Fiction*. 1st ed. (Routledge Advances in Corpus Linguistics). New York: Routledge.

Moretti, Franco. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London; New York: Verso.

Mosteller, Frederick & David L. Wallace. 1963. Inference in an Authorship Problem. *Journal of the American Statistical Association* 58(302). 275–309. doi:10.2307/2283270 (8 December, 2011).

Nerbonne, John & Wybo Wiersma. 2006. A measure of aggregate syntactic distance. *Proceedings of the Workshop on linguistic Distances*, 82–90. Association for Computational Linguistics.

Quiniou, Solen, Peggy Cellier, Thierry Charnois & Dominique Legallois. 2012. What about Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics? In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, 166–177. (Lecture Notes in Computer Science 7181). Springer Berlin Heidelberg.

Ramsay, Stephen. 2008. Algorithmic Criticism. *Companion to Digital Literary Studies*. (Blackwell Companions to Literature and Culture). Oxford: Blackwell Publishing Professional.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. *Proceedings of the international conference on new methods in language processing*. Manchester, UK.

Schmid, Helmut. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Semino, Elena & Mick Short. 2004. *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. 1st ed. Routledge.

Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60(3). 538–556.

Stein, Achim. 2003. French TreeTagger part-of-speech tags. Web-page. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html (25 March, 2015).

Vogel, Carl & Gerard Lynch. 2008. Computational Stylometry: Who's in a Play? In Anna Esposito, Nikolaos G. Bourbakis, Nikolaos Avouris & Ioannis Hatzilygeroudis (eds.), *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction: COST Action 2102 International Conference, Patras, Greece, October 29–31, 2007. Revised Papers*, 169–186. Berlin, Heidelberg: Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-540-70872-8_13.

Wiersma, Wybo, John Nerbonne & Timo Lauttamus. 2011. Automatically Extracting Typical Syntactic Differences from Corpora. *Literary and Linguistic Computing* 26(1). 107–124. doi:10.1093/llc/fqq017 (26 June, 2015).

Dominique Longrée and Sylvie Mellet

Towards a topological grammar of genres and styles: a way to combine paradigmatic quantitative analysis with a syntagmatic approach

Abstract: We analyze a corpus of classical Latin texts, comprising various literary genres and authors. Two Correspondence Analyses (CA) are based on discrete units (used by Biber 2006). The first one represents the distances between the main works in the classical Latin corpus according to the parts of speech used in the different texts, the second according to the distribution of verb tenses and moods. The paradigmatic approach is efficient for automatically classifying the texts, but provides little new information for the linguist or philologist.

We therefore assess the impact on genre characterization of taking the integration of the parts of speech (POS) and grammatical categories in syntactic structures (the syntagmatic approach) into account. However, even when the syntactic dimension is taken into consideration, this method does not really account for the sequential structure of the text's linearity. Moreover, the choice of the syntactic structures studied depends upon the knowledge already acquired by the Latinist and their detection is always supervised.

We therefore propose the new concept of *motif* in order to handle the different tokens of a given structure and to model them in a single pattern whose identification is based on its unified text dynamics function, disregarding surface variations. As a general pattern, the motif is able to characterize a genre; but its different realizations or tokens may be specific to different authors in a given genre. This claim is exemplified by a contrastive analysis of the style of two Latin historians who both lived at the close of the classical literary period, Caesar and Tacitus.

In order to contribute to the discussion herein about what makes a "Grammar of Genres and Styles", we would like to submit a methodological study based on textual analysis whose aim is to identify formal criteria for distinguishing between different discursive genres or authors' styles and characterizing them

Dominique Longrée, LASLA, Université de Liège et SeSLa, Université Saint-Louis Bruxelles, Belgium
Sylvie Mellet, Université Côte d'Azur, CNRS, BCL, France

https://doi.org/10.1515/9783110595864-007

according to their linguistic properties and textual dynamics[1]. In our previous work, we have used methods relying not only on a paradigmatic, quantitative analysis but also on syntagmatic approaches: sequences (Longrée and Luong 2003, 2005), text segmentations (Longrée, Luong, and Mellet 2004, 2006; Longrée and Mellet 2007), neighbourhoods (Mellet and Barthélemy, 2007; Luong, Julliard, Mellet and Longrée, 2007; Barthélemy, Longrée, Luong, and Mellet 2009) and bursts (Longrée, Luong, and Mellet 2008; Longrée and Mellet 2016). This work has led to a theoretical proposal to consider the text as a topological space and to introduce a new analytical unit that we call the "motif" (Longrée, Luong and Mellet 2008; Mellet and Longrée 2009, 2012; Longrée and Mellet 2013, 2014). With this methodological background in mind, we would like to assess here the benefits and limitations of both approaches – paradigmatic and syntagmatic – in the characterization of textual genres and author's styles.

1 The corpus and the methodology

As our previous work did, this study involves the analysis of a corpus of Latin classical texts, made up of literary works of various genres and authors. It follows a long philological tradition. For decades, classical philologists have tried to characterize styles and genres according to their lexical, lexico-grammatical, morphological, or even syntactic particularities, and they have therefore often used methods involving exhaustive counting, e.g. counting Sallustius' narrative infinitives, Caesar's historical presents (Mellet 1980), clausulas in the Latin prose (Aumont 1996), or Tacitus' postponed subordinate clauses (Seitz 1958; Kohl 1959). A book of J.P. Chausserie-Laprée, *L'expresssion narrative chez les historiens latins, Histoire d'un style*, published in 1969, is particularly representative of this kind of work: he studied a large sample of literary texts, from Caesar to Tacitus, in order to describe the evolution of historical narrative prose; he highlighted the interest of counting recurrent linguistic and stylistic phenomena; in particular, he counted occurrences of certain syntactic phrases according to their sentence positions in order to characterize different types of sentence structures.

Analysing Latin texts offers great advantages. First of all, it is a well-known, closed corpus. Secondly, since the 1960s, this corpus has been digitalized and

1 We are grateful to Peter Follette for his careful reading of our text.

tagged,[2] which allows for automatic counting and statistical processing.[3] This makes it possible to enhance philological studies with the modern methods of Textual Data Analysis, although it is still necessary to carry out a reflection on the theoretical concepts required for such an analysis.

We will first show that the paradigmatic approach as used by Biber (1988, 2006)[4] is useful for identifying pertinent generic classifications, but that these results are often rough and poorly informative (e.g. they build a tri-partition history / discourse / poetry). We will also show that the results of the classifications are better and more refined by taking into account a syntagmatic dimension, and that this second approach offers more accurate text characterization.[5]

Second, we will examine the available conceptual tools for the text syntagmatic approach, such as the grammatical n-grams; we will also introduce the notion of motif and we will address its relevance for our purposes.

This methodological exploration will allow us to select new, effective tools that we will use to characterize the respective styles of Caesar and Tacitus within the framework of a topological (Mellet and Barthélemy 2007) modelling of the texts. In this way, we will try to offer a new option to go beyond the paradigmatic / syntagmatic opposition of text analysis.

2 The potential and the limitations of the paradigmatic approach

The paradigmatic approach is herein illustrated by two Correspondence Analyses (CA). By "paradigmatic approach," we mean an analysis applied to grammatical features, building up closed classes. The first of the two CAs presented below represents the distances between the main works of the LASLA classical Latin

2 Digitalized, lemmatized and tagged corpus of the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) of the University of Liege (http://www.cipl.ulg.ac.be/Lasla/tlatins.html).

3 The software programs Hyperbase-Latin and Hyperbase Web Edition ("Bases, corpus, langage", Université Côte d'Azur) offer tools for such quantification and analysis (http://ancilla.unice.fr/ et http://hyperbase.unice.fr/).

4 However, we must note that, in a paper published in 2009, Biber took into account the existence of multi-word patterns, including a sketch of the syntagmatic dimension.

5 These two approaches overlap with the opposition between "language in the Mass" and "language in the Line". See Pawlowski 1999.

Corpus[6] according to the parts of speech used in the different texts; the second CA shows the same works analysed according to the distribution of verb tenses and moods.

This first CA (Figure 1) is applied to 21 parts of speech and 36 textual partitions. The POS are the following: substantive, verb, adjective, numeral, personal pronoun, possessive pronoun, reflexive pronoun, possessive reflexive pronoun, demonstrative pronoun, relative pronoun, interrogative pronoun, indefinite pronoun, adverb, relative adverb, interrogative adverb, negative adverb, interrogative-negative adverb, preposition, coordinating conjunction, subordinating conjunction, interjection.[7] The 36 textual partitions correspond to a very large sample of all Latin textual genres: theatre, treatise, poetry, history, speech and novel.[8] The first CA dimension (36% of information) opposes the speeches and the treatises (often written as a conversational debate) on the right side of the graph to all the other texts. Inside the theatre genre, it also opposes Plautus' comedies on the right ("Plaute") to Seneca's tragedies on the left ("Tragédies"). The second dimension (29% of information) opposes strongly in the left part of the graph history (above) to poetry (below). These two dimensions can account for 65% of the inertia. As expected, the atypical works are located near the crossing of the axes; indeed, the method cannot associate with other works the only novel of the corpus (Petronius' *Satyricon*) or the only philosophical poem (Lucretius' *De natura rerum*). In this way, this CA confirms the well-known generic classifications, and its informative power is weak.

The second CA is based on more specific grammatical criteria, i.e. the distribution of verb tenses and moods.

6 Plaute = Plautus (*Amphitruo, Asinaria, Aulularia, Bacchides, Captiui, Casina, Curculio, Epidicus*); Caton = Cato (*De Agricultura*); Catulle = Catullus; Lucrèce = Lucretius; Gaules = Caesar, *Bellum Gallicum*; civile = Caesar, *Bellum civile*; 1_Discours, 2_Verrines, 3_Discours, 4_Discours, 5_Discours, 6_Discours, Philipp. = Cicero, *Orationes*, (all speeches divided into 7 chronologic groups); Traités_C = Cicero, *De Amicitia, De Officiis, De Senectute*; Salluste = Sallustius, *Catilina* and *Jugurtha*; GéorgEglog = Virgilius, *Georgicae* and *Eclogae*; Enéide = Virgilius, *Aeneida*; Horace = Horatius, *Carmina, Carmen Saeculare, Sermones, Epistulae;* Tibulle = Tibullus, Properce = Propertius, 1_Ovide, 2_Ovide = Ovidius (all works excepted *Metamorphosis*, *Tristes* and *Ponticae* divided into 2 chronologic groups), Quinte-Cur = Curtius; Consolatio = Seneca, *Consolationes*, Colère = Seneca, *De ira*; Bienfaits = Seneca, *De Beneficiis*; 1_Lucilius, 2_Lucilius = *Epistulae ad Lucilium* (divided into 2 chronologic groups); Traités_S = Seneca, all other treatises; Tragédies = Seneca, all tragedies; Juvénal = Iuvenalis; Pétrone = Petronius; Mineures = Tacitus, *Agricola, Germania, de Oratoribus*; Histoires = Tacitus, *Historiae*; Annales = Tacitus, *Annales*.

7 LASLA categories.

8 See note 5.

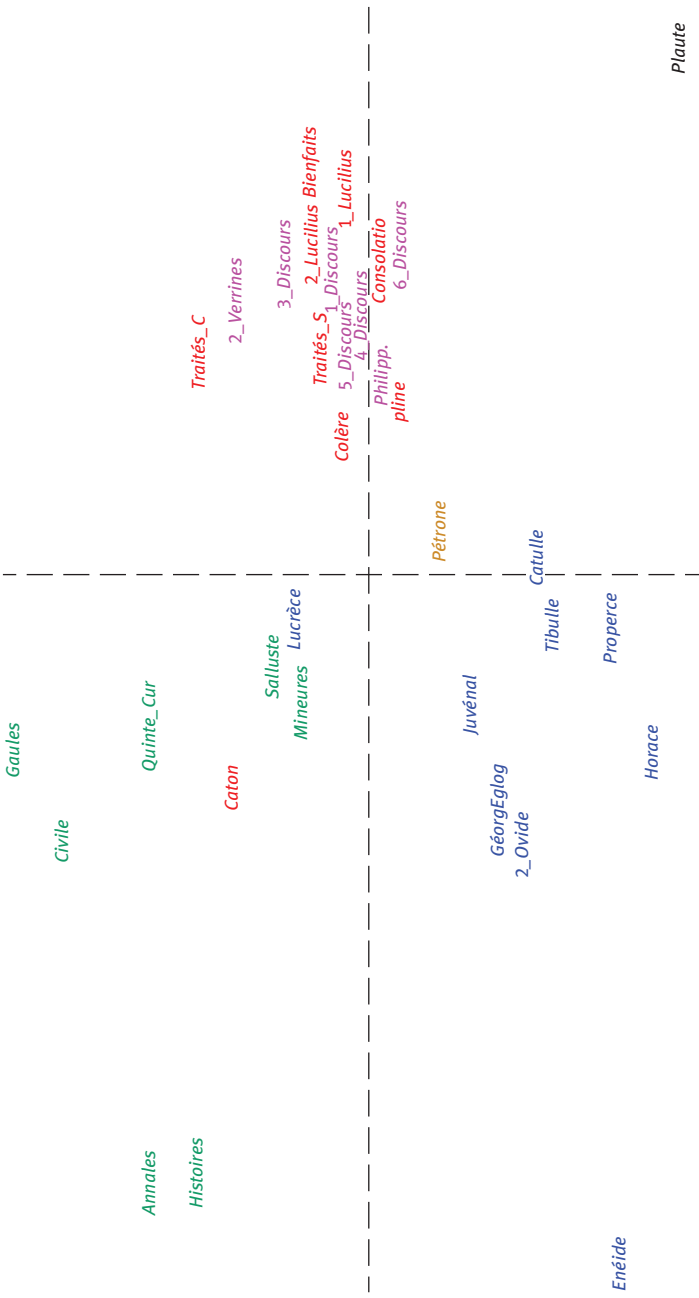


Figure 1: Correspondence Analysis – 36 textual partitions and 21 parts of speech (displaying only the distribution of textual partitions).

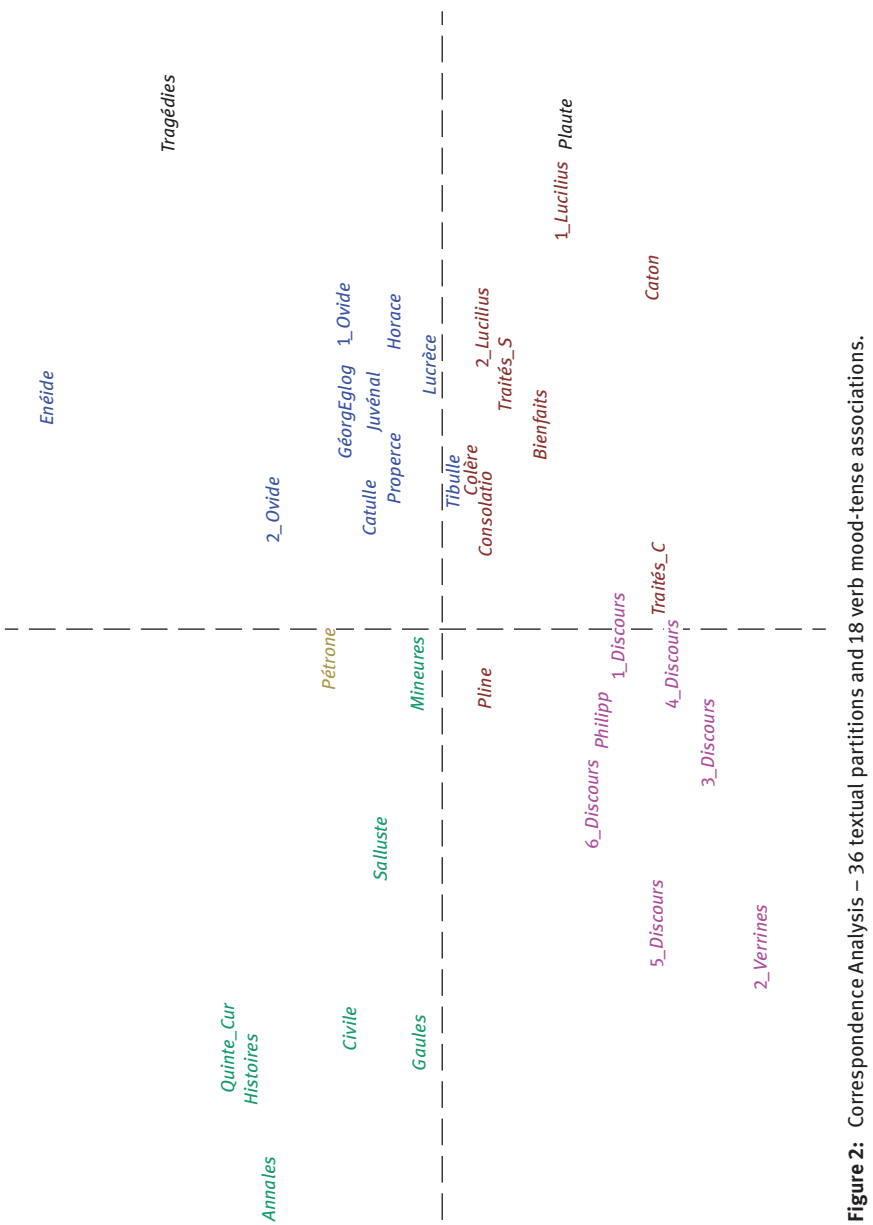


Figure 2: Correspondence Analysis – 36 textual partitions and 18 verb mood-tense associations.

This CA (Figure 2) more clearly shows groupings by text types, since each of the four most documented genres is located in a different quadrant of the graph: history to the upper left, speech to the lower left, treatise to the lower right, and poetry to the upper right. In addition, in a more refined way, we can also detect groupings by discourse modes: regarding the theatre genre, Plautus's comedies are close to the treatises (they have in common the use of the two indicative futures), whereas Seneca's tragedies are located in the same quadrant (even if off-centre) as the other versified texts. Near the crossing of the axes, we still find atypical works (Petronius's *Satyricon*, Tacitus's minor works), although Lucretius's *De natura rerum* has joined the other versified texts.

As we can see, the paradigmatic approach is effective for automatically classifying the texts, but it provides little new information to the linguist or the philologist. We thus made the assumption that introducing a syntagmatic approach would allow a more precise and original classification of the texts to be obtained and, in addition, would permit the characterization of the genre and style of each text.

3 The contribution of the syntagmatic approach

The syntagmatic approach requires the definition of new analysis objects. Various types of such objects are available (Gledhill 2007): repeated segments, clusters, phraseological phrases, syntactic constructions, anaphoric or isotopic networks, etc. Those objects can be detected by a range of methods that can be grouped into two types: supervised and unsupervised (Ganascia 2001). Supervised detection applies to objects whose interest and meaning for textual analysis are already known, according to philological tradition. Such detection allows for a noiseless inventory and a thorough statistical investigation. While the second type of method, unsupervised detection, can reveal unanticipated structures, they are not always significant (Legallois, Quiniou, Cellier and Charnois 2012; Quiniou, Cellier, Charnois & Legallois 2012).

3.1 One isolated POS vs. one POS imbedded in a syntactic structure

The first test for validating the benefit of a syntagmatic approach lies within the ambit of the supervised method. We will assess the impact on genre characterization of taking into account the integration of parts of speech (POS) and grammatical categories into syntactic structures. With this aim in view, we will study the distribution of two features across the different texts of the corpus: first,

the distribution of indicative perfects functioning as the predicate of a relative clause; and second, the distribution of the ablative forms used in the participial structure named *ablativus absolutus* 'absolute ablative'. We will compare the characterization power of these grammatical categories when considered in isolation (Figures 3 and 5) with the power of the same categories when included in the syntactic structures defined above (Figures 4 and 6).

Figure 3 shows the significant overuse[9] of Latin perfect indicative occurrences in texts belonging to various genres: history, speech, poetry, theatre, novel. This distribution does not allow a clear generic characterization of the texts. By contrast, the distribution of the perfect indicatives in the 3rd person singular functioning as predicates of relative clauses can clearly discriminate speeches and, to a lesser extent, treatises.

The seven first vertical bars with a positive value represent the speeches of Cicero, and the other positive bars the treatises of Cicero and Seneca. See examples (1) and (2), including strings of Latin perfect indicative verb forms in the 3d person singular functioning as predicates of relative clauses, mined, in (1), from a speech of Cicero and, in (2), from a treatise of Seneca:

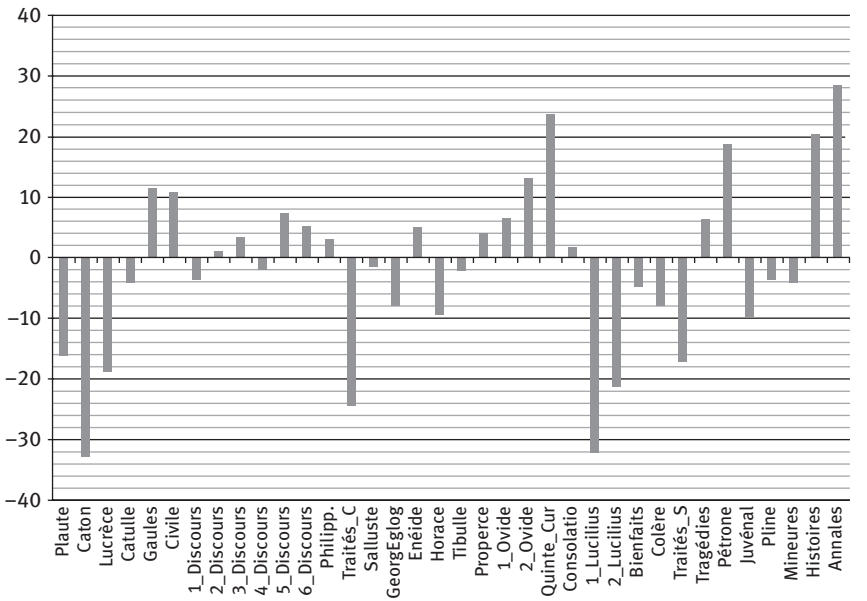


Figure 3: Distribution of Latin perfect verb forms.

9 Above the dotted line, the values have less than 5% chance of occurring by chance.

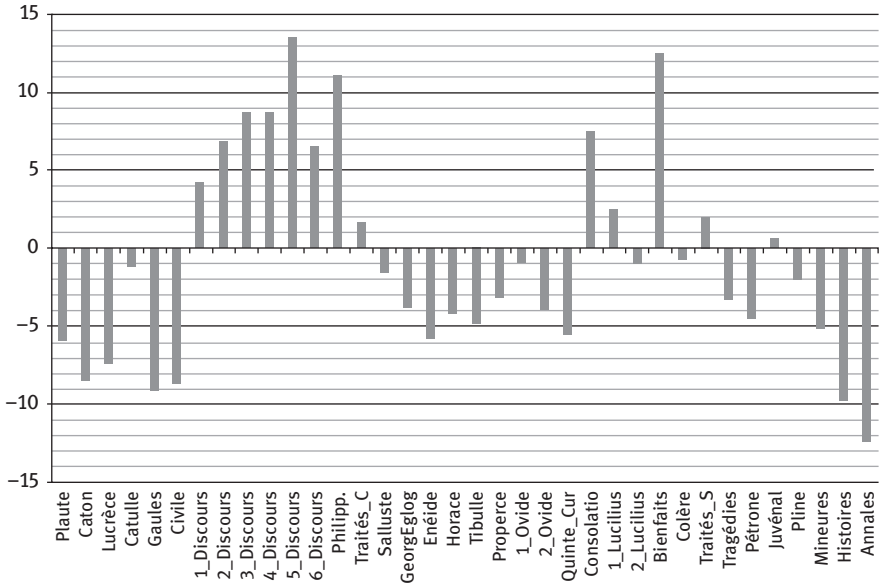


Figure 4: Distribution of Latin perfect indicative verb forms in the 3d person singular functioning as predicates of relative clauses.

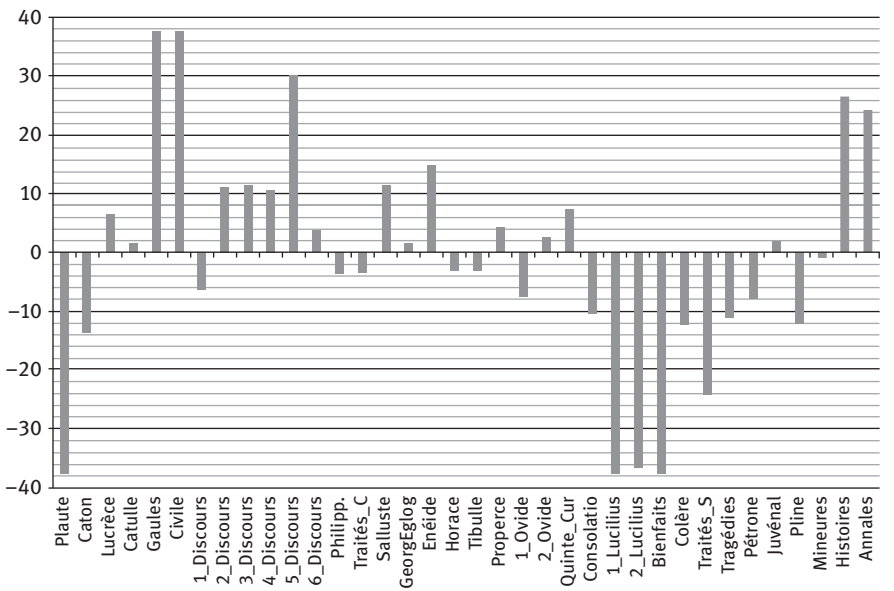


Figure 5: Distribution of ablative case occurrences.

(1) *Cuius ut omittam innumerabilia scelera urbani consulatus,* ***in quo*** *pecuniam publicam maximam* ***dissipauit,*** *exsules sine lege* ***restituit****, uectigalia* ***diuendidit****, prouincias de populi Romani imperio* ***sustulit****, regna* ***addixit*** *pecunia, leges ciuitati per uim* ***imposuit,*** *armis aut* ***obsedit*** *aut* ***exclusit*** *senatum, ut haec, inquam, omittam…*

'For, to say nothing of his countless acts of wickedness during his consulate in the city, **during which** he **has squandered** a vast amount of public money, **restored** exiles without any law, **sold** our revenues to various people, **removed** provinces from the empire of the Roman people, **given** kingdoms for bribes, **imposed** laws on the city by violence, **besieged** the senate or **excluded** from it by force of arms, to say nothing, I say, of all this…'
(Cicero, *Philippica oratio*, 7, 15)

(2) *Nonnunquam enim magis nos obligat* ***qui dedit*** *parua magnifice,* ***qui*** *regum* ***aequauit*** *opes animo,* ***qui*** *exiguum* ***tribuit*** *sed libenter,* ***qui*** *paupertatis suae* ***oblitus est,*** *dum meam respicit,* ***qui*** *non uoluntatem tantum iuuandi* ***habuit*** *sed cupiditatem,* ***qui*** *accipere se* ***putauit*** *beneficium, cum daret,* ***qui dedit*** *tamquam non recepturus,* ***recepit*** *tamquam non dedisset,* ***qui*** *occasionem qua prodesset et* ***occupauit*** *et* ***quaesiit.***

'For sometimes indeed we feel under greater obligations **to one who has given** small gifts out of a great heart, **who matched** the wealth of kings by his spirit, **who bestowed** his little, but **gave** it gladly, **who** beholding my poverty **forgot** his own, **who had**, not merely the willingness, but a desire to help, **who though** he received a benefit when giving it, **who gave** it with no thought of having it returned, who, when it was returned, **had** no thought of having given it, **who** not only **sought**, but **seized**, the opportunity of being useful.'
(Seneca, *De Beneficiis*, 1, 7)

In the same way, the distribution of the ablative case is not clearly significant, as this case can function as a marker of numerous different syntactic functions, as for instance in (3) as a marker of the complement of the intransitive verb *frui* 'to enjoy':

(3) ***Insolito spectaculo*** *fruebantur…*

'They enjoyed **the strange spectacle…'**
(Tacitus, *Historiae*, 4, 62)

In Figure 5, we only observe underuses in all of Seneca's works, in the tragedies as well as in the treatises.

On the contrary, when the ablative is used in the particular participial construction called *ablativus absolutus* 'absolute ablative', as in (4),

(4) *Galli,* ***re gognita*** *per exploratores obsidionem relinquunt…*

'The Gauls, **the matter having been discovered** through their scouts, abandon the blockade'.
(Caesar, *Bellum gallicum*, 5, 49)

its distribution (Figure 6) strongly isolates and characterizes the historical works (from left to right, Caesar's *Gallic War* and *Civilian War*, Sallustius and Quintus Curtius' works, Tacitus' *Histories* and *Annals*).

As a conclusion of those two tests, we can note that the taking into account of the syntactic dimension produces far better results, although at this stage the results are not really unexpected and do not bring new information about the generic grouping of the corpus partition. This is not really surprising: even by leaning in this way on a syntactic dimension, this method does not take into account the real sequential structure of the text's linearity. Moreover, the choice of the studied syntactic structures relies upon the already acquired knowledge of the Latinist, and their detection is always supervised.

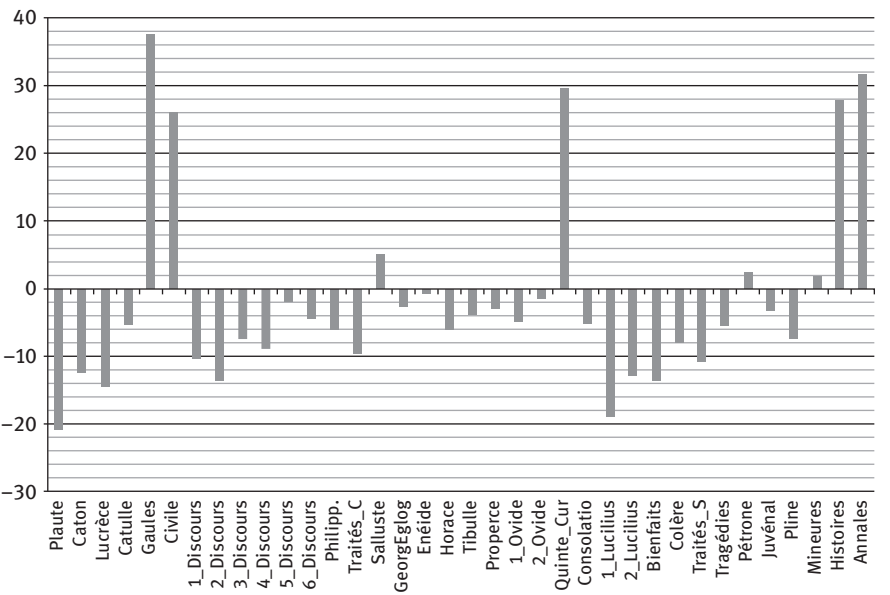


Figure 6: Distribution of the predicates (in the ablative case) of the participial construction called *ablativus absolutus* 'absolute ablative'.

Therefore, studying the distribution of one grammatical category in one given structure in a supervised way is not enough. It is thus important to also analyse strings of several automatically detected grammatical categories. With this aim in view, the texts will of course be reduced beforehand to a string of morphosyntactical tags. Leaning on the POS-part of these tags, we will automatically and in an unsupervised way mine repeated strings of three POS-tags (POS-3-grams) and submit the results to a Correspondence Analysis. Then, we will focus our research on one particular POS-3-grams.

3.2 The POS-n-grams

The first analysis with the POS-3-grams results in a CA showing a clear bipartition between prose and poetry (Figure 7).

This CA seems less informative than Figure 1 based on the distribution of isolated POS-tags, but in fact this bi-partitioned CA highlights other similarities or distances which can easily be interpreted by the philologist. For example, Lucretius' work integrates with the group of other poetic works (on the left), and Cicero's different works are more closely grouped together but without masking the distinction between treatises ("Traités_C") and speeches ("Discours", "Verrines", "Philipp").

Here, we can wonder which POS-3-grams are specific to poetry and help distinguish it from prose. When we display the POS-3-grams on the above CA (Figure 8), we observe on the far left that the POS-3-grams "cca" /adjective – adjective – substantive/ contributes strongly to this CA and shows a great proximity to the poetic works.

It was possible to confirm this by a graph distribution (Figure 9).

In this graph, there is only one non-poetic text, Tacitus's *Histories*, that presents a slightly significant overuse of the POS-3-Gram. This is not completely surprising: philologists have long emphasized the "poetical colour" of Tacitus's writing. With a longer POS-n-gram, the difference between the Poets and Tacitus is strengthened: for instance, the distribution of the POS-5-gram /adjective – adjective – substantive – verb – substantive/ (Figure 10) presents positive, significantly reduced variation for all poetic texts, while both of Tacitus' works show negative variations.

POS-n-grams can also characterize authors' styles (Longrée, Mellet & Poudat 2010). The results are far better when they take into account not only the POS, but all of the information provided by the morphosyntactical tags. For instance, the string /adverb – adjective, 1st class, Nominative singular – adjective, 1st class,

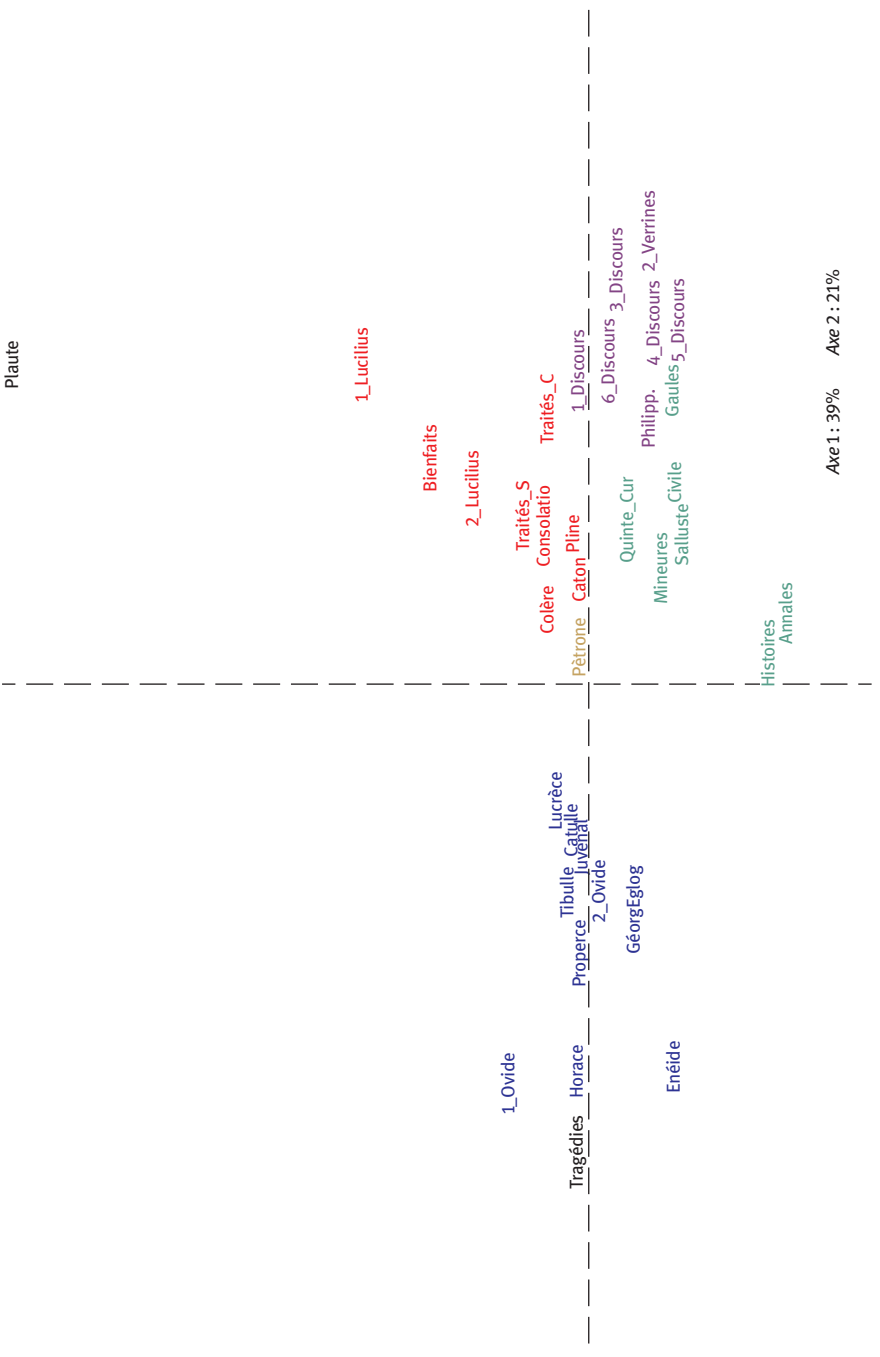


Figure 7: Correspondence Analysis – Distribution of POS-3-grams in 36 textual partitions (displaying only the texts distribution).

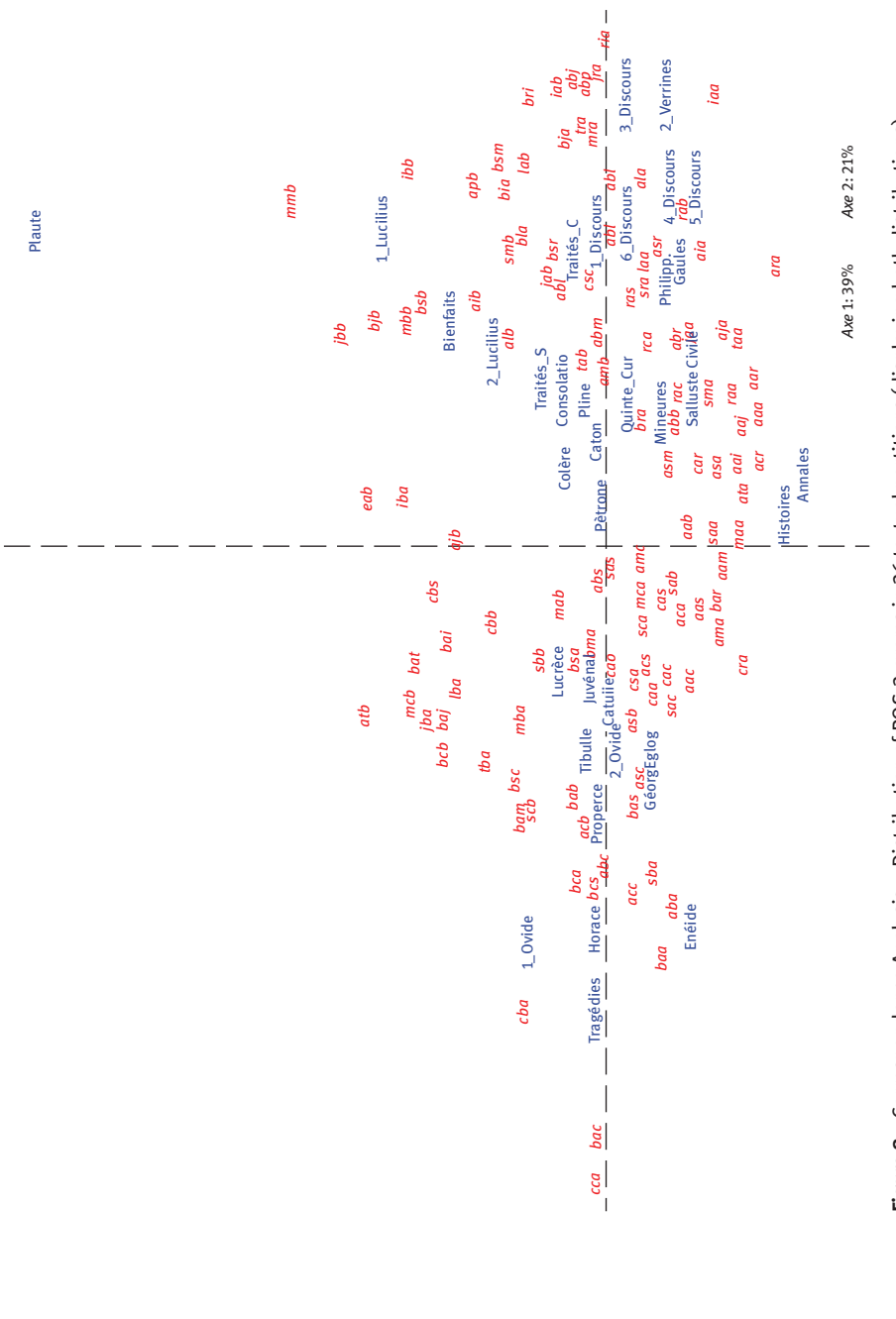


Figure 8: Correspondence Analysis – Distribution of POS-3-grams in 36 textual partitions (displaying both distributions).

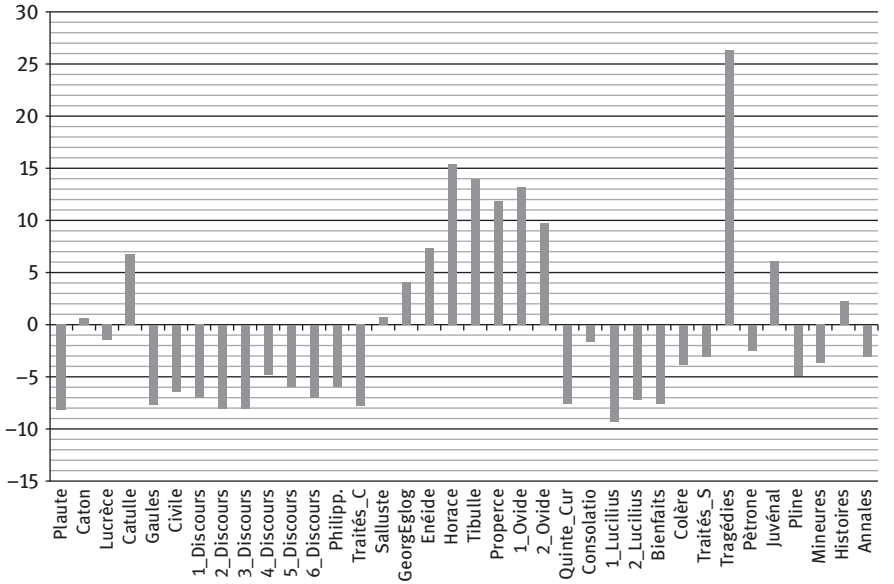


Figure 9: Distribution of the POS-3-grams /adjective – adjective – substantive.

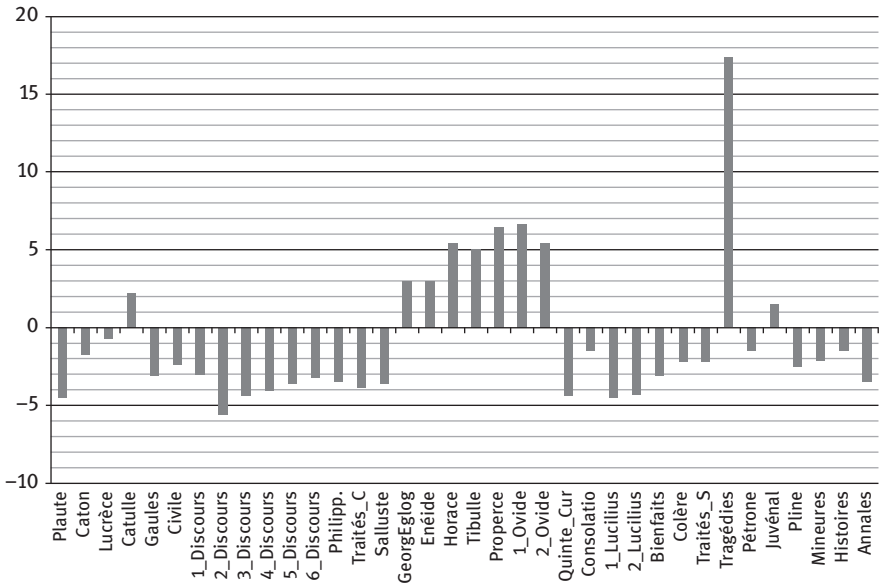


Figure 10: Distribution of the POS-5-grams /adjective – adjective – substantive – verb – substantive.

Nominative singular/ with or without coordination[10] between the two adjectives is a feature that is automatically and statistically detected[11] as characteristic of Sallustius's style, whereas this string is totally missing from Caesar's works. Therefore, this string helps distinguish between the writing of these two contemporary historians.

At the same time, the meaning of the POS-n-grams can be difficult to interpret. Indeed, POS strings do not necessarily correspond to a syntactic structure; there are not always grammatical links between the different elements of the string: for instance , the POS-4-grams /substantive 2d decl. accusative singular – coordination – preposition – substantive 2d decl. ablative singular/ is specific to historians, it does not correspond to a particular syntactic structure; the detection is the result of a pure statistical phenomenon due to the high frequencies of each of the constituents of the repeated segment. With our tagging and mining methods,[12] we extract different types of strings: some correspond to syntactic patterns, some to phraseological patterns, some are a pure succession of POS without lexical nor grammatical coherence. And for the time being, the Treebanks do not seem to offer a reliable solution.

In addition, the n-gram (or repeated string) is a frozen structure, which authorizes no variation, no addition, and no suppression. It is therefore not a suitable theoretical framework for collecting all of the various-shaped tokens in the texts of a syntagmatic pattern. Yet, the specific structures characterizing a genre or an author's style generally authorize some variations. For example, the two initial subordinate *dum*-clauses *Dum haec per provincias a Vespasiano ducibusque partium geruntur* [While these events are taking place in the provinces at the instigation of Vespasian and the party leaders] and *Dum ea geruntur* [While these events are taking place], both with the same predicate *geruntur* (indicative present of *gero*) and the same kind of subject (anaphoric pronoun neuter and plural) referring to previously narrated events are two tokens of the same pattern, but they are situated at the extreme ends of a continuum which goes from the simplest and shortest structure to the longest and the most complex. The generic pattern is characteristic and exclusive of historical prose: from the point of view of the grammar of genres, it is the same distinctive feature; therefore, to not be

10 This variation suggests that the repeated segments method is not sufficient to detect all the specific patterns of a text, style, or genre.

11 By the method of the repeated segments applied to the texts reduced to a string of morpho-syntactical tags.

12 TXM, Hyperbase, Sdmc (Sequential Data Mining under Constraints).

able to recognize and count all of its various-shaped occurrences would be a major drawback for the syntagmatic approach of genre characterization.

Finally, the n-gram corresponds to an exclusively sequential and localized approach to the text. This approach is not capable of globally comprehending the dynamics of the entire text, on which a grammar of genres and styles is based[13].

4 The notion of "motif"

We will therefore call now for the concept of "motif" in order to handle the different tokens of a given structure and to model them in one unified pattern. The identification of a unified pattern as a "motif" is legitimated by the fact that this "motif" always has the same textual function, regardless of its surface variations. By way of its repetition, the motif is indeed strongly related to the text dynamics and is one of its main meaning components.

What is a "motif"? Formally, the "motif" is defined as an ordered subset of the textual ensemble, formed by the recurring combination of n elements provided with its linear structure. Thus, if the text is formed by a certain number of occurrences of elements A, B, C, D, and E, a "motif" can be the recurring micro-structure ACD or AAA, etc., without here prejudging the nature (lexical, grammatical, metrical, ...) of the elements A, B, C, D and E in question: the 'motif' is only the framework – or the collocational pattern – accommodating a range of parameters to be defined and which are capable of characterizing the diverse texts of a corpus, or even the different parts of a text.

The concept of "motif" is one of the foundation stones of a topological approach to texts. This approach aims to account for the global dynamics of texts, both in their linearity and their reticularity (Viprey 1997, 2002a, 2002b; Legallois 2006).

The "motif" properties of recurrence and stability behind the surface variations make it a pivotal element in textual structuration. The motif is involved in particular in the temporal dynamics of the narration, in the relations between sentences, and between the different textual sequences such as descriptions, narrations, argumentations, and so on.

As general pattern, the "motif" is able to characterize a genre[14]; but its different realizations or tokens, – which we will call from now on "motif variations" – may be specific to different authors in a given genre (Longrée and Mellet 2014).

13 See the mixed conclusion in which resulted Magri and Purnelle (2012).
14 See also Stubbs and Barth (2013).

We will exemplify this assertion by characterizing, in a contrastive way, the style of two Latin historians living at either end of the classical literary period, Caesar (1st century B.C.) and Tacitus (end of 1st century A.D.). With this aim in view, we selected three books of the *Gallic War*,[15] one book of the *Civilian War*, four books of Tacitus's *Annales* as well as the *Life of Agricola*, and a biography from the same author: the selection criterion was the size of the texts, in order to make a comparison possible. In this corpus, we will mine sets of characteristic motif tokens – verb tense sequences, sentence structures – and then we will observe their distribution across the texts and their meaningful collocations.

The sentence structures we will study are amongst those Chausserie-Laprée (1969) has detected as the most characteristic of narrative sentences: the variation in their use has been analysed as a marker of the diachronic evolution of narrative expression. He distinguished two main types of narrative sentences: the so-called "typical narrative sentences," and sentences with an "appended element".

All of the "typical narrative sentences" begin with a set of syntactical structures whose function is to describe the circumstances of the main action signified in the main clause. This set forms a narrative framework for the following elements and is mainly made up of participial constructions (*ablativus absolutus*) and circumstantial subordinate clauses (*cum*-clauses in the subjunctive), as in example (5):

(5) *Postridie eius diei,* ***refractis portis****,* ***cum*** *iam* ***defenderet*** *nemo, atque* ***intromissis militibus nostris****, sectionem eius oppidi uniuersam Caesar uendidit.*

'The day after, **the gates having been broken open**, **while** no one more **defended** them, and **our soldiers having been sent in**, Caesar sold the whole spoil of that town.'
(Caesar, *Bellum gallicum*, 2, 33)

We have selected one "framing motif" made up of at least one occurrence of one of these two circumstantial elements: in our corpus, we have detected seven different sufficiently frequent realizations of this motif; the variations rely on the expansion of the motif to two, three, or four circumstantial elements and the intrusion of another element into the sequence.

All of the sentences with an "appended element" end with an unexpected circumstantial element which brings a complementary afterthought that provides more information about the action described in the preceding main clause, as in (6):

15 *Gal.* 4 and 5 written by Caesar himself and *Gal.* 8 written by one of his legates, Hirtius.

(6) *Pisonem Verania uxor ac frater Scribonianus, Titum Vinium Crispina filia composuere, / **quaesitis redemptisque capitibus** quae uenalia interfectores seruauerant.*

'For Piso, the last rites were performed by his wife Verania and his brother Scribonianus, for Vinius, by his daughter Crispina, **their heads** which the murderers had reserved for sale **having been searched out and purchased**.' (Tacitus, *Historiae*, 1, 47)

We have detected three different sufficiently frequent realizations of this "appendage motif". With the seven realizations of the "framing motif", we have a set of 10 sequences that are potentially able to characterize the texts of the corpus. The data we collected have been treated by way of a Tree analysis, which allows the visualization of the grouping of the texts according to the chosen parameter, in our case, the distribution of the 10 sequences.

This classification method (Figure 11) succeeds in regrouping all of Tacitus's works, including the *Life of Agricola*, vs. all Caesarian works including Hirtius's 8th book of the *Gallic War*.

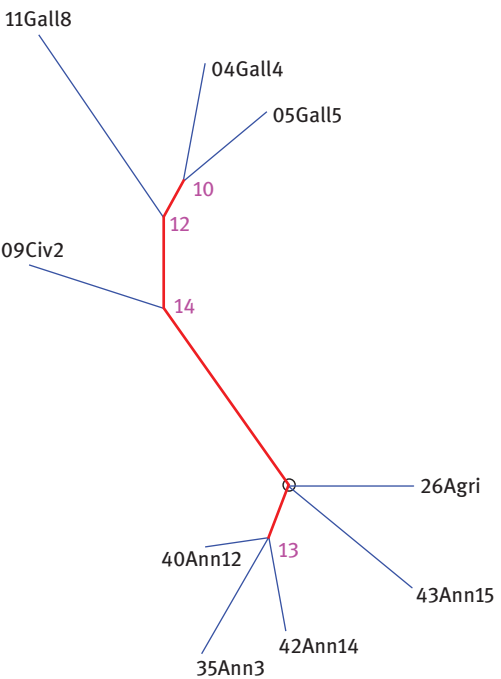


Figure 11: Classification of Caesar's and Tacitus texts according to the distribution of seven instances of a "framing motif" and three instances of an "appendage motif".

These opposite uses of the "appendage motif" can be made obvious by comparing the linear distribution of this motif, for instance, across Caesar's *Civilian War* 2 and Tacitus's *Annales 12*. We will make use here of the "neighbourhood method" (Mellet and Barthélemy, 2007; Luong, Julliard, Mellet and Longrée, 2007; Barthélemy, Longrée, Luong, and Mellet 2009), a method we borrowed from topology. We reduce the text to a chain of codes symbolizing two types of sentence: with or without an appended element, respectively code 1 and code 0. Then, we determine a contextual sliding span of an arbitrary size, here of size 11. In each sliding span, called a neighbourhood, we count the number of codes 1. This number corresponds to the density of the motif in the span and can be graphically represented. In Figure 12, we observe that the maximal density in Caesar's book is 2 and that it is 6 in Tacitus's book.[16] In addition, the global number of "appended element sentences" is far greater in Tacitus's book than in Caesar's, and is also far more regular, whereas appended elements are completely absent in Caesar's book between sentences 92 and 169.

The same method can be used to study the linear distribution of the "framing motif sentences".

Figure 13 shows a situation opposite to that present in Figure 12: the maximal density in Caesar's book is 6, and in Tacitus's book it is only 3. The global number of "framing motif sentences" is far greater in Caesar's book than in Tacitus's, and is also far more regular, while framing motifs are completely absent in Caesar's book between sentences 85 and 141.

This study highlights the capacity of our motifs to distinguish between and to characterize two different writing styles within the same literary genre. The method takes into account both a micro-syntagmatic dimension (the structure of the motifs themselves) and a macro-syntagmatic one (the overall dynamics of the text). However, it also adds a paradigmatic dimension by way of the set of so-called "motif variations": for instance, our framing motif includes seven different patterns, BCCCM, BCCM, BCCx, BCM, BCx, BxCM and BCxC, where B stands for "beginning of the sentence", C for "circumstantial element, *ablativus absolutus* or *cum*-clause", M for main clause and x for any "intruding" element. We create in this way a new closed class of syntactic structures and thereby define a new type of paradigmatic list that includes a syntagmatic dimension and that is able to characterize the grammar of an author's style. In doing this, it is truly possible to go beyond the opposition between the paradigmatic and syntagmatic approaches.

16 Please note that, for graphic reasons, the scale of theY-Axis has been increased tenfold.

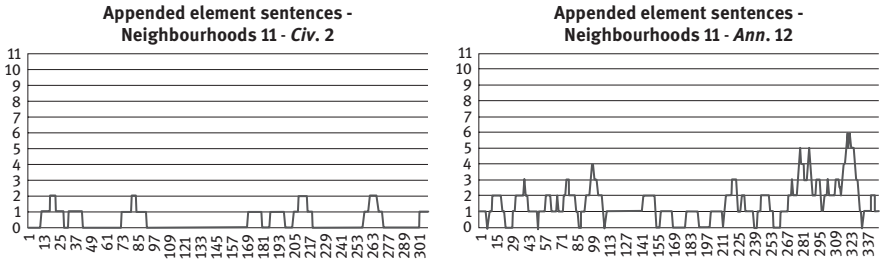


Figure 12: Linear distribution of the "appended element sentences" throughout Caesar's *Civilian War* 2 and Tacitus' *Annales* 12.

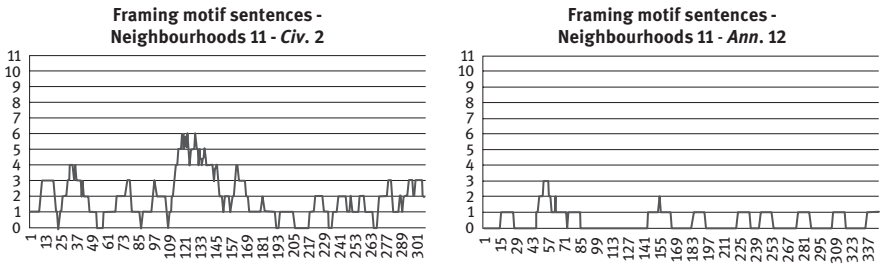


Figure 13: Linear distribution of the "framing motif sentences" throughout Caesar's *Civilian War* 2 and Tacitus' *Annales* 12.

5 Conclusion

The various analyses proposed here raise a methodological question: because they include syntagmatic strings at various linguistic levels, the motifs that we have studied imply that the texts have been reduced to various schematic layouts, such as POS-tags or clause-type tags ... We have to wonder to what extent text analysis may deconstruct and reconstruct its object.

The consistency of the results obtained through this crossed research based on various criteria at least partially legitimates this kind of deconstruction / reconstruction process. Going back to the real text brings an additional guarantee.

This methodological exploration results in the following assessment: a topological approach makes it possible to go beyond a purely sequential and localized approach to the text that is incapable of globally capturing the grammar of genres and styles. It allows the detection and mining of repeated and characteristic textual structures, as well as their grouping into paradigmatic lists of syntagmatic patterns that are able to characterize a genre or a style. This far more

effective method therefore leads to an association of both the syntagmatic and paradigmatic approaches in a cross-fertilization process that goes beyond the initially observed complementarity of the two approaches.

References

Aumont, Jacques. 1996. *Métrique et stylistique des clausules dans la prose latine. De Cicéron à Pline le Jeune et de César à Florus*. Paris: Champion.

Barthélemy, Jean-Pierre, Dominique Longrée, Xuan Luong & Sylvie Mellet. 2009. Représentations du texte pour la classification arborée et l'analyse automatique de corpus: application à un corpus d'historiens latins. *Mathematics and Social Sciences* 187 (3). 107–121.

Biber, Douglas. 1988. *Variation across language and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English. Multi-word patterns in speech and wrinting. *IJCL* 14 (3). 275–311.

Chausserie-Laprée, Jean-Pierre. 1969. *L'expression narrative chez les historiens latins. Histoire d'un style*. Paris: de Boccard.

Ganascia, Gabriel. 2001. Extraction automatique de motifs syntaxiques. In *Actes de Traitement* Automatique *du Langage Naturel* 2001 (*TALN* 2001). Tours, 2-5 juillet 2001.

Gledhill, Christopher & Pierre Frath. 2007. Collocation, phrasème, dénomination: vers une théorie de la créativité phraséologique. *La Linguistique* 43 (1). 63–88.

Kohl, Alfred. 1959. *Der Satznachtrag bei Tacitus*, Diss., Wurtzbourg.

Legallois, Dominique. 2006. Des phrases entre elles à l'unité réticulaire du texte. *Langages* 164. 56–70.

Legallois, Dominique, Solen Quiniou, Peggy Cellier & Thierry Charnois. 2012. What About Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics ? In *Lecture Notes in Computer Science*. Berlin, Heidelberg, Dordrecht: Springer.

Longrée, Dominique & Xuan Luong. 2003. Temps verbaux et linéarité du texte: recherches sur les distances dans un corpus de textes latins lemmatisés. *Corpus* 2. 119-140. http://corpus.revues.org/33 (accessed 18 May 2017).

Longrée, Dominique & Xuan Luong. 2005. Spécificités stylistiques et distributions temporelles chez les historiens latins: sur les méthodes d'analyse quantitative d'un corpus lemmatisé. In Geoffrey Williams (ed.), *La Linguistique de Corpus* (Rivages Linguistiques), 141–152. Rennes: Presses Universitaires de Rennes.

Longrée, Dominique, Xuan Luong & Sylvie Mellet. 2004. Temps verbaux, axe syntagmatique, topologie textuelle: analyses d'un corpus lemmatisé. In Gérald Purnelle, Cédric Fairon & Anne Dister (eds.), *JADT 2004, Le poids de mots, Actes des 7e Journées internationales d'Analyse statistique des données textuelles*, 743–752. Louvain-la-Neuve. http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_071.pdf (accessed 18 May 2017).

Longrée, Dominique, Xuan Luong & Sylvie Mellet. 2006. Distance intertextuelle et classement des textes d'après leur structure: méthodes de découpage et analyses arborées. In

Jean-Marie Viprey, Claude Condé, Alain Lelu & Max Silberztein (eds.), *JADT 2006, Actes des 8èmes Journées internationales d'Analyse statistique des Données Textuelles*, 643-654. Besançon: Presses universitaires de Franche-Comté. http:// lexicometrica.univ-paris3.fr/jadt/jadt2006/PDF/II-057.pdf (accessed 18 May 2017).

Longrée, Dominique, Xuan Luong & Sylvie Mellet. 2008. Les motifs: un outil pour la caractérisation topologique des textes. In Serge Heiden, Bénédicte Pincemin (eds.), *JADT 2008, Actes des 9èmes Journées internationales d'Analyse statistique des Données Textuelles*, 733–744. Lyon: Presses de l'ENS. http://lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/longree-luong-mellet.pdf (accessed 18 May 2017).

Longrée, Dominique & Sylvie Mellet. 2007. Temps verbaux et prose historique latine: à la recherche de nouvelles méthodes d'analyse statistique. In Gérald Purnelle, Joseph Denooz (eds.), *Ordre et cohérence en latin*, 117–128. Genève: Droz.

Longrée, Dominique & Sylvie Mellet. 2013. Le motif: une unité phraséologique englobante? Etendre le champ de la phraséologie de la langue au discours. *Langages* 189. 65–79.

Longrée, Dominique & Sylvie Mellet. 2014. Les variantes des motifs chez les prosateurs latins, Entre récurrence générique et spécificité d'auteur, des formes révélatrices et caractérisantes. In Dominique Longrée, Sabine Fialon & Paul Pietquin (eds.), *Langues anciennes et analyse statistique: cinquante ans après – Distances textuelles et intertextualités = Les Etudes Classiques*, 82. 65–88.

Longrée, Dominique & Sylvie Mellet. 2016. A Text Structure Indicator and two Topological Methods: New Ways for Studying Latin Historic Narratives. *Digital Scholarship in Humanities*. Oxford: Oxford University Press. doi: 10.1093/llc/fqw021. http://dsh.oxfordjournals.org/content/early/2016/04/27/llc.fqw021.full?ijkey=wDyZkoG1iV8aqRa&keytype=ref (accessed 18 May 2017).

Longrée, Dominique, Sylvie Mellet & Céline Poudat. 2010. Les taggers, auxiliaires heuristiques en ADT ? In Sergio Bolasco (ed.), *Actes des 10èmes Journées internationales en Analyse statistique des Données Textuelles,* 1195–1206. Milan: LED. http://lexicometrica.univ-paris3.fr/jadt/jadt2010/allegati/JADT-2010-1195-1206_027-Longree.pdf (accessed 18 May 2017).

Luong Xuan, Marcel Juillard, Sylvie Mellet & Dominique Longrée. 2007. Trees and after: The Concept of Text Topology. Some applications to Verb-Form Distributions in Language Corpora, *Literary and Linguistic Computing,* 22, 2, 167–186.

Magri, Véronique & Gérald Purnelle. 2012. Mot à mot, brin par brin: les suites [Nom préposition Nom] comme motifs. In Anne Dister, Dominique Longrée et Gérald Purnelle (eds.), *JADT 2012, Actes des 11èmes Journées internationales d'analyse statistique des données textuelles*, 659–673. Liège: Université de Liège. http://lexicometrica.univ-paris3.fr/jadt/jadt2012/tocJADT2012.htm (accessed 18 May 2017).

Mellet, Sylvie. 1980. Le présent "historique" ou de "narration". Quelques remarques à propos de César, *Guerre des Gaules VII* et Charles de Gaulle, *Mémoires de Guerre*. *L'Information grammaticale* 4. 6–11.

Mellet, Sylvie & Jean-Pierre Barthélemy, 2007. La topologie textuelle : légitimation d'une notion émergente. *Lexicometrica* 7. http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Magri,%20Veronique%20et%20al.%20-%20Mot%20a%20mot,%20brin%20par%20brin.pdf (accessed 18 May 2017).

Mellet, Sylvie & Dominique Longrée. 2009. Syntactical Motifs and Textual Structures. *Belgian Journal of Linguistics* 23 (New Approaches in Textual Linguistics). 161–173.

Mellet, Sylvie & Dominique Longrée. 2012. Légitimité d'une unité textométrique: le motif. In Anne Dister, Dominique Longrée & Gérald Purnelle (eds.) *JADT 2012*, *Actes des 11èmes*

Journées internationales d'Analyse statistique des Données Textuelles, 716–728. Liège: *Université de Liège. http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Mellet,%20Sylvie%20et%20al.%20- %20Legitimite%20d'une%20unite%20-textometrique.pdf* (accessed 18 May 2017).

Pawlowski, Adam. 1999. Language in the Line vs. Language in the Mass: On the Efficiency of Sequential Modelling in the Analysis of Rhythm. *Journal of Quantitative Linguistics* 6 (1). 70–77.

Quiniou, Solen, Peggy Cellier, Thierry Charnois & Dominique Legallois. 2012. Fouille de données pour la stylistique: cas des motifs séquentiels émergents. In Anne Dister, Dominique Longrée & Gérald Purnelle, *JADT 2012, Actes des 11èmes Journées internationales d'analyse statistique des données textuelles,* 821–833. Liège: Université de Liège. http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Quiniou,%20-Solen%20et%20al%20-%20Fouille%20de%20donnees%20pour%20la%20stylistique.pdf (accessed 18 May 2017).

Seitz, Konrad.1958. *Studien zur Stilentwicklung und zur Satzstruktur innerhalb der Annalen des Tacitus*. Diss., Marbourg.

Stubbs, Michael & Isabel Barth. 2013. Using recurrent phrases as text-type discriminators. *Functions of Language* 10 (1). 65–108.

Viprey, Jean-Marie. 1997. *Dynamique du vocabulaire des* Fleurs du Mal. Paris: Champion.

Viprey, Jean-Marie. 2002a. *Analyses textuelles et hypertextuelles des Fleurs du mal*. Paris: Champion.

Viprey, Jean-Marie. 2002b. Dynamisation de l'analyse micro-distributionnelle des corpus textuels. In Annie Morin, Pascale Sébillot (eds.), *JADT 2002, Actes des 6èmes Journées internationales d'Analyse statistique des Données Textuelles,* 779–790. Saint-Malo: IRISA/INRIA. http://lexicometrica.univ-paris3.fr/jadt/jadt2002/PDF-2002/viprey.pdf (accessed 18 May 2017).

Dominique Legallois, Thierry Charnois and Meri Larjavaara

The balance between quantitative and qualitative literary stylistics: how the method of "motifs" can help

Abstract: In this chapter, we study 60 novels by twelve 19th century French authors, aiming to show the complementarity between a stylistics of identification (based on stylometry techniques), and a stylistics of characterization (adopting a qualitative approach). The two stylistics belong to very different traditions, but by taking *motifs* as units of analysis, it is possible to identify some of the lexico-grammatical patterns typical of each author. The study presents in detail the method of extraction of patterns (motifs) and proposes many examples of stylistic features, especially among the novelists Hugo, Balzac, Flaubert, and Gaboriau. Very often, these features have not been identified by traditional stylistics.

At a time when computational methods of extraction of the morphosyntactic specificities of a given author or a literary genre are becoming more and more accessible thanks to new computational tools, linguists and specialists in stylistics and literary discourse can no longer remain unaware of the contribution of quantitative data. Admittedly, a quantitative approach is only one aspect of the stylistic analysis of texts, but this criterion has to be taken into account since statistically overrepresented features of language in texts can be considered as linguistic singularities, peculiar to an author, to a text or to a genre. Some analyses aiming to identify the linguistic specificity of a corpus give extremely interesting results, as evidenced by the small but growing number of monographs or articles demonstrating the approach: for example, in the French textometric tradition, Brunet (1985) on E. Zola, Magri (2009) on the comparison between travel writing and fiction, Kastberg Sjöblom (2006) on J.M.G. Le Clézio, or in other traditions, Burrows (1987) on J. Austen, Hoover (2007) on H. James, Ho (2011) on J. Fowles, Oakes (2014) for various applications, etc.

Note: We are grateful to the ANR Programme franco-allemand en Sciences humaines et sociales (FRAL) – Projet PHRASEOROM – for financial support.

Dominique Legallois, University of Paris 3 Sorbonne Nouvelle
Thierry Charnois, University of Paris 13
Meri Larjavaara, Åbo Akademi University

https://doi.org/10.1515/9783110595864-008

However, computational stylistics remains marginal for "academic" literary specialists in stylistics, despite the current development of digital humanities. The reasons for this marginal position may stem from the complexity of the methods used, the technical nature of analyses, or a difference of "culture" between text practitioners. There may however be other, deeper reasons than these: namely, the fact that statistical and automated analyses are invariably paradigmatic. They concern *discrete* descriptors, that is to say, units (lexical forms, lemmata, morpho-syntactic categories, punctuation marks) with no direct syntagmatic relations between one another, even if correlations can always be calculated. The contextual interpretation of these descriptors remains vague, therefore, since the main focus is commenting on tables of quantified units, rather than utterances. Highlighting the specificity of a descriptor is of course essential, but to grasp the role of the descriptor in its textual environment is a step that is still too rarely taken. This is the limitation of the discrete approach and may explain why its adoption by specialists in stylistics has been rather lukewarm so far.

The main aim of this contribution is to show how the study of an important area within stylistics, namely the characterization of an author's style, can benefit from a new method in corpus linguistics, the discovery of sequential patterns or "motifs", i.e. contiguous strings of word forms/lemmas/POS tags. Motif analysis can be viewed as complementary to discrete approaches and constitutes a more powerful paradigm than other non-discrete analyses such as lexical bundles or clusters because motifs combine annotations at different levels of abstraction.

As will be seen from the analyses given below, this type of complex unit, while not claiming to be exhaustive since a style cannot be reduced to a set of lexico-grammatical patterns, nonetheless provides a more accurate vision of the stylistic characteristics of an author.

The study is based on a corpus of 60 novels by 12 19th-century French writers (see below). The aim is to identify the syntactic motifs favored by each author.

In the first part of the chapter, we examine the issue of the definition of style proposed in a recent article by Herrmann and colleagues (Herrmann et al. 2015), in which the quantitative approach features prominently. Through a critical discussion of this study, we show that there are in fact two major trends in stylistics, and even two types of stylistics: a stylistics of identification, which has its roots in the stylometry of the nineteenth century, and a stylistics of characterization. The first one (essentially quantitative) is a stylistics based on discriminatory features; the second (essentially qualitative) is interpretative. In order to ensure that quantitative and qualitative approaches are complementary, it is necessary to take not only discrete units, but also non-discrete units such as motifs into account in the analysis.

The second part deals with the presentation of the method of motifs. We formulate a definition of the term and we introduce three possible statistical

methods of analysis: an endogenous method, an exogenous method, and a combination of the two.

In the third part, we apply the method of motifs to a literary corpus, and we extract some of the syntactic patterns favored by each author. For reasons of space, the motifs of only five authors (Balzac, Hugo, Gaboriau, Stendhal, and Flaubert) will be discussed.

The fourth part focuses on the qualitative interpretation of motifs. Taking the example of Hugo, we show that some groups of motifs contribute to realizing the same aesthetic project.

1 Stylistics, stylometry and style

1.1 Two stylistics: identifying and characterizing

Stylometry and corpus stylistics are still unpopular among literary stylisticians. The quote below, although formulated 25 years ago, still resonates in the "non-computational" stylistic tradition:

> When stylistic features of a text have been transformed into numerical form, they acquire a status that actually prevents them from being perceived as language-for-communication as such. That is to say, in the very act of transforming textual qualities into counts, their essential process-like character is irretrievably lost. [. . .] Thus no level of (mathematical) sophistication is able to overcome the problem that the processes of meaning constitution have been eliminated before the analysis is undertaken. (Van Peer, 1989: 302)

The problem of the reception of stylometry can be explained therefore by the loss of the qualitative dimension in data interpretation. To solve this problem, it is necessary to take both quantitative and qualitative orientations into account when considering the notion of style.

As it is impossible in this contribution to give an overview of the history of stylistics, even less of style, we will draw on a very recent and interesting study by Herrmann et al. (2015) that relates the development of the notion of style in three different traditions, those of German, Dutch and French language and literary studies, since 1945. The paper identifies 6 basic conceptions of style across these traditions. Style can be seen:

1. as constituting a higher-order artistic value (assessed through aesthetic experience);
2. as a holistic gestalt of single texts: style cannot be reduced to descriptive categories or classes;

3. as an expression of the individuality, subjectivity and/or emotional attitude of an author or speaker;
4. as an artifact that presupposes (hypothetical or factual) selection/choice among a set of (more or less synonymous) alternatives;
5. as a deviation from some type of norm, involving (quantitative or cognitive) contrast;
6. as any property of a text that can be measured computationally. (Herrmann et al. 2015: 30)

After considering the various criticisms levelled at these conceptions, and how they have been addressed in the three different traditions, the authors propose a definition of style that is an attempt to provide a common ground for both mainstream and computational literary stylistics:

> Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively. (Herrmann et al. 2015: 44)

While this definition may be relevant for a conception of style that encompasses any textual genre[1], we claim that it does not seem sufficiently precise for the analysis of literary style, for three reasons.

Our first criticism is that the vast majority of computational approaches are based on comparative methods (comparisons of genres, or authors, or texts). This comparative approach is usual for a stylistics whose aim is identification: identification of the author, the date of the text, or a particular type or school (Holmes 1998). As is well known, the primary aim of stylometry is identification. The term stylometry was coined in 1897 by the Polish philosopher Lutosławski (1863-1954)[2], who worked on the dating of Plato's dialogues (Lutosławski 1897a, 1897b). The concept of stylometry had of course some forerunners before the term *stylometry* itself became established, but Lutosławski was the first to define stylometry as the science of measuring stylistic affinities:

> of two works of the same author and of the same size, that is nearer in time to a third, which shares with it the greater number of peculiarities, provided that their different importance is taken into account, and that the number of observed peculiarities is sufficient to determine the stylistic character of all three works (Lutosławski 1897a: 152)

1 The authors defend a conception of style which is assumed to apply to any text (whether literary or not).

2 Pawłowski and Pacewicz (2004).

Lutosławski considered that the stylistic characteristics of an author constitute a kind of fingerprint that escapes the author's consciousness. As such, the features are objective and not subject to interpretation. Although it is similar to a fingerprint, the style of an author can obviously evolve.

One must however keep in mind a statistical principle due to the specificity of language: comparative methods *necessarily* produce results, that is to say, they systematically highlight differences (Kilgarriff 2005; Loiseau 2010). Statistical methods guarantee the detection of identifying features, but not necessarily that of characterizing features.

The difference between these two types of features is fundamental. For example, in our corpus of French 19th century writers (see Section 2.1 below for the composition of the corpus), the forms *ça* (this/that) and *on* (we/one), and verbs in the imperfect (imparfait) are over-used in Zola compared to the eleven other writers. How is one to interpret this result? While these data may help to identify a novel by Zola, do they say anything about the style of this author? Simply knowing that *ça* and *on* are comparatively overused in the novel, is not enough then. Without a precise examination of these items in context or a concordance analysis, it is difficult to give these units a status: over-used features (keywords) do not necessarily have an interpretative value.

The distinction between identifying and characterizing features was theorized by the French linguist Rastier (2001) who (in reference to the work of Carlo Ginzburg) differentiated morellian features from spitzerian features. The former owe their name to the Italian doctor Morelli, who revolutionized the attribution of paintings in the late nineteenth century by identifying clues, including anatomic features such as ear lobes, which are unattended to by counterfeiters and experts. Morellian features are not necessarily interpretable, they simply make it possible to identify a text, an author, or a genre. Spitzerian features, from the name of the stylistician L. Spitzer, contribute to characterizing a text, an author, or a genre: they should be considered as creating a system and expressing an aesthetic project. That is the reason why – and this is our second criticism – the definition by Herrmann et al. lacks the notion of the convergence and interaction of features. For formal features to be considered as interpretive features, it is necessary to show that they create more or less closed systems of interrelated elements. Another criticism is the disjunction "quantitatively *or* qualitatively". We argue to the contrary that a stylistic analysis concerning a text, a type, or an author, may be exclusively quantitative if its aim is identification; if the aim is characterization, however, quantitative and qualitative approaches are necessary and complementary: a qualitative approach cannot claim to analyze one or more features if they are not characteristic in terms of frequency (over-used or under-used in the corpus).

Traditional qualitative analysis cannot allow any shortcuts and must use at least a minimal amount of quantitative information. Moreover, a mere quantitative

approach would remain purely descriptive, and does not constitute a literary stylistic analysis. To highlight correlations between features and dimensions (by factorial analysis) is not enough to discover aesthetic effects and communicative intentions.

We therefore replace the definition of style given above by the following one, which applies to literary stylistics that aims at characterization:

> Style is a property of texts constituted by an ensemble or several ensembles of interrelated formal features which can be analyzed quantitatively AND qualitatively.

There again, this definition may not be sufficient; it provides nevertheless a consistent frame of analysis for our study of the 12 writers.

1.2 Discrete and non-discrete descriptors

We call "descriptors" all the units generally considered in stylometric analysis, regardless of their nature, i.e. word forms, lemmata, POS, but also average sentence or word length, average word length, type/token ratio (vocabulary richness), and vocabulary growth (homogeneity of text). These descriptors are discrete since they are atomic units. They are completely relevant as morellian features; they can also contribute to characterizing style, but they are difficult to interpret because these units are decontextualised abstractions, even if correlations between them are possible. More syntagmatic units or non-discrete units (continuous or discontinuous) are still little used in analyses of style. Salem (1987) recommended considering repeated segments, that is to say repeatedly occurring sequences of words. These segments are also referred to as 'n-grams' or lexical bundles or clusters. Other non-discrete units such as P-Frames (Fletcher, 2003) are more flexible than lexical bundles, since they provide systematic groupings of lexical bundles, which vary in only one position. Repeated segments and P-Frames have proved to be relevant and sufficient for the study of phraseology. Römer (2010) takes these units into account to establish the phraseology of a text. This kind of research is centered on particular genres or academic registers, but there are few studies that take repeated segments into account in a literary stylistic perspective (see however Mahlberg 2013 on Dickens). Table 1 shows the key lexical bundles in *Madame Bovary* by Flaubert with respect to the eleven other writers. Three association measures are used: Calculation of Specificities (Lafon 1984), log-likelihood[3], and T.score.

3 See Bertels and Speelman (2013) for a comparison between log-likelihood and Specificities.

Table 1: The first 12 key lexical bundles in *Madame Bovary* by Flaubert.

key lexical bundles	sub freq	tot freq	specificities	loglikelihood	t.score
de_temps_à_autre,	31	157	84.53	165.87	5.41
les_uns_après_les	9	38	27.50	51.69	2.93
ce_n'est_rien_!	6	9	26.51	50.54	2.43
elle_se_mettait_à	5	5	25.79	51.64	2.22
se_mit_à_lui	8	34	24.53	45.82	2.76
au_clair_de_lune,	5	7	22.74	43.28	2.22
;_et,_à_travers	4	4	20.63	41.31	1.99
à_la_croix_rouge,	4	4	20.63	41.31	1.99
dans_la_côte_du	4	4	20.63	41.31	1.99
haut_de_la_côte	4	4	20.63	41.31	1.99
se_passa_la_main	4	4	20.63	41.31	1.99
temps_à_autre,_comme	4	4	20.63	41.31	1.99

However, the question of the granularity of the linguistic forms inevitably arises: lexical bundles can be considered too specific since they are multiword sequences. They do not permit any generalization. For example:

(1) il eut un geste de (lit. He had a gesture of)[4]
(2) elle fit un mouvement de (lit. She made a movement of)

are two different lexical bundles, but it would probably be illuminating to group them together as instances of a more abstract unit, in order to investigate the syntactic characteristics of a text. A part of speech n-gram could be useful in an identification task. For example:

PRONOUN + VERB + DET + NOUN + PREP

could be a 5-gram POS. The problem is that such n-grams are too generic, and may match sentences that are very different from (1) and (2)[5], such as:

(3) il mangea une pomme en (deux minutes)
Lit. He ate an apple in (two minutes)

In view of these respective drawbacks, it is necessary to adopt a mixed and partly non-supervised method that we call the *motif method*. "Motif" is a term that is rather problematic since it is used by several scholars with different

4 "Lit." means that the English translation is a literal one.
5 This method can however give satisfactory results; cf. Frontini et al., this volume.

but closely-related meanings (Ganascia 2001, Quiniou et al. 2012, Köhler 2015, Longrée and Mellet 2013, Longrée and Mellet, this volume).

2 Motifs: definition and method

2.1 The corpus

The literary corpus consists of 60 novels by 12 19th century French authors. Five novels per writer were selected:

Balzac (*La Cousine Bette, Les Illusions perdues, Le Lys dans la vallée, Le Colonel Chabert, Le Père Goriot*), Dumas (*Les trois mousquetaires, Vingt ans après, Les Quarante-cinq, La Reine Margot, Le Prince des Voleurs*), Flaubert (*Madame Bovary, Les trois contes, L'éducation sentimentale, Bouvard et Pécuchet, Salammbô*), Gaboriau (*Le Crime d'Orcival, Le Dossier 113, Monsieur Lecoq, La Clique dorée, L'Affaire Lerouge*), Hugo (*Notre Dame de Paris, Les Misérables, Les Travailleurs de la mer, L'Homme qui rit, Quatre-vingt-treize*), Huysmans (*Les Sœurs Vatard, À rebours, En route, Là-bas, La Cathédrale*), Maupassant (*Bel Ami, Fort comme la mort, Pierre et Jean, Mont Oriol, Une vie*), Sand (*La petite Fadette, La Mare au diable, Indiana, Consuelo, François le Champi*), Stendhal (*Le Rouge et le Noir, Armance, La Chartreuse de Parme, Lucien Leuwen, Le Vert et le Rose*), Sue (*Le Juif Errant, Atar-Gull, Le Morne au Diable, Les Mystères de Paris, Mathilde – Mémoires d'une jeune femme*), Verne (*De la Terre à la Lune, 20000 lieues sous les mers, Deux ans de vacances, L'Île Mystérieuse, Michel Strogoff*), Zola (*Nana, L'Œuvre, l'Assommoir, la Terre, Germinal*).

2.2 The method

The motif method is relatively simple. It consists of several steps: specific annotation, the extraction of segments, statistical calculation of segments, and finally identification of motifs from segments.

1st step
Identifying patterns that may be composed of different types of elements: word forms, lemmata and POS. Specifically, the results of a tagger (the French language tagger *Cordial*[6]) are modified with regular expressions as follows:

6 http://www.cordial.fr/

- we keep invariable forms such as prepositions, conjunctions, frequent adverbs, etc.;
- we keep the lemmata of several frequent verbs (aspectual and modal verbs, auxiliary verbs, etc.) and reduce personal pronouns to their canonical forms;
- we keep the POS labels (common noun, proper noun, adverb, verb, adjective, etc.) of the other words. Verb tense categories are kept. Some semantic categories for nouns and adverbs are also retained. For example, adverbs of manner are labelled ADVMAN, nouns expressing parts of the body are labelled NCCOR, abstract nouns NCABS[7].

For example, this extract (from *Madame Bovary*)

> *Nous étions à l'étude, quand le proviseur entra, suivi d'un nouveau habillé en bourgeois et d'un garçon de classe qui portait un grand pupitre. Ceux qui dormaient se réveillèrent, et chacun se leva comme surpris dans son travail.*
>
> (we were at prep when the Headmaster came in, followed by a 'new boy' not wearing school uniform, and by a school servant carrying a large desk. Those who had been asleep woke up, and we all rose to our feet as though we had been interrupted at our work.)[8]

is annotated as follows:

nous être à le NC, quand le NC VPS , PASS de un NC PASS en NC et de un NC de NC qui VIMP un ADJ NC . celui qui VIMP se VPS , et chacun se VPS comme ADJ dans DETPOSS NC.

2nd step

Extracting segments of variable length for each novel. For example, see Table 2 from *Madame Bovary.*

Only segments which appear at least twice in all the novels by each author are kept.

3rd step

Calculating the segments.

This step concerns statistical calculation. There are three possible methods:

1 For every author, segments are extracted on the basis of Mutual Information (MI); calculation is therefore said to be endogenous because it does not presuppose a comparison between texts. See table 3 for an example from Balzac.

7 The lexicon of abstract nouns is based on syntactic tests; for example: *faire preuve de N* (to show N), *manifester du N* (lit. demonstrate N), *ressentir du N*, *éprouver du N* (to feel N).

8 Translation by M. Mauldon (Oxford world's classics).

Table 2: Segments from *Madame Bovary*.

Segments	Rank	Frequency	Segments
le NC de le NC	1	918	le NC de le NC
NC de le NC ,	2	428	NC de le NC ,
le NC de le NC ,	3	353	le NC de le NC ,
à le NC de le	4	244	à le NC de le
	...		
avec de le NC ADJ	1127	9	avec de le NC ADJ
avec le NC de DETPOSS	1128	9	avec le NC de DETPOSS
ce être comme un NC	1129	9	ce être comme un NC
	...		
plus ADJ de le NC	2525	6	plus ADJ de le NC
plus ADJ et plus ADJ	2526	6	plus ADJ et plus ADJ
	...		
si ADJ , si ADJ ,	2540	6	si ADJ , si ADJ ,
	...		
VPS un NC , un NC ,	10050	3	VPS un NC , un NC ,

Table 3: The top 12 segments in Balzac extracted with MI.

Segments	Frequency	Mutual Information
je ne savoir quoi de	18	23,175499
. INT ! oui ,	23	20,530413
je ne avoir jamais PASS	17	19,915905
je ne vouloir pas être	11	19,040236
! dire il en PRES	26	19,010582
. je ne avoir jamais	12	18,271523
NC pour le ADJORD fois	10	18,194517
NC . INT ! oui	10	18,152967
, je ne vouloir pas	18	17,848886
, dire il en PRES	85	17,552335
il se mettre à INF	13	17,337274
environ NUM NUM NC de	5	17,289822

Note that Mutual Information tends to highlight strong associations between elements without taking into account how frequent, or infrequent, the associations between the elements are.

2 For every author, segments are extracted with their raw frequency. Then, the frequency of the segments of every author is compared with that of the others; the calculation is exogenous since it is based on a comparison. Calculation of Specificities is therefore used (see Table 4).

Table 4: The top 12 segments in Balzac calculated with Specificities.

Segments	Specificities	sub freq	tot freq
le NC de le NC	Inf[a]	3896	55482
NC de le NC ,	Inf	1349	21739
le NC de DETPOSS NC	Inf	1145	11015
le NC de le NC ,	Inf	982	16062
à le NC de le	Inf	891	13822
NC de le NC .	Inf	798	13030
à le NC de le NC	Inf	737	11664
le NC , le NC	Inf	697	10418
le NC de un NC	Inf	686	7730
NC de le NC de	Inf	662	8995
le NC de le NC .	Inf	604	9796
le NC de NP ,	Inf	566	6404

[a] "Inf" means that the association is very strong.

Table 5: The top 12 segments in Balzac (extracted with MI and calculated with specificities).

Segments	Specificities	sub freq	tot freq
je ne savoir quoi de	Inf	18	18
! dire il en PRES	Inf	26	42
, dire il en PRES	Inf	85	126
ne pouvoir être PASS que	Inf	12	12
dire il en PRES DETPOSS	Inf	18	26
dire il en PRES le	Inf	49	49
un NC ... INT !	Inf	13	13
, dire il en se	Inf	40	61
en lui PRES un NC	Inf	21	39
, dire il avec un	Inf	12	12
, VPS il en PRES	Inf	15	15
je me être PASS ,	Inf	16	16

3 The previous two methods can be merged: segments calculated by MI are extracted for each author, and then the authors are compared with one another in order to define the key segments (see Table 5).

In sum, various statistical strategies can be adopted to present different views of the data. In the present analysis, all three strategies were used.

If the aim is identification, the non-discrete method of sequences proves to be much less competitive than the discrete method. A clustering analysis (with

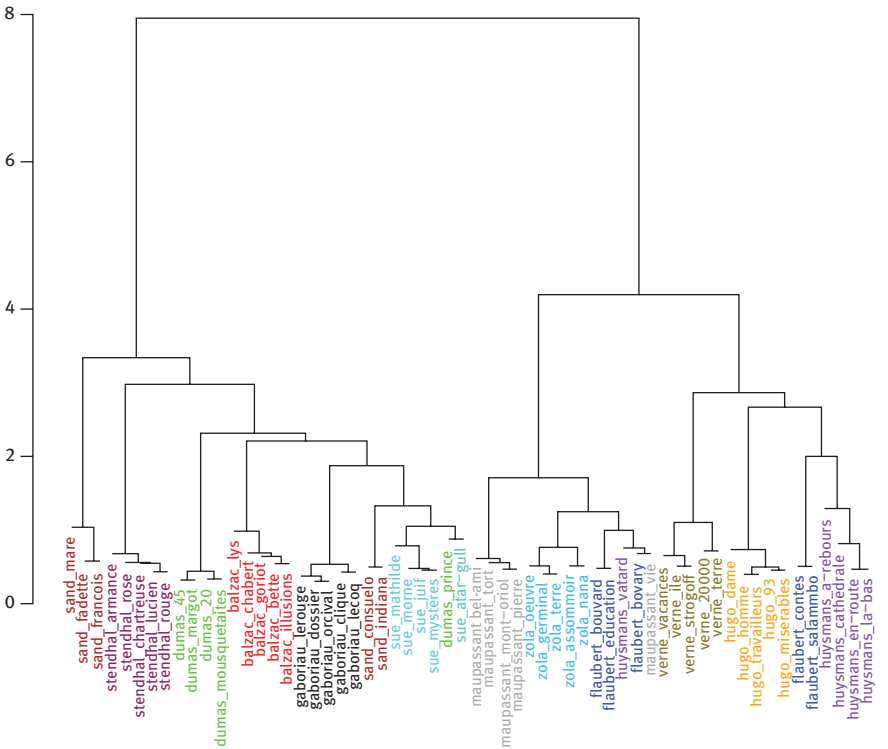


Figure 1: Cluster analysis of the 60 novels based on discrete descriptors.

the classic delta distance)[9] based on discrete descriptors (and not on sequences) gives a consistent representation of the corpus (see Figure 1).

The two main clusters conform to a time division: the former includes Stendhal, Sand, Sue, Balzac, Dumas, that is to say, writers of the first half of the 19th century. The latter includes all the writers of the second half. Only Gaboriau (1832-1873) belongs to the first cluster. This clustering can be considered to provide an acceptable consistency (relative to the chronology).

On the other hand, a clustering based on sequences of five elements gives less consistent results[10] (see Figure 2).

In the first main cluster, novels from the first half of the century (plus Gaboriau) are again grouped together, but in a sub-cluster one also finds novels by

9 The R package Stylo was used. See Eder et al. (2016).

10 The clustering is automatically made from sequences and not from motifs.

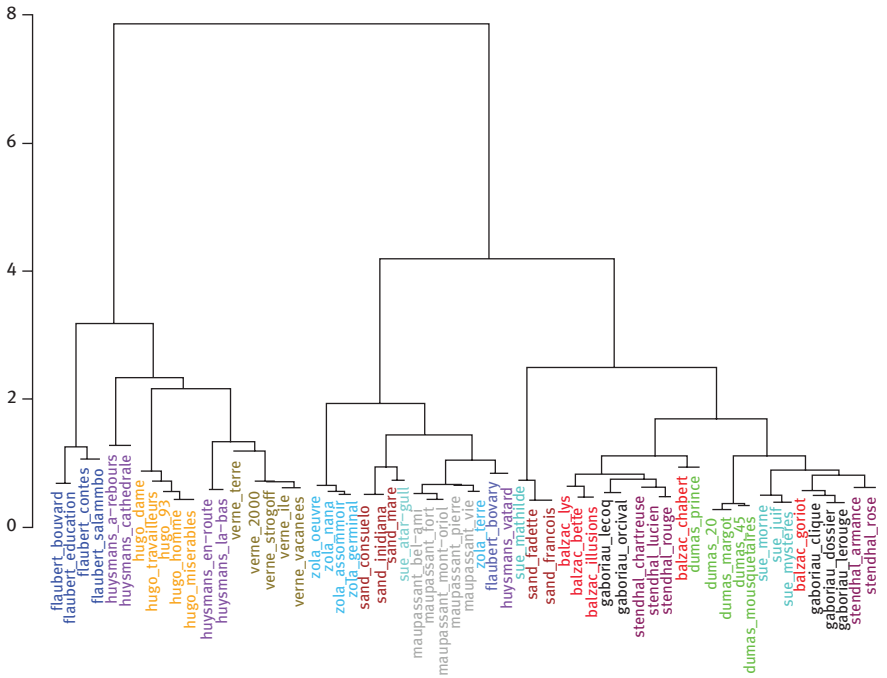


Figure 2: Cluster analysis of the 60 novels based on 5-grams.

Maupassant and Zola (second half of the century). The second main cluster remains homogenous.

The clustering based on seven elements (7-grams) is rather similar to the previous one, but some novels change places (see Figure 3). For instance, *Madame Bovary* was isolated in the five-elements based clustering while in this clustering, it is now grouped with the other novels by Flaubert. This demonstrates that clustering is relatively sensitive to the size of sequences and that sequences are perhaps not the best unit to use when undertaking a stylistics of identification.

4th step

Whatever the method, it is necessary not to confuse segments and motifs. We consider as motifs only the sequences that are syntactically well formed and interpretable (because, for example, the same lexical paradigm is present in the pattern), or that have a recognizable functional or expressive role. For example, Zola uses the sequence **le NC ADJ de un NC qui VP** in order to characterize a character's typical behavior:

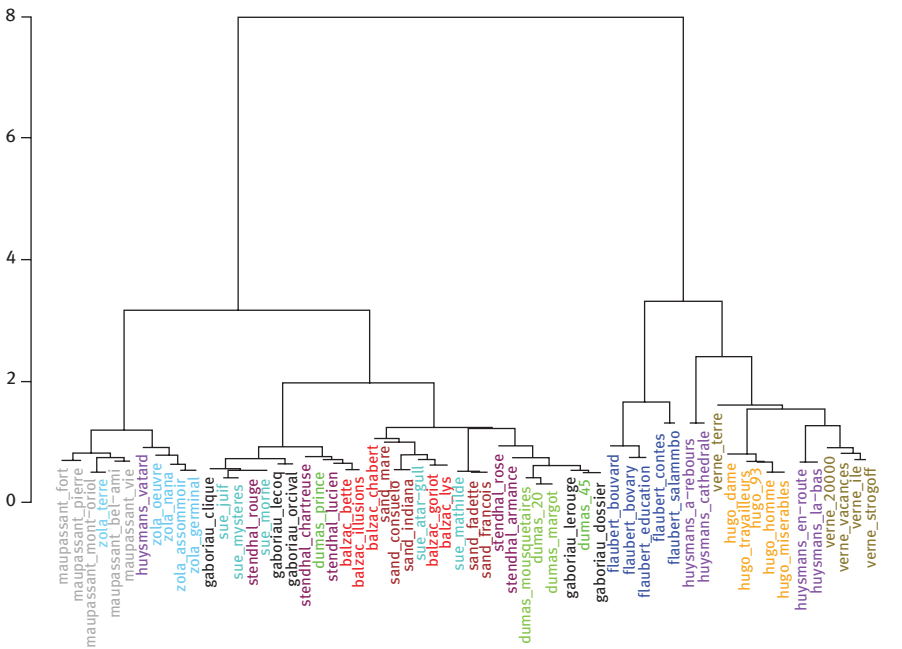


Figure 3: Cluster analysis of the 60 novels based on 7-grams.

(4) Maheu seul marchait pesamment, butait à chaque tour contre le mur, de **l'air stupide d'une bête qui ne voit plus sa cage** (*Germinal*)

Maheu alone was walking heavily up and down the bare room, stumbling against the wall at every turn, with **the stupid air of an animal which can no longer see its cage**

(5) Bijard, (...) roulait toujours la tête, **du mouvement ralenti d'un animal qui a de l'embêtement**. (*L'Assommoir*)

Lit. Bijard was still rolling his head **with the slow movement of an animal that is being goaded**

(6) Il resta quelques minutes encore, se pâma devant d'autres études, fit le tour de l'atelier avec **les coups d'œil aigus d'un parieur qui cherche la chance** (*L'œuvre*)
He remained for a few minutes longer, going into raptures before other sketches, while making the tour of the studio with **the keen glances of a speculator in search of luck**

Very interestingly, Mahlberg (2013) shows that the five-word cluster "with the air of a" appears 46 times in Dickens:

(7) ...and offered Mr. Pickwick a pinch of snuff *with the air of a* man who had made up his mind to a Christian forgiveness of injuries sustained (*Pickwick Papers*).

This cluster is functionally identical to Zola's motif but our method can identify a more lexically flexible and variable pattern than a simple lexical bundle.

It should be noted that the method generates a very large number of sequences (hundreds or even thousands for a single novel); distinguishing the motifs is therefore very time-consuming.

However, a sequence such as **le NC, et,** for instance, cannot be a motif, because it is not coherent and interpretable; for example, this sequence in Verne's novels:

(8) Nous marchions directement vers l'**Ouest, et,** le 11 janvier, nous doublâmes ce cap Wessel (*Vingt-mille lieues sous les mers*)

We were traveling due **west, and,** on January 11 we rounded Cape Wessel

(9) Alors Cyrus Smith éleva **la voix, et,** à l'extrême surprise de ses compagnons, il prononça ces paroles (*L'île mystérieuse*)

Then Smith raised **his voice, and,** to the extreme surprise of his companions, pronounced these words

In general, sequences extracted by MI are best able to be interpreted as motifs.

3 Results

For reasons of space, it is not possible to give exhaustive results, nor even to give a satisfactory synthesis of motifs for each of the twelve authors. We will therefore present an illustrative but very limited set of results for only five novelists: Balzac, Hugo, Stendhal, Flaubert, and Gaboriau, the author of popular novels. The motifs are in bold. The discourse function (DF) they express is in italics. All are statistically over-represented in the author's work. Their frequency in the author's texts (x) and their frequency in the entire corpus (y) are indicated as a ratio (x/y).

3.1 Balzac

DF: Reducing a set of references to an evaluative category

– , **enfin tous le NC** (+ extension) (lit. That is to say all the N/in short all of the N) 8/8

This motif is typical of Balzac since it was not identified in the other authors' novels. It is a formula in which the noun phrase subsumes, at the same time as it qualifies, all of the previously mentioned references:

(10) Rien ne démontrera mieux la singulière puissance que communiquent les vices, et à laquelle on doit les tours de force qu'accomplissent de temps en temps les ambitieux, les voluptueux, enfin tous les sujets du diable (*La Cousine Bette*).

Nothing can demonstrate more completely the strange capacity communicated by vice, to which we owe the strokes of skill which ambitious or voluptuous men can occasionally achieve or, in short, any of the Devil's pupils.

(11) – Si tu vas chez tes belles madames, je veux que tu effaces ce monstre de De Marsay, le petit Rastignac, les Ajuda-Pinto, les Maxime de Traille, les Vandenesse, enfin tous les élégants (*Les Illusions Perdues*).

If you are going to your fine ladies' houses, you shall eclipse that monster of a de Marsay and young Rastignac and any Ajuda-Pinto or Maxime de Trailles or Vandenesse, in short all of the dandies.

(12) – …je m'aperçus que la femme, autrefois si imposante par ses sublimités, avait dans l'attitude, dans la voix, dans les manières, dans les regards et les idées, la naïve ignorance d'un enfant, les grâces ingénues, l'avidité de mouvement, l'insouciance profonde de ce qui n'est pas son désir ou lui, enfin toutes les faiblesses qui recommandent l'enfant à la protection (*Le Lys dans la Vallé*e).

I now perceived that the woman, once so dignified in her bearing, showed in her attitude, her voice, her manners, in her looks and her ideas, the naive ignorance of a child, its artless graces, its eager movements, its careless indifference to everything that is not its own desire, in short all the weaknesses which commend a child to our protection.

The marker *enfin* expresses an enunciative stance, and therefore subjectivity, not so much at the level of the class in question as in the very process of categorization.

DF: Contrasting two categories

Two motifs fulfill this function:

– **il être un NC** (+ sub. relative) (there be a N) 32/371

The existential sentence with the verb *être* (*to be*) is more salient in Balzac than the existential sentence with the verb *avoir* (*to have*) when the discourse function is to contrast two categories (usually categories of people):

(13) Il est des femmes qui s'éprennent de la grandeur comme d'autres de la petitesse (*Les Illusions Perdues*)

Lit. There are women who are as much attracted by greatness as others by pettiness

(14) Il est des personnes que nous ensevelissons dans la terre, mais il en est de plus particulièrement chéries qui ont eu notre cœur pour linceul (*Le Lys dans la Vallée*)

Lit. There are people we bury in the ground, but there are some particularly cherished ones whose shroud is our heart.

The motif is of course also used in situations where there is no contrast but a strong assertion:

(15) Mon cher monsieur Crevel, répliqua la Lorraine, il est des noms qu'on ne prononce pas ici (*La Cousine Bette*)

My dear Monsieur Crevel, replied Lisbeth, there are certain names we never utter here.

– **plus ADJ que ne le être** (+ sujet) (more ADJ than is) 12/17

This segment is a comparative device expressing a relation of superiority:

(16) Le militaire, en temps de guerre, n'est-il pas également réservé à des spectacles encore plus cruels que ne le sont les nôtres ? (*La cousine Bette*).

Lit. Is not the soldier in time of war brought face to face with sights even more dreadful than those that we see?

(17) Je croyais à de pures amitiés, à des fraternités volontaires, plus certaines que ne le sont les fraternités imposées (*Le lys dans la vallée*).

I believed in pure friendship, in a voluntary brotherhood, more real than the brotherhood of blood.

DF: Characterizing by a remarkable property

– **un de ce NC ADJ** (one of this N ADJ) 55/257

The referent is said to belong to a class with a highly distinctive feature; this feature is supposedly known by the addressee or by the reader.

(18) Votre fille est une de ces beautés effrayantes pour les maris (*La cousine Bette*).

Your daughter is one of those beauties who rather alarm intending husbands

(19) La duchesse tourna sur Eugène un de ces regards impertinents qui enveloppent un homme des pieds à la tête (*Le Père Goriot)*

The Duchess gave Eugene one of those insolent glances that measure a man from head to foot.

Weinrich (1974) considers this stylem as specific to Balzac. It has been analyzed by Bordas (2001) who clearly showed its usage in the 19th century novel.

– **VPS le plus ADJ NC** (lit. $V_{\text{simple past}}$ the more ADJ N) 10/21

The motif marks the high degree of a property. In this structure, the adjective is anteposed instead of being in its usual postposed position:

(20) Je subis alors une conversation folle, pendant laquelle il me fit les plus ridicules confidences. (*Le Lys dans la Vallée*).

Lit. I then had to put up with a mad dialogue, during which he told me the most ridiculous secrets.

(21) Elle éprouva la plus vive émotion de sa vie, elle sentit pour la première fois la joie inondant son cœur (*La Cousine Bette*).

She felt the deepest emotion of her life; for the first time she felt the full tide of joy rising in her heart.

3.2 Hugo

DF: Amplification for expressive purposes

– **PASS , PASS , PASS , PASS ,** (+ etc.) 23/33; **ADJ , ADJ , ADJ , ADJ ,** (+ etc.) 167/402; **VP ,VP ,VP , VP ,** (+ etc.) 10/13

This set of patterns is highly characteristic of Hugo's style: the juxtaposition can reach spectacular lengths.

(22) C'est en faisant la bouche en cœur du côté de ma vieille caboche que tu as taillé, coupé, tourné, viré, traîné, limé, scié, charpenté, inventé, écrabouillé, et fait plus de miracles à toi tout seul que tous les saints du paradis (*Les Travailleurs de la Mer*)

It was in the midst of all this misery, alongside of my old craft, that you shaped, and cut, and turned, and twisted, and dragged about, and filed, and sawed, and carpentered, and schemed, and performed more miracles there by yourself than all the saints in paradise

(23) Gauvain avait devant lui l'impossible devenu réel, visible, palpable, inévitable, inexorable (*Quatre-vingt-treize*).

Lit. Gauvain had before him the impossible become real, visible, palpable, inevitable, inexorable

(24) L'émeute est une sorte de trombe de l'atmosphère sociale qui se forme brusquement dans de certaines conditions de température, et qui, dans son tournoiement, monte, court, tonne, arrache, rase, écrase, démolit, déracine, entraînant avec elle les grandes natures et les chétives, l'homme fort et l'esprit faible, le tronc d'arbre et le brin de paille (*Les Misérables*).

Revolt is a sort of waterspout in the social atmosphere which forms suddenly in certain conditions of temperature, and which, as it eddies about, mounts, descends, thunders, tears, razes, crushes, demolishes, uproots, bearing with it great natures and small, the strong man and the feeble mind, the tree trunk and the stalk of straw.

The sentence grows by adding synonymous expressions, which creates an effect of variation in the meaning (Wulf 2014).

DF: Expressing the existence of a property in a human patient

– **il y avoir dans le NC** (+ GN) (there be in the N) 87/158

(25) Il y a dans le désespoir de la femme on ne sait quoi de faible qui est terrible (*Quatre-vingt-treize).*

There is in a woman's despair some indescribable weakness which is terrible to behold.

(26) Il y avait dans la femme le fond d'une brute et dans l'homme l'étoffe d'un gueux (*Les Misérables*)

In the wife there was a brutish streak and in the husband, the makings of a ruffian.

– **il y avoir de le NC dans** (+GN) 59/81

The impersonal construction here "massifies" count nouns:

(27) Il y avait de la crinière dans sa perruque (*Quatre-vingt-treize*)

Lit. There was horsehair in his wig

(28) Il y a du songe dans le tonnerre (*Les travailleurs de la mer*)

Lit. There is dream in thunder

– **le NC avoir un NC, le NC** (the N have a N, the N) 17/22

The informational structure places the focus on the identification of the "possessed" object:

(29) La révolution a un ennemi, le vieux monde, et elle est sans pitié pour lui, de même que le chirurgien a un ennemi, la gangrène, et est sans pitié pour elle (*Quatre-vingt-treize*).

Revolution has an enemy, the old world, and it feels no pity for it, just as the surgeon has an enemy, gangrene, and feels no pity for it.

(30) Le Passé a un synonyme, l'Ignoré (*L'Homme qui Rit*)

Lit. The past has a synonym, the Ignored

DF: Metaphorizing

Hugo is a writer of metaphor, a trope which he uses in all his novels, and not only in those that are sometimes referred to as philosophical (*Les Travailleurs de la mer* and *L'Homme qui Rit*). There are countless occurrences of this motif in the corpus:

– **, ce être le NC** (+ expansion) (, it be the N) 516/1386

This form is often used to express an opposition or at least a differentiation:

(31) La guerre étrangère, c'est une écorchure qu'on a au coude ; la guerre civile, c'est l'ulcère qui vous mange le foie (*Quatre-vingt-treize*).

A foreign war is a scratch on the elbow; a civil war is an ulcer which eats away your liver

(32) Dédalus, c'est le soubassement ; Orpheus, c'est la muraille ; Hermès, c'est l'édifice (*Notre Dame de Paris*).

Lit. Daedalus is the foundations; Orpheus is the wall; Hermes is the building

– **le NC être un NC** (+ expansion) (the N be a N) 254/332 *;* **un NC être un NC** (+expansion)(a N be a N) 79/113

(33) La jeunesse est un plan incliné (*l'Homme qui Rit*)

Youth is an inclined plane

(34) Une prison est un habit de pierre (*Quatre-vingt-treize*)

Lit. A prison is a garment of stone

DF: Constructing contrasts

– **le NC VP, le NC VP** (the $NV_{simple\ present}$, the N $V_{simple\ present}$) 26/29 ; **le NC être ADJ, le NC être ADJ** (the N be ADJ, the N be ADJ) (35/41)

Contrast is created by simple juxtaposition, exploiting antonymic relations and lexical chiasmus:

(35) le savant conjecture, l'ignorant consent et tremble (*Les Travailleurs de la Mer*)

Lit. the scholar speculates, the ignoramus agrees and trembles

(36) le petit est grand, le grand est petit (*Les Misérables*)

Lit. the little one is tall, the tall one is little

(37) le ciel est noir, l'océan est blanc (*L'Homme qui Rit*)

Lit. The sky is black, the ocean is white

– **il y avoir le NC** (+ et/comme) **il y avoir le NC** (there be the N (and/like there be the NC)14/19

Here the existential construction is used as a device to contrast two realities:

(38) Il y a les fumées paisibles et il y a les fumées scélérates (*Quatre-vingt-treize*)

Lit. There is peaceful smoke and there is villainous smoke

(39) Il y a la bravoure du prêtre comme il y a la bravoure du colonel de dragons (*Les Misérables*)

Lit. There is the bravery of the priest and there is the bravery of the colonel of dragoons

(40) Dans la nuit il y a l'absolu ; il y a le multiple dans les ténèbres (*L'homme qui Rit*).

Lit. At night there is the absolute; there is multiplicity in darkness

3.3 Gaboriau

Gaboriau is the French father of the detective novel.

DF: Constructing reasoning

The first three motifs correspond to hypothetical constructions; two motifs have the canonical form of *si* (*if*) hypothetical sentences. The third uses a different device: a thematic infinitive followed by *ce être se INF.*

– **si** (+ P), **ce être que il** (+ P) (if (+P), it be that he) 72/247

(41) S'il se tait, c'est qu'il n'a rien trouvé de plausible (*Le Crime d'Orcival*)

Lit. If he remains silent, it is because he found nothing plausible

(42) Or, pensait Lecoq, s'il accepte cette lutte, c'est qu'il entrevoit quelque chance d'en sortir vainqueur (*Monsieur Lecoq*)

Lit. However, Lecoq thought, if he accepts this struggle, it is because he foresees some hope of emerging victorious from it

– **si** (+P), **ce en être faire de NP/ DETPOSS NC** (if (+P), it be the end of NP/ detposs N) 17/36

(43) Si Trémorel est jugé, c'en est fait de Laurence (*Le Crime d'Orcival*)

Lit. If Trémorel is judged, it is the end of Laurence

(44) Si vous faites du bruit, dit-il, si vous donnez l'éveil, c'en est fait de nos espérances (*Le Dossier 113*)

If you make any noise, he said, or raise an alarm, all our hopes are ruined.

– (Inf. + expansion), **ce être se INF** (It be to Inf) 10/41

(45) Rester, c'était s'exposer à une explication pénible, à des insultes, à une collision peut-être … (*L'Affaire Lerouge*)

Lit. To stay was to run the risk of a painful explanation, insults, maybe a confrontation

(46) Laisser voir ses opinions, c'était se créer sans nécessité ni utilité une situation impossible (*Le Dossier 113*)

Lit. Revealing his own opinions would only needlessly create an impossible situation

DF: Introducing direct speech

The following two patterns are specific to the introduction of direct speech:

– **, et après un NC** (+ expansion) **:** (, and after a N :) 9/16

(47) D'un coup d'œil, le vieux brocanteur avait embrassé ces détails, et après un sourire de remerciement à sa sœur : (*La Clique dorée*)

Lit. At a glance, the old bric-à-brac dealer had taken in these details, and after a smile of thanks to his sister:

(48) Il se redressa un peu interdit, et après un moment de méditation : (*Monsieur Lecoq*)

Lit. Surprised, he sat up, and after a moment of thought:

– **, et ce être de un NC ADJ que** (+ P) **:** (, and this be of a N ADJ that(+P) :) 8/8

(49) …, et c'est d'une voix étranglée qu'il murmura : (*La Clique Dorée*)

Lit. …and it was with a choked voice that he murmured

(50) …, et c'est d'une voix brutale qu'il répondit à sa mère : (*Le Dossier 113*).

Lit. …, and it was in a violent tone of voice that he answered his mother:

3.4 Flaubert

DF: Expressing the invasion of feelings

– **un NC ADJ le VIMP** (a N ADJ him Vpast) (23/76)

(51) Une angoisse permanente l'étouffait (*L'Education Sentimentale*)

Lit. A constant anxiety stifled him

(52) Une tranquillité singulière l'occupait (*Salammbô*)

Lit. He was filled with a singular tranquility

(53) Des réflexions douloureuses l'assaillaient (*Bouvard et Pécuchet*)

Lit. Painful thoughts assailed him

DF: Making an experience known thanks to an approximation

– (transitive verb) **comme le NC de un NC** (like the N of a N) (58/242)

This hedging construction mitigates the representation of an experience:

(54) Les enfants qui chantaient des hymnes, les gerbes de lilas, les festons de verdure, lui avaient donné comme le sentiment d'une jeunesse impérissable (*Bouvard et Pécuchet*)

Lit. The children singing hymns, the wreaths of lilac, the festoons of greenery, gave him a sensation like that of imperishable youth

(55) Frédéric éprouva comme la sensation d'un coup de fouet (*L'Education Sentimentale*)

Lit. Frederick felt a sensation like that of a whiplash

(56) j'ai senti dans mon cœur comme le froid d'une épée (*Salammbô*)

Lit. I felt in my heart the coldness of a sword

The motifs briefly presented in this section illustrate the relevance of the method: the motifs emerged from an automated and unsupervised analysis. Only a few of those listed above have been previously identified by stylisticians. Others require comments and developments.

4 Discussion

The motif is, we contend, a necessary and fundamental unit to link qualitative and quantitative stylistics. The examples introduced above have already shown that motifs are associated with expressive aims. It is however necessary to go

further in order to highlight the convergence of some motifs towards the author's aesthetic project or world-view. We will take the example of Victor Hugo, in whose work we have identified a "family" of motifs that all in various ways express the same theme: the need to adjust categories and denominations to objects. A characteristic feature of Hugo's aesthetics is the difficulty of describing referents or experiences using pre-established categories. Hence the need to adjust categorization, designation, or even perception.

A. The categorization of the object or experience is based on an approximate category. Several motifs express this type of categorization:

– **quelque chose comme le NC** (something like the N) 24/28 **/ on ne savoir quoi de ADJ** (one does not know what ADJ) 24/24 **/ on ne savoir quel NC** (one does not know what NC) 99/104

(57) C'était quelque chose comme le lever d'une montagne d'ombre entre la terre et le ciel (*Les Travailleurs de la Mer*).

Lit. It was something like the rising of a mountain of shadow between earth and heaven

(58) Elle se souvenait d'on ne sait quoi de lumineux et de chaud (*L'Homme qui Rit*)

Lit. She remembered something indescribably bright and hot

(59) Leur souffle sous le voile ressemble à on ne sait quelle tragique respiration de la mort (*Les Misérables*)

Lit. Their breath under the veil resembles some indescribably tragic breathing of death

– **Avec un sorte de NABS** (with a sort of $N_{abstract}$) 23/78

(60) Cosette considérait la poupée merveilleuse avec une sorte de terreur (*Les Misérables*)

Lit. Cosette looked at the marvelous doll with a sort of terror

(61) Il lui touchait les cheveux avec une sorte de précaution religieuse (*Les travailleurs de la mer*)

He touched her hair with a sort of religious awe

– **être un espèce de NC** (be a kind of N) 25/46 / **être un sorte de NC** (be a sort of N) 39/83

(62) Ce qui fait qu'une mère est sublime, c'est que c'est une espèce de bête (*Quatre-vingt-treize*)

Lit. what makes a mother sublime is that she is a kind of beast

(63) C'était une sorte de ruche monstrueuse qui y bourdonnait nuit et jour (*Notre Dame de Paris*)

It was a sort of monstrous hive, which buzzed night and day.

B. The categorization is corrected

– **être plus que un NC (ADJ), ce être un NC** (be more than a N (ADJ), it be a N) 7/10

(64) Tellmarch était plus qu'un homme isolé, c'était un homme évité (*Quatre-vingt-treize*)

Lit. Tellmarch was more than an isolated man, he was avoided

(65) C'est plus qu'une morte, c'est une sainte (*Les Misérables*)

Lit. She is more than a dead person, she is a saint

C. The denomination is adjusted

– **ce qu'on pouvoir INF** (what one can + Inf) 17/52

(66) Notre-Dame de Paris n'est point du reste ce qu'on peut appeler un monument complet, défini, classé.

Lit. Notre-Dame of Paris is not what one can call a complete, defined, classified monument

– **ce NC que on VP** (this N that we $V_{\text{simple present}}$) 27/56

(67) C'est cette décadence qu'on appelle renaissance. (*Notre Dame de Paris*)

Lit. It is this decadence which is called revival

(68) Il y a un dieu pour ces ivrognes qu'on appelle les amoureux (*Les Misérables*)

Lit. There is a god for the drunkards who are called lovers

In these examples, the referent is first classified in a detrimental category (decadence, drunkard), and then it is positively recategorized.

D. The perception is adjusted

This indirect and approximate characterization of an object or situation is also present in a seemingly different motif (which is mostly used in *Quatre-vingt-treize*):

– **qui être le NC de (le NC/NP)** (that be the N of) 77/140

This relative clause is very particular: it re-categorizes with a definite noun phrase a previous indeterminate phrase that corresponds to the initial perception of a whole:

(69) Un rassemblement se pressait devant un perron de quelques marches qui était l'entrée de la mairie (*Quatre-vingt-treize*)

Lit. A crowd gathered in front of a flight of a few steps which was the entrance to the town hall

(70) À l'est apparaissait une blancheur qui était le lever du jour, à l'ouest blêmissait une autre blancheur qui était le coucher de la lune (*Quatre-vingt-treize*)

Lit. In the East a whiteness appeared which was the sunrise, and in the West another whiteness paled which was the setting of the moon

(71) Derrière cette tour se perdait dans la brume une grande verdure diffuse qui était la forêt de Fougères (*Quatre-vingt-treize*).

Lit. Behind this tower a broad expanse of greenery which was the forest of Fougères was lost in the mist

A flight of a few steps, a whiteness, a broad expanse of greenery are parts (or, better, signs) of a whole that cannot be directly identified.

Like other patterns, but in a very different way, this motif expresses the hugolian theme of the difficult relationship between the two modalities of perception and knowledge.

These examples show how a stylistics of motifs can contribute not only to identifying authors, but also to characterizing their writing.

5 Conclusion

To sum up, motifs are units chosen from segments of annotated units. They are both quantitative and qualitative in nature, in that they allow computational stylistics not to be confined to the selection of "morellian" features.

If the method is to make a real contribution to the knowledge of styles, it is advisable to list the cases where it could provide the specialist of literary texts with a genuine device for observation and analysis. We mention just briefly two possible directions for future research: How can motifs be made more informative? What types of analysis does the method allow?

It is obvious that the annotation could be enriched, notably by semantic properties. Speech and perception verbs are already considered in the method, but it would be possible to take other categories into account as well, such as psychological verbs (*émouvoir* (to move) *effrayer* (to scare), etc.). Similarly, action nouns, deverbal or even event nouns could be specifically annotated, provided reliable lists were available. This improvement would no doubt make it possible to identify more semantic specificities than those illustrated above.

Moreover, the method still lacks indications about the topological location of motifs in the text: do some motives tend to appear at the beginning / end of the novel, chapter, paragraph, or sentence? This is a matter of "textual colligation" (Hoey 2005), which is still technically difficult to implement, but necessary to study because it can be assumed that a relationship exists between certain forms and the narrative and textual phases of novels.

Finally, correlations between motifs should also be calculated: is a pattern preferentially used with some other motif in the same context (chapter, paragraph)? This calculation would make it possible to identify not only single motifs, but sets of motifs, giving rise, perhaps, to significant constellations.

The application of the method is a crucial question, as it is essentially the applications that can contribute to its validation.

References

Bertels, A. & D. Speelman. 2013. 'Keywords Method' versus 'Calcul des Spécificités'. A comparison of tools and methods. *International Journal of Corpus Linguistics* 18:4, 536–560.

Bordas, Eric. 2001. Un stylème dix-neuviémiste. Le déterminant discontinu un de ces… qui… *L'Information Grammaticale* 90 :1, 32–43.

Brunet, Etienne. 1985. *Le Vocabulaire de Zola*, suivi de l'Index complet et synoptique des Rougon-Macquart. Genève-Paris: Slatkine-Champion. 3 volumes.

Burrows, John. 1987. *Computation into Criticism*: *A Study of Jane Austen's Novels*. Oxford: Clarendon Press.

Eder, Macie M., Mike Kestemont & Jan Rybicki. 2016. Stylometry with R: A package for computational text analysis. *R Journal* 16(1). Advance access available at: https://journal.r-project.org/archive/

Fletcher, William. 2003. Exploring words and phrases from the British national corpus. http://kwicfinder.com/BNC/

Ganascia, Gabriel. 2001. Extraction automatique de motifs syntaxiques. *Actes de Traitement Automatique du Langage Naturel* 2001 (TALN 2001). Tours, 2–5 July.
Herrmann, Berenike J., Karina van Dalen-Oskam & Christof Schöch. 2015. Revisiting Style, a Key Concept in Literary Studies. *Journal of literary theory* 9/1, 25–5.
Ho, Yufang. 2011. *Corpus Stylistics in Principles and Practice. A Stylistic Exploration of John Fowles' The Magus*. London/New York: Continuum.
Hoey, Michael. 2005. *Lexical Priming: A new theory of words and language*. London: Routledge.
Holmes, David I. 1998. The Evolution of Stylometry in Humanities scholarship. *Literary and Linguistic Computing* 13: 3, 111–117.
Hoover, D. L. 2007. Corpus Stylistics, Stylometry, and the Styles of Henry James. *Style* 41, 174–203.
Kastberg Sjöblom, Margareta. 2006. *L'écriture de J.M.G. Le Clézio. Des mots aux thèmes*. Paris: Honoré Champion.
Kilgarriff, Alan. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1-2, 263–276.
Köhler, Reinhard. 2015. Linguistic motifs. In George K. Mikros & Ján Macutek (eds.), *Sequences in Language and Text*, 89–108. Berlin–Boston: De Gruyter Mouton.
Lafon, Pierre. 1984. *Dépouillements et statistiques en lexicométrie*. Genève–Paris: Slatkine-Champion.
Loiseau, Sylvain. 2010. Les paradoxes de la fréquence. *Energeia* 2, 20–55.
Longrée, Dominique & Sylvie Mellet. 2013. Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours. *Langages* 189, 68–80.
Lutosławski, Wincenty. 1897a. *The Origin and Growth of Plato's Logic*. London: Longmans.
Lutosławski, Wincenty. 1897b. On stylometry. *Classical Review* 11, 284–286.
Magri-Mourgues, Véronique. 2009. *Le Voyage à pas comptés. Pour une poétique du récit de voyage au XIX*e *siècle.* Paris: Honoré Champion.
Mahlberg, Michaela. 2013. *Corpus Stylistics and Dickens's Fiction.* New York: Routledge.
Oakes, Michael. 2014. *Literary Detective Work on the Computer*. Amsterdam–Philadelphia: Benjamins.
Pawłowski, Adam & Artur Pacewicz. 2004. Wincenty Lutosławski (1863–1954). Philosophe, helléniste ou fondateur sous-estimé de la stylométrie? *Historiographia Linguistica* 31: 2–3, 423–447.
Quiniou, Solen, Peggy Cellier, Thierry Charnois & Dominique Legallois. 2012. What about sequential data mining techniques to identify linguistic patterns for stylistics? *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent text Processing*, Mar. 11–17, CICLing, New Delhi, India. 166–177.
Rastier François. 2001. Vers une linguistique des styles. *L'Information Grammaticale* 89, 3–6.
Römer, Ute. 2010. Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction* 3, 95–119.
Salem, André. 1987. *Pratique des segments répétés. Essai de statistique textuelle*. Publication de l'INALF, collection « Saint-Cloud ». Paris : Klincksieck.
Van Peer, Willie. 1989. Quantitative studies of literature: a critique and an outlook. *Computers and the Humanities* 23 (4/5), 301–307.
Weinrich, Harald. 1974. *Le Temps*. Paris: Le Seuil.
Wulf, Judith. 2014. *Étude sur la langue romanesque de Victor Hugo. Le Partage et la composition.* Paris: Classiques Garnier.

Categorial symbols

ADJ	adjective
DETPOSS	possessive determiner
INF	infinitive
NC	common noun (N)
NCABS	abstract noun
NP	proper name
VPS	simple past verb
VP	present verb
PASS	past participle

Sandra Augendre, Anna Kupść, Gilles Boyé
and Catherine Mathon

Live TV sports commentaries: specific syntactic structures and general constraints

Abstract: The aim of this chapter is to examine the linguistic specificities of live TV sports commentary as a discourse genre. We first analyze syntactic structures from the commentary of a rugby match. In order to provide a more fine-grained analysis of the data, we adopt the division of the discourse into "play-by-play" (simultaneous narration of game actions) and "color-commentary" ("non-activity-tied freely spoken part"). Results show that the distribution of structures is not identical in the two cases. They we compare these syntactic structures with those extracted from the commentaries of a soccer match and a 4x100 meter relay race. This comparison enables us to reveal some interesting differences among these three sports, but the analysis shows that, in order to get a better grip on the nature of linguistic invariants in live TV sports commentary as a discourse genre, exploring the oral dimension of live sports commentary (especially prosodic features such as rhythm, pitch and intensity) is necessary.

The conclusion is that both syntactic analysis and recognition tests show that linguistic structures are not sufficient to define sports commentary as a discourse genre, even when linguistic productions are linked to contextual information. Because of the visual dimension of live TV sports commentary, the rhythm of the sport and the sequence of actions must be taken into account as constraints which influence the structures in the discourse.

1 Introduction

The aim of this study on French live TV sports commentaries is to examine the behaviour of the linguistic units from a functional point of view, in relation to communicative purpose and production contexts.

First of all, we consider live TV sports commentary as a discourse genre and its linguistic specificities as one of the distinctive features of this genre.

Sandra Augendre, CLLE-ERSSàB (UMR 5263)
Anna Kupść, Université Bordeaux Montaigne, CLLE-ERSSàB (UMR 5263)
Gilles Boyé, Université Bordeaux Montaigne, CLLE-ERSSàB (UMR 5263)
Catherine Mathon, Université Bordeaux Montaigne, CLLE-ERSSàB (UMR 5263)

https://doi.org/10.1515/9783110595864-009

A discourse genre has been addressed by several authors: Bakhtine (1979: 269), Canvat (1999: 115), Petitjean (1991: 352), Maingueneau (1998: 51) and Adam (2001: 41) among others. Following their ideas, Richer (2005, 2011) proposes to summarize genre definitional criteria along six dimensions:

- A material dimension which includes the medium's nature (written, spoken, reading...), its size (e.g. length of an abstract), its duration (5-minute weather forecast or 80-minute rugby match), its periodicity (daily column, TV series).
- A socio-pragmatic dimension which refers to questions such as "who is speaking?","To whom?", "For what purpose?", "In which context?" etc.
- A textual dimension which concerns the general text outline and identification of regular organisation patterns.
- A stylistic dimension which includes register (formal/casual...), lexical choices, syntactic structures. This dimension is of most interest for this study since we discuss various syntactic structures trying to define a discourse genre.
- A thematic dimension, i.e. the thematic content specific to a genre, the choice of specific themes (sports moves, tactics and strategies for sports commentary).
- A cultural dimension which includes all the other dimensions.

Previous studies have favored both stylistic and socio-pragmatic approaches by analyzing syntactic structures specific to sports commentary (Deulofeu 2000; Krazem 2011) or by relating prosodic features to contextual traits (Audrit *et al.* 2012; Pršir *et al.* 2013). Most of these studies, even if they acknowledge the influence of both the nature and rhythm of the sport on the discourse, focus on the discursive productions only, and don't take into account the constraints of the sport.

Our study, presented in this contribution, is based on a detailed analysis of the opening France-Argentine match of the Rugby World Cup in 2007. The match broadcasted on French TV (TF1) was commented by three speakers: Thierry Gilardi, journalist; Thierry Lacroix, sports expert; Fabrice Landreau, touchline reporter. The overall duration of the match is 108 minutes but the speech time corresponds to 55 minutes only, distributed as follows: journalist 40 minutes; sports expert 13 minutes; touchline reporter 2 minutes. The match was orthographically transcribed and the text was aligned with spoken commentary, see Lortal & Mathon (2008) for initial annotations.

More recently, following Hartmann (2013), we distinguished two kinds of discourse in our corpus: "play by play" (simultaneous narration of game moves) and "color commentary" ("non-activity-tied freely spoken part", Hartmann 2013:12). We also indicated syntactic structures in the corpus. Finally, the text was aligned with video images and all game moves were annotated, see Mathon *et al.* (2015). The main goal of these annotations is to provide rich data which will allow us to investigate interactions between linguistic and extralinguistic factors on discourse production.

Our study of the genre explores two types of non-discrete linguistic units: syntactic structures and prosodic units. The main focus of the contribution is on syntactic units: first, we provide an inventory of syntactic patterns present in this type of discourse (genre) and second, we analyze the influence of the genre itself on the use/choice of these patterns. We provide only sketchy observations concerning prosodic patterns. Nevertheless, these spoken production units obviously play a part in the definition of the oral genre discussed in the study, i.e., live TV sports commentary.

The organisation of our contribution is as follows. First (Section 2), we analyze syntactic structures from the French commentary of the rugby match described above. Then (Section 3), we compare these syntactic structures with those extracted from the French commentaries of a 4x100 meter relay race, a soccer match and a basketball match summary. This comparison enables us to reveal some interesting differences between these sports with respect to production context (live commentary vs. a match summary). Finally (Section 4), in order to assess the importance of syntactic structures in the definition of this discourse genre, we explore the oral dimension of live sports commentary. For this purpose, we provide results of two recognition tests based on prosodic information only.

2 Specific structures in sports commentary

Several authors (e.g. Deulofeu 2000; Krazem 2011) studying sports commentary have characterized this genre, from a syntactic point of view, by a number of specific structures. In particular, the following structures have been mentioned: isolated proper names, simple and complex nominal and prepositional phrases, i.e., structures where the verb is either absent or is not the head element of the group.

In our rugby corpus these structures are present as well; their distribution is presented in Table 1.

This distribution corresponds to the data (speech turns) in the entire corpus. Even though verbal turns (SS: simple sentence, CS: complex sentence) are still quite frequent, one third of all turns corresponds to specific verbless turns:

– Isolated proper names (X, 8%), or proper names followed by a relative clause (*Xqui* 'Xwho', 2%)

Table 1: Distribution of syntactic structures in the entire corpus.

Structure	SS	CS	NP	X	N	PrepX	PP	Xqui	Inf	Part	OTHERS	Total
Total	841	277	179	142	118	79	52	39	30	25	80	1862
Percentage	45%	25%	10%	8%	6%	4%	3%	2%	2%	1%	4%	100%

- Prepositions followed by proper names (PrepX, 4%)
- Noun phrases with (NP, 10%) or without (N, 6%) an article
- Prepositional phrases (PP, 3%)
- Participle (Part, 1%) and infinitive clauses (Inf, 2%).

Their presence can be explained by several factors. First, each structure possesses its own informative contribution: proper names (X) indicate the ball possessor, prepositions followed by a proper name (PrepX) describe ball movements from/to a player, nominal phrases (NP and N) introduce a move, whereas move localisation and progression is expressed by prepositional phrases (PP). Then, each structure provides an answer to various extralinguistic constraints that influence the discourse. First of all, we consider the impact the two discourse types linked to game development, "play by play" (simultaneous narration of game moves) and "color commentary" ("non-activity-tied freely spoken part", Hartmann, 2013:12), may have on linguistic (syntactic) production. Then, we analyze the structures in relation to other parameters such as speakers, rhythm of the game (number of moves within a 5-second time window) and type of move to be described (long/short, decisive/not decisive, etc.).

2.1 Syntactic structures according to "play by play" and "color commentary" distinction

When we provide a more fine-grained analysis of the data, adopting the division of the discourse into "play by play" (40% of the entire corpus) and "color commentary" (60% of the entire corpus), the repartition of the structures present in our corpus is contrasted. Although the distribution of structures in "color commentary" is similar to that in the entire corpus, their frequency in "play by play" commentary is strikingly different: the overall frequency of specific structures is almost four times higher and, at the same time, some structures (ex. PrepX) appear only in this part.

This comparison shows that only the discourse directly affected by the game is syntactically marked. Indeed, it seems that extralinguistic constraints (time and rhythm of the game) force the commentator to adopt a non-standard format in order to follow the game more efficiently. For example, prepositions with proper names (PrepX) allow commentators to succinctly identify ball movements and possessors. If game development allows, every structure can be further elaborated. For example, in addition to its basic format, PrepX is found in our corpus followed by a relative clause, a prepositional or participial phrase which provides more details on the current move.

In the following sections we will analyze our corpus data in more detail, focusing on the three structures considered in the literature (Deulofeu 2000,

Krazem 2011) as characteristic of sport commentary: structures involving proper names, structures based on nouns and structures based on prepositions.

2.1.1 Structures involving proper names

As shown in Table 2, proper names (X) and related structures (PrepX and *Xqui* 'Xwho') are most frequent in "play by play" and correspond, respectively, to 90%, 100% and 80% of all occurrences of these phrases. Indeed, we find them in quick sequences of moves, for which the commentary is limited to identification of the agent. Typically, the ball possessor is indicated by X alone or preceded by the preposition *avec* ('with X') while its receiver is indicated by *pour X* ('for X'). *Avec* and *pour* appear to be the two main prepositions used in PrepX structure but their distribution is not homogenous: *avec* 'with' shows up in 75% of PrepX whereas *pour* 'for' only in 20% of PrepX.

As Tables 3 and 4 show, for more than 60% of the occurrences, the speaker is limited to the identification of player(s) implied in the current move(s). However, if the move development allows, the player's name can be completed by other

Table 2: Distribution of syntactic structures according to play by play/color commentary distinction.

Structure	SS	CS	PP	N	X	PrepX	Xqui	PART	INF	NP	OTHERS	Total
Play by Play	215	37	35	77	125	78	29	21	14	95	12	738
	29%	5%	5%	10%	17%	11%	4%	3%	2%	13%	2%	100%
Color	626	240	17	41	17	1	10	4	16	84	68	1124
Commentary	56%	21%	2%	4%	2%	0,1%	0,9%	0%	1%	7%	6%	100%
Entire corpus	841	277	52	118	142	79	39	25	30	179	80	1862
	45%	15%	3%	6%	8%	4%	2%	1%	2%	10%	4%	100%

Table 3: Proper Name (X) and its extensions.

Structure	Examples	Distribution
X	a. *David Skrela* 'David Skrela'	111 (61%)
Xqui	b. *Mignoni encore qui insiste au ras* 'Mignoni snipes down the near side again'	39 (22%)
XPP	c. *Contepomi pour la transformation* 'Contepomi for the conversion'	16 (9%)
XPrepX	d. *Pichot pour Hernandez* 'Pichot to Hernandez'	9 (5%)
XPart	e. *Heymans bien pris par trois argentins* 'Heymans well closed down by three Argentinians'	6 (3%)

Table 4: PrepX and its extensions.

Structure	Examples	Distribution
PrepX	a. *Avec Dominici / Pour Skrela* 'With Dominici' / 'To Skrela'	54 (68%)
PrepXqui	b. *Avec Heymans qui a les appuis* 'With Heymans who is side-stepping'	21 (26%)
PrepXPP	c. *Avec Harinordoquy en bout de ligne* 'With Harinordoquy at the back of the lineout'	3 (4%)
PrepXPart	d. *Avec Martin lancé par Mignoni* 'With Martin launched by Mignoni'	2 (3%)

information. An extension with a subject relative clause *qui* ('who') seems to be the most frequent strategy adopted by commentators (see examples Tab. 3 ex. b and Tab. 4 ex. b).

2.1.2 Structures based on nouns

Concerning the structures based on nouns, we distinguish two types: nominal phrases with an article (NP, 10% of all structures) and, less frequent, nominal phrases without an article (N, 6% of all structures).

The nominal phrase is another short format privileged in play by play (65% of all NPs occur in this type of discourse). An NP is used to name a move (Tab. 5 ex. a), associated with an agent, by adding a PP as in (Tab. 5 ex. b), or to give details on its development (using a relative clause or a past participle, see (Tab. 5 ex. c and d) .

The reduced variant of NP, based on a noun without an article (N), is even more specific. This structure mostly appears in play by play commentary (71% of occurrences) and shares the informational goal with full NPs as they both focus on moves, see (Tab. 6 ex. e and f).

Table 5: NPs and their variants.

Structure	Examples	Distribution
NP	a. *Un bon lancer // Une bonne prise de balle* 'A good throw // Safe catch'	65 (36%)
NP PP	b. *Le dégagement de Dominici* 'Dominici's clearance'	57 (32%)
NPqui	c. *La bonne chandelle en arrière qui peut profiter à Michalak* 'Nice up-and-under, going backwards, which may work for Michalak'	37 (21%)
NP Part	d. *Un ballon bien pris quoique relâché* 'A ball first caught then spilled'	21 (12%)

Table 6: N and its variants.

Structure	Examples	Distribution
N	a. *Pénalité* 'Penalty'	35 (29%)
N PP	b. *Faute des français* 'French infringement'	29 (24%)
NPrepX	c. *Récupération de Raphael Ibanez* Lit.: Retrieval by Raphael Ibanez 'Raphael Ibanez retrieves the ball'	24 (20%)
NPart	e. *Ballon récupéré par Borges* Lit.: Ball taken by Borges 'It's a turn over by Borges'	18 (15%)
N X	d. *Remise en jeu Ibanez* 'Line-out throw by Ibanez'	10 (8%)
Nqui	f. *Tentative qui a échoué* Lit.: Attempt which failed 'An Attempt which failed'	4 (3%)

Table 7: PP and its variants.

Structure	Examples	Distribution
PP	a. *Au point de chute // dans l'axe peut-être* Lit.: at the landing point // Along the axis maybe 'Where the ball landed' // 'Down the middle maybe'	36 (67%)
PPprepX	b. *Avec le départ de Roncero* Lit.: With the start of Roncero 'With Roncero's charge'	9 (17%)
PP PP	c. *Dans l'axe pour le coup de pied d'Hernandez* Lit. : Along the axis for this kick by Hernandez 'Down the middle for this kick by Hernandez'	4 (7%)
PPPart	e. *Avec un arrêt de volée accordé par monsieur Spreadbury* 'With a mark awarded by mister Spreadbury'	3 (6%)
PPqui	d. *Dans ce jeu de pilonnage qui va peut-être libérer des espaces* 'In this aerial bombardment which may create some chances'	2 (4%)

2.1.3 Structures based on prepositions

The last structure, a general PP, also focuses on moves, especially its localization and development. The three main prepositions are *à* 'to' (33%), *dans* 'in' (15%) and *avec* 'with' (29%). The first two mostly specify move localization (77% of PP[*à*] and 75% of PP[*dans*] are locative), cf. Tab. 7 ex. a and d. The remaining dominant preposition, *avec* 'with', is linked to game description (Tab. 7 ex. b–c and e).

2.2 Syntactic structures of various speakers

As mentioned above, our rugby match is commented by 3 speakers: a sports journalist and a rugby expert, both present in the broadcast studio, and a touchline reporter. The match commentary is essentially produced by the first two speakers (their discourse represents respectively 39% and 24% of the entire recorded time) whereas the touchline commentator intervenes much less.

Not only does the intervention time of each commentator vary, but so do the linguistic resources they employ, as indicated by the use of syntactic structures in our corpus. Table 8 below shows the distribution of six main structures present in our data, according to each speaker. Four of these structures have been discussed in previous sections as syntactically marked (a preposition followed by a proper name, a proper name alone, a proper name with a relative clause and a nominal phrase without an article) whereas the other two, although the most frequent, correspond to standard (simple and complex) sentences.

As far as the four syntactically marked structures are concerned, they are mainly or almost exclusively used by the sports journalist. In particular, the proper name alone or preceded by a preposition never appears in the discourse of the touchline reporter. On the other hand, the proportion of unmarked structures, i.e., French standard sentences, is relatively balanced in the discourse of the two main speakers, even if the ratio of simple to complex sentences is higher for the sports journalist whereas the opposite is true for the expert.

If we consider absolute frequencies, the dominance of verbal turns is clear for all speakers whereas marked structures seem to remain marginal. It should be noted however that the ratio of unmarked to marked structures present in the discourse of the expert is much lower than their proportion in the journalist's. This asymmetry can be explained by distinct production constraints of both speakers, see (Augendre *et al.* 2014). The expert mostly provides a color commentary, i.e., gives additional information concerning players, game strategies, rules etc., when the game is stopped or during long moves. On the other hand, the game

Table 8: Distribution of most frequent syntactic structures according to speaker.

	Journalist	Expert	Touchline	Total
PrepNP	73 (24%)	3 (1%)	0 (0%)	308 (100%)
NP	92 (30%)	9 (3%)	0 (0%)	
NPqui	32 (10%)	2 (1%)	2 (1%)	
N	75 (24%)	17 (6%)	3 (1%)	
PS	333 (38%)	280 (32%)	23 (3%)	880 (100%)
PC	80 (9%)	143 (16%)	21 (2%)	

in real time is almost exclusively commented by the journalist (even if he participates in color commentary as well) when he is forced to closely follow game and time constraints. As discussed in (Augendre *et al.* 2014), color commentary includes longer turns (up to 17–30 words, depending on the speaker), and therefore using "classical" (simple or complex) sentences. On the other hand, play by play comments are shorter (at most 13–19 words) and the journalist has a general tendency to use more verbless turns (up to 40% of his turns), which explains a higher rate of marked structures in his discourse.

Among unmarked structures, we want to mention one structure in particular, namely right dislocation, see (1). Interestingly, this structure is more often used by the journalist (24 times vs. only 7 times by the expert):

(1) *Et il joue au pied David Skrela*
Lit.: And he plays with foot David Skrela
'And David Skrela kicks the ball'

This kind of verbal turn presents a good informational strategy for the journalist: it permits to describe the move and the player, if move constraints make it possible (the game slows down, the player becomes recognizable, etc.). As the sentence is already complete, more detailed information (the player's name) can be provided by the dislocated element but is not required from a syntactic point of view (for more details, see Augendre *et al.* 2014).

2.3 Syntactic structures and game rhythm

In our previous study (see Mathon *et al.* 2015), we defined three game rhythms according to the number of moves realized in a given time window. In particular, we showed how the main syntactic structures are distributed in these three game rhythms (see Mathon *et al.* 2015: 24, Figure 8): when the rhythm is slow (rhythm 1), the discourse is dominated by (simple or complex) sentences; however, when the game rhythm accelerates, short and verbless turns (proper names, PPs, NPs, etc.) are much more frequent.

To go farther, we explore here how the distribution of syntactic structures varies not only with respect to game rhythm but also their complexity. Therefore, we analyze the distribution of a simple structure (for example, a proper name, X) in the three types of rhythm and compare it with the distribution of its different syntactic extensions, varying in complexity (ex., X completed by a relative clause, an adjective, a participle or PP).

For this study we have chosen two (marked) structures based on proper names, X with extensions (cf. Tab. 3 and Tab. 9 ex. a–c) and PrepX with extensions

Table 9: Distribution of selected marked structures (X, PrepX and RD) according to the game rhythm.

Structure	Examples	Rhythm 1	Rhythm 2	Rhythm 3
X	a. *David Skrela*	46	41	14
	'David Skrela'	45,5%	40,6%	13,9%
Xqui	b. *Michalak qui va être un peu seul*	27	8	1
	'Michalak who is going to be a little isolated'	75%	22%	3%
Other X structures	c. *Contepomi toujours dans les airs*	17	9	2
(+ADJ/PP…)	'Contepomi with another kick'	61%	32%	7%
PrepX	d. *Avec Heymans*	18	23	14
	'With Heymans'	33%	42%	25%
PrepXqui	e. *Avec Michalak qui a les appuis*	9	8	0
	'With Michalak who is side stepping'	53%	47%	0%
Other PrepX	f. *Avec Traille au contact*	2	1	0
structures (+inf/PP)	'With Traille into contact'	67%	33%	0%
Right dislocation	g. *Il revient agressif Fabien*	35	9	1
(RD)	Lit.: He comes back aggressive, Fabien	78%	20%	2%
	'An aggressive response from Fabien'			

(cf. Tab. 4 and Tab. 9 ex. d–f) and one particular type of unmarked structure, right dislocation (see ex. 1 and Tab. 9 ex. g). The results are presented in Table 9.

The results above indicate a clear correlation between game rhythm and the syntactic complexity of each structure. In general, the faster the rhythm, the fewer structures are used. Among various candidates considered here, only two base structures (X and PrepX, without extensions) adapt efficiently to commenting quick sequences of moves (respectively, 14% and 25% of these structures appear in rhythm 3). When sequences of moves are slower (rhythms 1 and 2), all types of structures are applied, including extensions, thus meeting one of the constraints of sports commentary which is to avoid silence and provide as much information as possible, cf. (Mathon *et al.* 2015). In particular, extensions (e.g., relative clauses) or structures (right dislocation) which include a verb can be used when there are less moves to comment: the majority of relative clause extensions of X and PrepX as well as RD structures appear in rhythm 1 (respectively, 75%, 53%, and 78%). In contrast, these structures are practically absent from comments in fast game rhythm (rhythm 3).

Another factor we took into account was the length (in terms of turns) of each structure with respect to game rhythm. In our corpus, syntactic structures often exceed a single turn and are spread over several successive turns. In order to verify whether game rhythm has an impact on structure length, we calculated the distribution of structures spread over several turns (indicated by a 'bis' suffix, meaning

a continuation of the same structure) in different rhythm types. The details for the three syntactic phrases discussed above as well as the overall number of all extended structures found in our corpus are presented in Table 10 below.

The numbers clearly show that extended structures are excluded from comments during the fast rhythm: almost no continuation is possible in rhythm 3 (only 0,7% of all extended structures occur in rhythm 3). The vast majority of extended structures concern rhythm 1, especially in the case of structures involving a conjugated verb (relative clause extensions or right dislocation). This observation confirms our hypothesis that speakers adapt their linguistic production to game constraints.

Table 10: Distribution of extended structures according to game rhythm.

Structure	Examples	Rhythm 1	Rhythm 2	Rhythm 3
XBis	No occurrence	0	0	0
XquiBis	a. *Hernandez à nouveau // qui lui va trouver la touche* Lit.: Hernandez again // who is going to find the sideline 'Hernandez again //he is going to kick it into touch'	24 83%	5 17%	0 0%
XotherBis	b. *Contepomi pour la transformation // de cet essai qui vaut cinq points* Lit.: Contepomi for the transformation // of this essay that is worth five points 'Contepomi for the conversion // to add to the five points from the try'	10 100%	0 0%	0 0%
PrepXBis	c. *Avec euh // Ostiglia et Fernandez Lobe* 'With er // Ostiglia and Fernandez Lobe'	1 33%	2 67%	0 0%
PrepXquiBis	d. *Avec Dominici // ah qui a pas trouvé le petit trou de la serrure* 'With Dominici // ah who did not manage to wriggle his way through'	14 64%	8 36%	0 0%
PrepXotherBis	No occurrence	0	0	0
RDBis	e. *Il faut qu'il ait du soutien vite // Christophe Dominici* 'He needs to have support soon // Christophe Dominici'	20 80%	5 20%	0 0%
Total for the three extended structures detailed here (X, PrepX, RD)		79 79,8%	20 20,2%	0
Total of all extended structures		793 89,6%	86 9,7%	6 0,7%

2.4 Syntactic structures and moves

In this section we discuss the relationship between individual game moves (rather than their rhythm) and syntactic structures involved in commenting them. Game annotations, described in (Mathon *et al.* 2015, Section 5.3: 18–19), allow us to view comments associated with separate moves and, in particular, to verify which moves are commented and, if this is the case, the type of syntactic structure applied.

As Table 11 below indicates, 63,6% of all moves in the entire match remain uncommented. Moreover, certain moves (e.g., a collapse) are never commented whereas others, e.g., a try, always are. We look for an explanation of such an unbalance directly in the conditions surrounding the commentary: on the one hand, the (tele)visual aspect of the commentary allows speakers to leave visible secondary moves, such as a scrum collapse, unmentioned; on the other hand, a primary move, such as a try, has to be verbalized in order to reflect the game's main developments.

In order to better understand the influence of individual moves on linguistic production, we present distribution of syntactic structures employed to comment two types of moves: passes and scrums, cf. Table 12 below.

The two moves are very different: a pass is a very frequent and quick move, it involves only two players and the ball is visible to the commentator; on the other hand, a scrum is less frequent, much longer, engages groups of players and the ball is not always visible.

Adopting this perspective, differences between the two graphs clearly follow. First, the frequency and speed of both moves can account for the fact that only a quarter of all passes are commented (a quick and frequent move) whereas

Table 11: Distribution of comments of the entire match moves.

Structures	Moves
No comment	970 (63,6%)
Comment	554 (36,4%)
X(...)	196 (35,4%)
PrepX(...)	87 (15,7%)
NP	60 (10,8%)
N(...)	37 (6,7%)
PP	31 (6,7%)
Part(...)	18 (3,2%)
INF	11 (2%)
ADJ	2 (0,4%)
SS	88 (15,9%)
CS	24 (4,3%)

Table 12: Distribution of comments on scrums and passes.

Structures	Scrums	Passes
No Comment	5 (23%)	145 (73%)
Comment	17 (77%)	54 (27%)
X	1 (6%)	34 (63%)
N	5 (29%)	
NP	6 (35%)	
PrepNP		8 (15%)
INF		5 (9%)
Interj	1 (6%)	
SS	3 (18%)	7 (13%)
CS	1 (6%)	

the opposite is true for a scrum (a longer and less frequent move), where only a quarter of all moves remains uncommented. Secondly, syntactic structures associated with both moves are adapted to the type of transmitted information. For a pass, the speaker usually mentions most of the time the name of the (new) ball possessor (X), sometimes preceded by a preposition (PrepX) to indicate ball movements (X and PrepX represent 77,8% of all comments on passes). In contrast, when commenting a scrum, which essentially concerns group movements, names of individual players (X) are much less frequent (about 7% of commented scrums). Indeed, other types of structures are more appropriate here: the majority of scrum comments (65%) are expressed by a nominal phrase (with or without an article, NP vs. N) and almost a quarter use verbal turns (simple and complex sentences correspond to 23% of commented moves).

3 Specific structures: impact of the sport and of the conditions of production

In order to get a better grip on the nature of linguistic invariants and the impact of extralinguistic constraints on sports commentary, we compared the data in our rugby match with two other sports commentary corpora and with the commentary of a match highlights.

First, in Section 3.1, we provide a comparison of the structures investigated in our rugby corpus with commentaries of other sports. The goal of this comparison is to demonstrate that even if we deal with the same discourse genre (live sports commentary), characteristic syntactic structures involved in the commentary are different or their distribution is different. Our discussion starts with a

very different type of sport, a relay race (corpus from Mathon 2014), and then we continue with a more similar sport, a soccer match (corpus from Audrit *et al.* 2012). In both sports commentaries, the characteristic structures identified differ from those noted in our rugby match. This observation is in line with the hypothesis put forward in (Mathon *et al.* 2015) that linguistic productions (i.e., syntactic structures here) are the result of content constraints (namely the commented sport) rather than imposed directly by the genre itself.

Next, in Section 3.2, we oppose live sports commentary (of our rugby match) and a match highlight in order to assess the impact of the conditions of production. In particular, we explore the commentary of the highlights of a basketball match from Audrit *et al.* 2012. This type of discourse, between a live sports commentary (simultaneous narration of real time moves) and a replay (the final outcome/score already known, selected images, etc.) allows us to vary production constraints and verify their influence on the commentary.

3.1 Different sports, different syntactic structures?

In this section, we compare our rugby corpus with two micro corpora consisting of the commentary of a relay race and that of a soccer match.

3.1.1 Relay race

Even if our relay race commentary is very short (2'48'' of speech time), it presents an interesting example of a very distinct type of sport: the competition is simultaneous, the time of performance is reduced to a few minutes and there are fewer events than in a rugby match.

The analysis of our relay race corpus revealed many differences as far as syntactic structures are concerned. What appears to be characteristic of this commentary are long and complex sentences (see (2)) and, more importantly, frequent and iterative repetitions (up to seven times, see (3)). This later structure contrasts sharply with other sports commentaries and should be considered the main specific structure of a race commentary.

(2) *Christophe Lemaître qui fait l'effort la Jamaïque est très bien les Etats-Unis aussi c'est pas une surprise // dans le virage Yannick Lesourd allez allez Yannick faut donner tout ce que t'as mon vieux*

'Christophe Lemaître is going for it Jamaica is going very well so are the United States no surprise there // round the bend Yannick Lesourd come on Yannick you have to give it all you've got, buddy'

(3) *elle va se battre elle va se battre // allez Floria allez allez allez allez // allez allez Floria // encore encore encore encore encore encore*

'She is going to fight she is going to fight // go Floria go go go go // go go Floria // again again again again again again'

While this is still a sports commentary, the structures are different from those of a rugby match. In particular, repetition, which is the most characteristic structure of the relay race, leads us to consider a new type of discourse in sports commentary: a cheering discourse. Indeed, the "play-by-play" comments moves, the "color commentary" provides secondary information but sequences such as *allez allez allez* 'go go go' do not fit either of these descriptions. Very frequent in our relay race corpus, this third type of discourse acts as a cheering, an emotional support for sportsmen. Due to its specific emotional and linguistic properties, this "third type" of discourse should be further examined in a separate study[1].

3.1.2 Soccer match

We also compare our data on rugby with a soccer commentary. Soccer is closer to rugby since both sports involve two teams in a confrontation during a match. From a small corpus of live soccer match commentary (13'58'' of speech time), we can see that the core distribution of syntactic structures (PC, PS, X, GN and N) is similar to our rugby corpora. Nevertheless, there are some interesting differences.

First, PrepX is nearly absent in the soccer corpus (two occurrences only) whereas this structure is very frequent in our rugby commentary (79 occurrences, corresponding to 11% of all turns). Secondly, the presence of subordinate clauses starting with *alors que* 'while', cf. (4), identified by Deulofeu (2000) as specific to soccer commentary, is confirmed in our soccer match (5 occurrences) whereas it is completely absent from our rugby commentary.

(4) a. *alors que Didier Deschamps // fait un début de match // tonitruant*
Lit.: While Didier Deschamps // does a beginning of match // resounding
'Didier Deschamps is shaving an extraordinary first few minutes'

b. *alors que l'on joue depuis trente-deux minutes*
Lit.: While we are playing for thirty two minutes
'We've been playing for thirty two minutes'

1 Initial considerations on this third type of discourse are presented in Mathon, Boyé & Kupść (to appear).

Next, the proportion of complex sentences in the soccer match (13%) is more than double from our rugby corpus (5%). Finally, interjections are much more frequent in the soccer corpus (29 occurrences, 10%) than in the rugby commentary (23 occurrences, 1%).

Although the two corpora have a different size, this comparison indicates some tendencies: the core of syntactic structures employed in both corpora is similar but their distribution is quite different (e.g., PrepX is almost absent in soccer commentary, contrary to rugby) and the structures used in both sports are not exactly the same.

3.2 Different production constraints, different syntactic structures?

In this section, we go beyond the influence of the sport on a sports commentary and we examine the impact of production constraints on syntactic structures found in the discourse. To this end, we analyze the highlights of a basketball match (from Audrit *et al.* 2012, 11'32'' of recorded time) as it presents an alternative type of sports report. A match summary differs from a live commentary by its post factum nature: there are no high stakes involved (the outcome of the game is already known) and only selected images and moves are commented.

Similarly to all previously studied corpora, the most frequent structure in our match highlights is a simple sentence (83 occurrences, corresponding to 35% of all structures). Interestingly, we find proportionally many more NPs in the summary (56 occurrences, corresponding to 26% whereas almost the same number, 57 NPs, takes up only 13% of our rugby corpus). Most of the time NPs indicate the score or point difference between the teams (30 occurrences):[2]

(5) a. *Zéro point pour Liège*
'No points for Liège'
b. *Six partout*
'Six all'
c. *Quatorze seize*
'Fourteen sixteen'

As far as marked structures from our rugby corpus are concerned (X, PrepX, XQui, N), only X still scores high (34 occurrences; 14% of all structures) but

2 In basketball, scoring is much more frequent than in soccer or rugby. Scoring is also more central to match highlights than to live comments. These two factors might play a part in the asymetric distribution of NPs here.

other structures are less frequent: 9 N (4%), 7 XQui (3%), 6 PrepX (2.5%). Although the ratios of marked structures (except for X) are roughly comparable with the numbers found in the overall rugby corpus, the absolute numbers are much lower, which is probably due to these sports' specificity, e.g., the dimensions of a basketball court lead to more dense moves and less visible passes (fewer PrepX), and to the production context, e.g., highlights lend themselves more to short compacted descriptions (number of N vs. NP structures) than a live commentary.

This brief comparison of different sports report corpora indicates that syntax alone cannot characterize the sports commentary genre. Even if some syntactic invariants might exist (e.g., proper names are present in all our corpora), variation in sports and production type result in distinct linguistic realisations (e.g., PrepX seems characteristic of a rugby "play-by-play" commentary but neither of a rugby match in general nor of other sports). The syntax alone does not define the sports commentary genre but, according to the external constraints of a specific sport and of a specific type of discourse (live or not), it allows to describe a sports commentary of this sport.

Given the size of our corpora, our conclusions are only indicative. In order to assess the exact influence of sport and production type on the discourse, a sports commentary of the same sport event should be considered with respect to different media (television, radio, live written commentary, summary, etc.) and commentaries of other sports should be included (e.g., figure skating, skiing). We leave this for our future study.

4 Genre beyond syntax and words

Sports commentary being a phonogenre, i.e. a genre of oral discourse, we investigated the weight of phonetic cues (especially prosodic features such as rhythm, pitch and intensity) in its identification as a genre. We conducted two perceptual studies on subjects who were asked to identify sports in different live TV commentaries by listening to a recording scrambled to make the actual text unrecognizable. For that purpose, we used two sets of low-pass filtered stimuli, which were first extracted from a corpus of different live TV commentaries (local and national elections, a military parade, a live report on demonstrations in Thailand) and live sports commentaries. For the latter, we included:

- Races: snowcross, relay race (4x100m)
- Matches: rugby (Lortal & Mathon 2008) and soccer (Goldman *et al.* 2014)
- Individual performances: slalom skiing, half-pipe snowboarding
- Artistic sport: figure skating

4.1 Sports recognition

The first perceptual experiment (Mathon 2014) was conducted with 4 sets of 12 stimuli, assessed by 63 French subjects.

The stimuli were extracted from various TV recordings. For the live sports commentaries part, we used a sample of our France-Argentina rugby match (1st match of the 2007 World Rugby Cup). The other sports stimuli were taken from TV recordings of Sochi's Olympic Games (2014): snowcross (1 race), slalom skiing (1 run), half-pipe snowboarding (1 run) and figure skating (1 program). For the other commentaries, we used a recording of the July 14th French Military Parade and live commentaries during a special evening TV program for the French general elections (2012).

The stimuli were selected on the basis of 3 criteria:

- Duration of the move reported
- Prominence of the move with respect to the whole performance
- Emotion involved in the move

5 categories were used to classify the 38 stimuli: Match (8 stimuli), Race (4), Individual Performance (10), Artistic Sport (4), Other Media Event (12). The subjects were asked to listen to each stimulus and choose the best fit among the tags (Match, Race, Individual Performance, Artistic Sport, Other) to characterize the type of event reported.

The results of this first experiment showed that subjects discriminated sports commentaries from other media event commentaries, but they made no distinction between the different types of sports (match, race, individual performance), except for figure skating (perhaps because of the musical dimension of this sport) (Figure 1).

4.2 Sport detection

Following this experiment, where sports were well detected but not really identified, we conducted a second experiment, focused on sport detection. We excluded musical sports like figure skating from the corpus of sports stimuli, as music seems to have been a very important identifying clue in the first experiment. We also extended the corpus of other media events, choosing media events much closer to sports commentary than the ones used in the first experiment regarding the emotional load (celebration of victory in presidential elections), rhythm and predictability (live report on the demonstrations in Thailand) (Table 13).

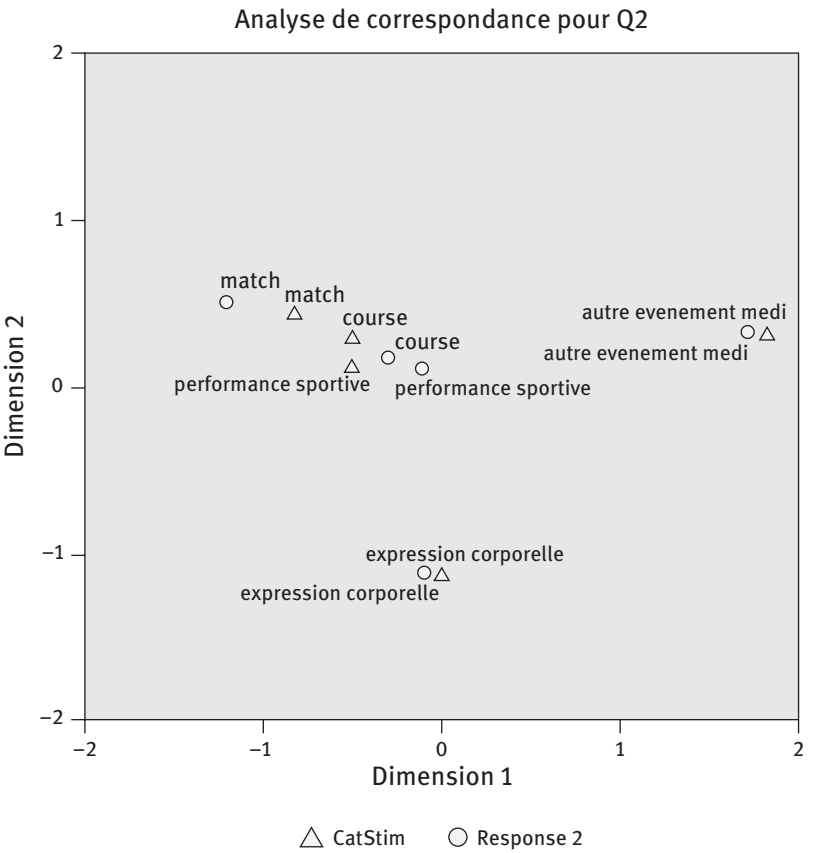


Figure 1: Sports recognition: multiple correspondence analysis (from Mathon, 2014)

Table 13: Corpus of Live TV commentaries used as material for the perceptive test stimuli.

TV live comments						
Sports comments				**Comments of other media events**		
Match		**Race**		**Other Media Event**		
Rugby (Lortal & Mathon, 2008)	Football (Goldman et al., 2014)	Relay race 4x100 m men	Relay race 4x100 m women	Sport's TV Emission	Elections	Live report demonstration in Thailand
World Rugby Cup 2007	World Football Cup 1998	Athletism World Championship 2011	Athletism European Championship 2014	July, 2000	French Presidential Elections 2012	Le Petit Journal, 2014/01/20
In French						

We extracted 24 stimuli from this corpus on the basis of 3 criteria:

- We defined 6 categories of live TV commentaries: rugby, football, relay race, sports TV show, elections and live report during demonstrations;
- We opposed "play-by-play" to "color commentary" (Hartmann 2013) and chose 12 stimuli of each type;
- We distinguished parts of discourse which contained specific syntactic structures (like those described in figure 2) from parts of discourse which contained "classic" syntactic structures (simple or complex sentence).

We asked 53 subjects to listen to 8 stimuli (among a pool of 24 stimuli) and decide if a sport was being reported or not (Figure 2).

The results show a better identification with "color commentary" stimuli, even if those contain far fewer specific syntactic markers than the play-by-play. We can hypothesize that, independently of the media event which is reported (sports or something else) when the constraints that govern discourse production are quite similar in rhythm, in predictability and in emotion, then the phonostyles (i.e. the prosodic productions) are very similar to sports styles.

In order to better understand these results, we have provided a study of speech rhythm (cf. Mathon et al. 2016 for details) in relation to moves and discourse types (play-by-play vs. color-commentary). Our initial analysis concerns just one rugby match and it is focused on the game rhythm. We have decided to study different prosodic factors separately in order to verify the impact of discourse production constraints on each of them. Hence, we have formed our first hypothesis that speech rhythm is most directly influenced by the medium's constraints and more specifically by the synchronicity of discourse and game moves.

In order to verify this hypothesis, we have aligned game rhythm (calculated in number of moves per 5 seconds) and speech rhythm (speech and elocution rate in number of syllables per second). The results of this analysis show that an

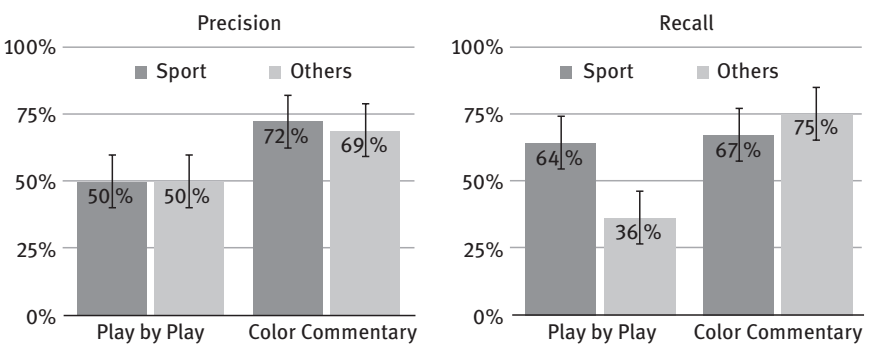


Figure 2: Sports detection: Color Commentary vs Play by Play.

increase, or on the contrary, a decrease in game rhythm does not systematically correspond to the same change in the commentator's speech rhythm. Even if we find sequences of play-by-play comments where the journalist's speech rhythm rises to follow the game rhythm, we find also play-by-play comments where the speech rate decreases or remains stable during a fast game phase. In such cases, the necessary adjustment is delegated to syntax: use of short nominal phrases allows the commentator to adapt his commentary to the game rhythm without necessarily increasing his speech rate. Similarly, during color-commentary sequences, the sports expert who wishes to provide as much information as possible (the constraint Max(Comp) in (Augendre et al. 2014)) has to accelerate his speech rate in order to fit all necessary comments in a relatively short period of time. This rapid speech rate could perhaps explain a better detection of sports commentary during color-commentary in the experiment described in sec. 4.2.

5 Conclusion

Both syntactic analysis and recognition tests showed that linguistic structures are not sufficient to define sports commentary as a discourse genre, even when linguistic productions are linked to contextual information. Because of the visual dimension of live TV sport commentary, the rhythm of the sport and the sequence of moves must be taken into account as constraints which influence the structures in the discourse.

As Maingueneau (2002) points out "in French discourse analysis, the category of 'discourse genre' is defined, as a rule, by situational criteria", which means that the socio-pragmatic dimension is preferred in order to characterize discourse genre. Thus, Pršir et al. (2013) for example select situational criteria to discriminate sport live report from seven other phonogenres such as liturgy, reading, or conversation. For some discourse genres however, like a live sport report, a definition of genre with respect to situational criteria only leads to a multiplicity of phonostyles depending for example on the nature of the sport (Audrit et al. 2012). It leads us to suggest a different point of view when defining discourse genres by integrating the notion of constraints.

The perceptive tests highlighted two things:

– Sport can be discriminated from other media events, as long as the production constraints are significantly different. The closer the constraints, the stronger the confusion. Consequently, the material dimension seems to have a strong impact on the perception of the genre, and must be taken into account.

– Characterization and detection of genre seem to have little to do with syntactic structures. Indeed, inside a genre such as a live sport report, syntactic structures vary, depending on the sport reported, the rhythm of the move reported etc. Moreover, when it comes down to discriminating sport from other media events, which are similar in terms of production constraints, then the best recognition score is on the text with the least specific syntactic structures (i.e. on color commentary).

This leads us to conclude that different dimensions used traditionally to define a genre (see Richer 2011) are not completely effective for a live sport report. Neither socio-pragmatic (context), nor stylistic (linguistic units used), nor the material dimension (medium characteristics) taken separately are sufficient to provide a satisfactory account of the data. We suggest that these different dimensions should be linked with three sets of constraints: medium, genre and content (see Mathon et al., 2015).

Medium constraints are linked to the material dimension. In fact it's more a list of material circumstances becoming constraints:

– The nature of the medium, for example, a TV show implies that there is a form of synchronicity between the sport's moves (what happens on the TV screen basically) and the discourse;
– The temporal and spatial dimensions of the sport's moves have an impact on how much time the journalist can spend on reporting a move, or on bringing more information about the sport's strategies;
– The global context in which the sport event takes place also has an impact on the discourse, like the importance of the competition (local, national, international competition or Olympic Games) or the fact that the national team is involved in the competition. Those circumstances influence the emotional charge conveyed in the discourse.

Content constraints are linked to the thematic dimension: sport is like a general theme and a sport report must be about sport. But more precisely, a sport report must be about the sport which is broadcast, and depending on the nature of the sport, the discourse will be adequately adopted. That's why we found some definitely different linguistic structures from rugby to soccer, or basketball, even if these sports seem to be very close (a competition between two teams, a ball involved, a field/court game, team strategies, etc.).

Genre constraints are in the center of all the dimensions. They are linked to socio-pragmatic, medium and thematic dimensions and of course they play a part on the textual and stylistic dimensions. Among genre constraints, we can note an emotional constraint. Considering that sport reports are a media genre,

a lot of emotional charge must be conveyed in the discourse in order to keep the audience's attention high.

At a syntactic level, we showed that the structures used in the sport commentaries studied are the result of these constraints, but also, according to the differences observed between three sports (rugby, soccer and relay race) and two production contexts (a live match commentary vs. match highlights), that the genre only imposes some constraints but not a list of structures. Indeed, a live (medium constraint) commentary (content constraint) implies the presence of unusual syntactic candidates just because the "normal" syntactic candidates do not match the constraints associated with this kind of discourse.

The confrontation between the analysis of syntactic structures and the perception tests confirms the impossibility to define the discourse genre in terms of syntactic choices only. For example, we showed that the specific structures mostly appear in the journalist play by play discourse whereas the perception test showed that the identification of the discourse genre is not promoted by this part of the discourse but by the color commentary.

Sports' report is a very constrained genre of discourse. However, it's not the only one and the model of constraints and dimensions we drew above can be applied to other genres. For example, let us consider newspapers' classified ads. The text has to be short because of fixed price per number of words or lines in the ad. The longer the ad, the more it costs. This swants to transmit and shorten his sentences. As a result, classified ads show specific structures like verbal phrases without pronouns. This is an example of a medium constraint applied to the stylistic dimension. Let's take now the same classified ad edited on a website which doesn't charge for the length of the published ad. The textual organization of the classified ad is very different: you can find a title and an introduction which have the same features as the not-free-of-charge classified ad, i.e. short, with the same specific syntactic structures, and then a much more descriptive text, with quite normal syntactic structures (complete verbal phrases). In this case, the medium constraint doesn't play any role, but it's integrated by the writer and the receiver as a genre constraint. In order to be considered a classified ad, the text has to be short, and contain these specific syntactic structures. TV sport commentary is thus one of the discourse genres in which the constraints (rhythm, sports characteristics, medium constraint, etc.) are very strong and specific to the event being described. Among the different discourse genres, some are also very constrained, such as classified ads or paid short text messages for which words/characters count is the main constraint (fewer words, lower price), whereas other genres, for example a thriller, might not obey such constraints.

To conclude, we can add that a discourse genre is in some way institutionalized: from production constraints, we obtained specific structures only valid for one type of discourse of the genre but not for all the discourses of the genre. However, a structure very frequent and associated to a popular event/sport tends to be institutionalized and thus recognized as specific for the genre "sport commentary" even if it is only representative of one type of sport event, a match for example (and not a race, an exhibition...). Here lies the limit of the definition of a discourse genre as a homogeneous set and thus the necessity, for each discourse, to consider its production constraints and the specific answers to these constraints provided at linguistic units level.

References

Adam, Jean-Michel. 2001. En finir avec les types de textes. In Michel Ballabriga (ed.), *Analyse des discours. Types et genres : Communication et Interprétation*, 25–43. Toulouse: Editions Universitaires du Sud.

Audrit, Stéphanie, Tea Pršir, Antoine Auchlin & Jean-Philippe Goldman. 2012. Sport in the media: a contrasted study of three sport live media reports with semiautomatic Tools. In Ma Qiuwu, Ding Hongwei & Daniel Hirst (eds.), *Proceedings of the 6th International Conference on Speech Prosody* (vol. 1), 127–130. Shanghai: Tongji University Press. http://www.speechprosody2012.org/page.asp?id=157

Augendre, Sandra, Catherine Mathon, Gilles Boyé & Anna Kupść. 2014. Influence des contraintes extra-linguistiques sur le discours : cas du commentaire sportif télévisé. *Actes du CMLF 2014 – 4ème Congrès Mondial de Linguistique Française*, 1905–1924. Paris: EDP Sciences. http://dx.doi.org/10.1051/shsconf/20140801381

Bakhtine, Mikhaïl. 1979. *Esthétique de la création verbale*. Paris: Gallimard.

Canvat, Karl. 1999. *Enseigner la littérature par les genres*. Bruxelles: De Boeck Duculot.

Deulofeu, José. 2000. Les commentaires sportifs constituent-ils un "genre", au sens linguistique du terme?. In Mireille Bilger (ed.), *Corpus: Méthodologie et applications linguistiques,* 271–295. Paris: Champion.

Goldman, Jean-Philippe, Tea Pršir & Antoine Auchlin. 2014. C-PhonoGenre: a 7-hour Corpus of 7 Speaking Styles in French: Relations between Situational Features and Prosodic Properties. *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC 2014), 202–305. Paris: ELRA.

Hartmann, Claudio. 2013. *Pre-fabricated Speech Formulas as Long-term Memory Solutions to Working Memory Overload in Routine Language*. PhD Thesis, Zurich: UZH.

Krazem, Mustapha. 2011. Représenter les relations entre grammaire et genres de discours: exemple des commentaires sportifs, *LINX* 64–65, 45–68.

Lortal, Gaëlle. & Catherine Mathon. 2008. Motion and Emotion or how to align emotional cues with game actions. *Proceedings of Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology EMOT – LREC 2008*, 72–78. Paris: ELRA.

Maingueneau, Dominique. 1998. *Analyser les textes de communication*. Paris: Dunod.

Maingueneau, Dominique. 2002. Analysis of an Academic Genre, *Discourse Studies* 4. 319–341.

Mathon, Catherine, Gilles Boyé, Sandra Augendre & Anna Kupść. 2015. Contraintes sur le discours et genre de discours contraint : le commentaire sportif télévisé en direct, *Discours* 17. http://discours.revues.org/9082; DOI: 10.4000/discours.9082.

Mathon, Catherine, Gilles Boyé & Anna Kupść. To appear. Commentaire sportif en direct: Etude des correspondances entre le rythme du jeu et le rythme de parole. *Proceedings of CMLF 2016*.

Mathon, Catherine. 2014. Perception des phonostyles et représentativité du phonogenre: le cas du commentaire sportif en direct. *Nouveaux cahiers de linguistique française* 31. 93–103.

Petitjean, André. 1991. Contribution sémiotique à la notion de "genre textuel". *Recherches linguistiques* XVI. 349–373.

Pršir, Tea, Jean-Philippe Goldman & Antoine Auchlin. 2013. Variation prosodique situationnelle: étude sur corpus de huit phonogenres en français. In Piet Mertens & Anne Catherine Simon (eds.), *Proceedings of the Prosody-Discourse Interface Conference 2013 (IDP-2013)*, 107–112, Leuven. http://wwwling.arts.kuleuven.be/franitalco/idp2013/Proceedings.html

Richer, Jean-Jacques. 2011. Les genres de discours : une autre approche possible de la sélection de contenus grammaticaux pour l'enseignement/apprentissage du F.L.E.?. *LINX* 64–65. 15–26.

Richer, Jean-Jacques. 2005. Le Cadre européen de référence pour les langues: Des perspectives d'évolution méthodologique pour l'enseignement/ apprentissage des langues?. *Synergies Chine* 1. 63–71.

Georgeta Cislaru and Thierry Olive

Bursts of written language as performance units for the description of genre routines

Abstract: We analyze the writing and re-writing processes which jointly participate in the configuration of a written genre, for instance social reports on at-risk children. Social reports on children at risk are examples of professional writing that are both institutionally and socially constrained. They are "routinized discourse genres" inasmuch as their form and content are pre-defined and their production is supposedly routinized and anonymized. The results of the study are discussed in terms of the communicative competence/performance contrast, along the lines of Hymes' proposals. We adopt a longitudinal viewpoint for data collection and analysis. In order to detect specific or recurrent linguistic structures, we perform a twofold analysis. First, we analyze the content of the bursts of writing. Second, we analyze the content of repeated segments (or n-grams), which are linguistic strings that are reiterated within a text or a corpus; repeated segments are assimilated to discourse routines. The goal is to determine whether the content of bursts and of repeated segments is similar. The real-time data analyzed show that lexical strings that are repeatedly used in finished written discourse are generally not produced as blocks or bundles. More often than not, the social workers automatically produce specific non-routinized strings. This means that the strings usually called "prefabs" in the literature and considered as memorized formulae are not part of a discursive stock, but – at least partly – the result of adaptation strategies.

1 Introduction

The production of texts and discourses obeys specific rules and constraints (see Plane et al. 2010) that are at least partly determined by genre parameters (Bakhtine 1984; Branca-Rosoff 1999). Text structure, lexical choices, syntactic patterns, pragmatic markers, and communicative frames are the units which help characterize genre affiliation. They are tacitly adopted, and sometimes taught (see, for example, Swales 1990 on academic genres). Schemes, scenarios, and routines are part of genre specificities, inasmuch as discourse genres articulate constraints on form and meaning.

Georgeta Cislaru, CLESTHIA, Université Sorbonne nouvelle, France
Thierry Olive, CNRS & Université de Poitiers, France

https://doi.org/10.1515/9783110595864-010

In this study we analyzed the writing and re-writing operations which conjointly participate in the configuration of a written genre, social reports on at-risk children (Branca-Rosoff and Torre 1993; Pugnière-Saavedra 2008). Social reports on children at risk are routinized, nonvariational discourse genres; their form and content are pre-defined and their production is routinized and anonymized. As such, social reports are both institutionally and socially constrained. To understand the way writers take into account the constraints related to this genre of text, we analyzed text segments that were identified through statistical exploration of textual data[1] (textometry) and keystroke logging tools. While textometric analyses rely on the finished texts only, keylogged writing allowed us to analyze text segments that were delimited by writers' behavior and thus may be defined as production units. Keystroke logging tools record all the textual operations carried out by a writer during the composition of his/her text (Spelman-Miller and Sullivan 2006), making visible the orthographic, lexical, or syntactical adjustments in written texts. We thus combined hybrid methods of corpus analysis based on real-situation text production.

The results of the present study will be discussed in terms of the communicative competence/performance contrast, along the lines of Hymes' (1984) proposals. We first discuss the concept of *genre* and the specificity of the genre of social reports. We then describe the methods used for real-time recording of the writing process and the parameters of the collected corpus. The last two sections focus on linguistic analyses of bursts of writing, *i.e.* segments of texts that were produced spontaneously and without interruption. In Section 4, we compare the contents of bursts and of the repeated segments, *i.e.* graphical chains of two or more units repeated at least twice in the finished texts, in order to determine the way discourse and genre routines are processed. Section 5 deals with coordinative bursts, and discusses their role in text organization and genre configuration.

2 Social reports on at-risk children, a case of online professional discourse

2.1 What's (in) a genre? Textual beginnings as genre outline

The question of discourse genre is not a straightforward one and is related to the social functioning of texts (see Miller 1984). Most theories of genre (Bakhtine 1984;

1 It includes the observation of lexical distribution and regularities across a corpus of texts, collocations, co-occurrences, etc.

Bhatia 1993; Swales 1990) postulate a form-function correlation, and thus place the definition of genre on the ground of the articulation between linguistic structures and communicative purposes[2]. A correlation between form and function does not imply that the same form corresponds to the same and unique function in a ubiquitous manner, but rather the specialization of a form or a group of forms meant to embody a function in a specific series of texts (cf. Bhatia 1993). Various constraints underlie genre configuration, and adjust form-and-function or, rather, form to function, given that discourse genres emerge from the differentiation of social practices (Rastier 1989: 37), and constitute socio-communicative patterns organizing discourse production (Adam 2011b: 33). Many definitions of genre grant an important role to collective acknowledgement and validation, based on the assessment of the felicitous or infelicitous fulfillment of the social purpose[3]:

> A genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community, and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style. Communicative purpose is both a privileged criterion and one that operates to keep the scope of a genre as here conceived narrowly focused on comparable rhetorical action. (Swales 1990: 58)

> [...] the term *genre* has been used to refer to a culturally recognized 'message type' with a conventional internal structure, such as an affidavit, a biology research article, or a business memo. *Genre* studies have usually focused on the conventional discourse structure of texts or the expected socio-cultural actions of a discourse community. (Biber, Connor and Upton 2007: 8)

This *socially-grounded* conception of genre logically leads to hypotheses about text-configuration strategies, and questions i) the ways in which collective validation is anticipated through text production, and ii) the place of language structures within the global mechanism. Considering social reports, what a judge, their direct addressee, expects is not only a text that has all the features of the "social report" genre, but also a text that meets its communicative purpose and will help decision-taking.

Some authors focus on the capacity of readers to recognize a genre either by using a prototype, an ideal text-model, or by exploiting learned dominant features; this expertise triggers readers' expectations as to the contents and structure

2 See Askehave (1999) for a critical review of the concept of communicative purpose.

3 See, for instance: "[...] each genre is an instance of a successful achievement of a specific communicative purpose using conventionalized knowledge of linguistic and discoursal resources." (Bhatia 1993: 16)

of a genre (Beghtol 2001; Bax 2011). The notion of readers' expectations is helpful for the projection of a text-in-process towards a model, a collection of features that meets these expectations. As noted by Flower (1979), while novice writers produce writer-based texts that "fail to communicate the same meaning to a reader", expert writers produce reader-based texts, that anticipate the reader's expectations (Scardamalia and Bereiter 1991). Although not directly referring to an expert writer's skills, Bhatia (1993: 19-20) offers an interesting overview of the "individual strategic choices made by the writer in order to execute his or her intention":

> These tactical choices, appropriately called **strategies**, exploited by a particular writer are generally used in order to make the writing more effective, keeping in mind any special reader requirements, considerations arising from a different use of medium or prerequisites or constraints imposed by organizational and other factors of this kind. Such strategies are generally non-discriminative, in the sense that they do not change the essential communicative purpose of the genre. Non-discriminative strategies are concerned with the exploitation of the conventional rules of the genre concerned for the purpose of greater effectiveness in a very specific socio-cultural context, originality or very special reader considerations. (Bhatia 1993: 19-20)

We come now to the processing and the functioning of language structures in relation to genre configuration. Some authors consider that contrastive genre analysis should not be based on strict linguistic differences or oppositions; they suggest rather that there are preferred/dispreferred linguistic structures for specific genres (Biber 1988; Bax 2011): some structures are statistically preferred in some texts for their capacity to strongly articulate the intended meaning and communicative purpose to structural expectations. The selection of specific linguistic structures, in order to present the information in the most appropriate way to fit the expectations and background of the reader, corresponds to what Bhatia (1993: 26) calls text-patterning, or textualization (following Widdowson 1979). Written segments of text that are produced in a burst, without any interruption, might be part of these strategies and thus become a relevant level of linguistic analysis which highlights the tactical aspect of conventional language use, where specific linguistic features textualize specific aspects of genre. Form and function are inseparably linked, but each type of form-function correlation is valid in the frame of a particular genre.

2.2 The genre of social reports: a situational and discursive description

The genre of social reports has seldom been studied (cf. Branca-Rosoff and Torre 1993; Léglise 2004; Pugnière-Saavedra 2008). In the present study, the identification and the definition of the genre were developed together with the group

of social workers we worked with by questioning them about their social and writing practices, as we recognize, with Bhatia (1993: 34), the interest of specialist information in genre analysis, inasmuch as it is the cumulative "experience and/or training within the specialist community that shapes the genre and gives it a conventionalized internal structure" (Bhatia 1993: 14). We conducted an ethnographic study of the genre of social reports, through meetings and discussions with the social workers, collecting information about their professional activity, and presenting the initial results of our linguistic analyses to them. This enabled us to better understand their professional practices, the writing process and the place social reports have in the social institution.

Social reports on children at risk are examples of professional writing that are both institutionally and socially constrained. They have explicit performativity and complex social purposes, due to the complex addressee. To sum up, the social reports we analyzed are contextually characterized by the following features[4]:

- Collective writing: Several authors wrote and revised a report. The body of the report was written by social workers, but the conclusion was (re)written by the section head of the child protection unit. Besides, social reports are professional writings closely linked to orality (see Delcambre 1997), each case being discussed during work meetings. Authorship therefore has to be understood as a collective responsibility, implying collective norms of writing and assessment;
- An official addressee in the shape of the judge, who requested the report in the first place. In our corpus of social reports, there were two kinds of texts: regular progress reports, and final reports that summarized the overall situation. Although the latter were written with the judge in mind, they might also be consulted and used by the families during the court hearings;
- An unofficial addressee in the shape of the family, who were given the right to access social reports in the 2002 Act reforming French social services. This means that while social workers must describe the child's situation as accurately as possible, with some judges even providing guidelines (cf. *Guide de la protection judiciaire de l'enfant*, Huyette and Desloges 2009), they must also protect the child and its family, and avoid offending them in the report.

The distinction between routine genres and variation-based genres (Adam 2011a; Maingueneau 2004 and 2014) places social reports among routine genres. Due

4 Detailed linguistic descriptions and discourse analysis of these social reports can be found in Cislaru et alii (2008), Cislaru et alii (2013), Née et alii (2014), Sitri (2013), etc.

to their institutional frame-and-functioning, the genre of social reports is less exposed to hybridization. Routine constraints may be identified at two main levels:

a) The situational level: Social reports are produced by an institution. They punctuate a social action, either making an intermediate assessment or providing a concluding evaluation of the measure taken. Each particular situation is subject to the same procedure, and the social report is part of this process.
b) The textual level: All the reports that constituted the corpus were produced within the same social service, the SAFE[5] of the City of Caen, France. All the reports have comparable structures: They all begin with the presentation of the family, followed by a brief history of the case, and all end with a conclusion. The subsections in the body of the report are almost identical: Education, health, relations with the parents/family, etc. A basic outline is offered by the direction of the service, although following it is not compulsory. The introduction of a frame of reference for all the social reports throughout France is part of a ministerial project in the past ten years. Social workers are generally strongly opposed to this idea, as they consider it impossible to strictly apply frames of reference to each particular case.

The classification of social reports as routine genres entails the issue of text writing routines and questions the nature and the place of routine units in the configuration of the genre. This will be addressed in Section 4.

2.3 Corpus data

The writers of the reports we collected were social workers and youth workers belonging to an association that worked with more than 20 foster children aged 5–21 years. The social workers collected field data while visiting the family / child, and during meetings with the parents. They then had to summarize this information and write regular progress reports and a final report describing the situation and progress of each child, and recalling the family history and the reasons for foster care. The reports were presented to the judge and were used to help him/her take a decision in accordance with the social protection measure.

All the reports had comparable structures, divided into the same sections and with comparable section headings (presentation of the family, history of

5 Service d'Accompagnement de la Famille et de l'Enfant (Family and Child Guidance Service).

the social measure, schooling, health, relationship with the parents and other members of the family, conclusions). There was even an institutional template, although this was not overtly followed. The reports had complex discursive features, since they combine different discourse modes (narration, description, evaluation, and argumentation). The production of social reports therefore requires specific linguistic competence.

We analyzed 10 computer-written social reports. We adopted a longitudinal viewpoint for data collection and analysis. The corpus was recorded and collected with Inputlog, a keystroke logging program (Leijten and van Waes 2006) that records all the actions that a writer performs with the computer when composing a text using a word processor, keyboard, and mouse. In the present case, each key press, along with its timing, was recorded, as well as each move of the mouse in the text or in the menu of the word processor that the writers used. All textual operations to modify the text were also recorded. The tool cannot be installed and used without the consent of the writer, and all the collected metadata were anonymized. Recording of the reports started from the first time the social workers began writing a report. We thus recorded all writing sessions and collected all the versions of the reports during the composing process.

3 A longitudinal study of spontaneous segmentation

3.1 Bursts of writing, the linguistic units of online writing

Roughly, a writer is busy writing about half of the time and pausing the other half (Alves et al. 2007). Bursts of writing are the result of a spontaneous segmentation of language units during the writing or revising process. This segmentation is highly comparable to segmentation in verbal production, i.e. sequences of text production (or transcription) that alternate with pauses. Pauses are generally used to mentally prepare the next sequences of text or to read segments of text previously produced. It is important to point out that a majority of pauses (at least 75%) occur for purely mechanical reasons. In handwriting, pauses occur for example when writers write the dot of an ‘i’, when they move the pen to write another word, or when they stop a sentence and begin the next one on a new line. The presence of a pause is also determined by cursive or script handwriting. In cursive handwriting, the majority of letters in a word are linked and so there are a few within-word pauses, whereas with script handwriting, several pauses occur within letters and within words. In typing, each key press, each action is separated by a pause. In both modes of writing, sequences of transcription are

therefore identified after these mechanical pauses have been removed by defining an appropriate pause threshold (see below). From a cognitive point of view, pauses which are not mechanical concern three main writing activities: planning the text content, preparing grammatical structures, revising the already written text. Pause duration is generally interpreted as resulting from the complexity of the underlying cognitive process. By contrast, transcription periods are moments during which the writer produces text continuously, i.e., without pausing. The text sequences – letters, words, phrases, clauses, etc. – that are produced during these transcription periods are called bursts of production (Chenoweth and Hayes 2001). In our corpus, the sentence *une cousine qui peut venir partager du temps avec elle pendant le week-end* was produced in seven bursts:

[pause] *une cousine qui* [pause] *peut venir partager du temps avec elle pendant* [pause] *le* [pause] *w* [pause] *eek* [pause] – [pause] *end*. [pause]
[pause] *a cousin who* [pause] *can come and spend some time with her during* [pause] *the* [pause] *w* [pause] *eek* [pause] – [pause] *end*. [pause]

Pauses are therefore boundaries that segment textual data and produce empirically relevant linguistic units, the bursts of production. As indicated above, pause duration is crucial for the identification of bursts of language. The question of how to define a threshold for discriminating between mechanical and cognitive pauses is still a matter of debate in psycholinguistics (see Chenu, Pellegrino, Jisa and Fayol 2014; Wengelin 2014), and further in-depth linguistic and psycholinguistic analyses are necessary to specify different relevant thresholds, adapted to specific styles of writing. Although there is no consensus about an ideal threshold for discriminating between mechanical and cognitive pauses, several studies that analyzed bursts of production used a 2-second threshold (see Chenoweth and Hayes 2001; Alves et al. 2007; Baaijen et al. 2012). Accordingly, we also used a 2-second threshold which allowed us to exclude all pauses that resulted from typing movements (for example, moving the hands and fingers on the keyboard to reach the next key, or preparing a combination of keys to type a diacritic character), and whose origin thus did not lie in the operations of one of the cognitive processes of writing.

Kaufer, Hayes, and Flower (1986) conducted the first study that investigated bursts of production in English. They showed that adult writers typically compose by producing segments of text with an average length of 9 words. They also observed that more skilled writers produced longer bursts (in length and in number of words: four words more on average) than less skilled writers. Since the texts written by the experts were generally rated of better quality than those composed by novices, the authors interpreted this increase in burst size and length as

evidence of more efficient translating (i.e., linguistic formulation in writing) processes. This interpretation was later confirmed by Chenoweth and Hayes (2001). Chenoweth and Hayes' (2003), and Hayes and Chenoweth's (2006) studies completed these findings by showing that impairment in verbal working memory, a cognitive system that is required in translating concepts into language, consistently decreased burst length and writing fluency. Bursts of writing seem thus to result from cognitive and graphomotor automatisms, and function as writing routines.

Kaufer et al. (1986) also showed that bursts have a strong tendency to end at clause boundaries, and less so at phrase boundaries. Thus, according to these authors, writers compose sentences by first selecting a topic, and then by producing and evaluating sentence parts that fit grammatically with the part of the sentence that has already been prepared. If the evaluation is negative, the writer has to either revise the current part or produce an alternative part to follow it. If the evaluation is positive, then the sentence part is added to the current sentence that has already been produced, or that is still in the writer's mind (i.e., in verbal short-term memory). The fact that a topic is first selected – but these data may not apply the same way to languages other than English – may suggest that communicative purposes shape the bursts. Therefore, analyzing the contents of bursts may open up new perspectives on the construction of meaning during text production.

To summarize, bursts of writing are text sequences that reflect the dynamics of textualization; they are the result of a spontaneous segmentation, which is cognitively shaped, but also determined by the functional features of each personal writing process.

3.2 Theoretical outline and methodology

Beyond variations of (re)writing (Grésillon and Lebrave 2008), the linguistic study of bursts raises the question of a process-grammar that frames writing competence and performance. The notion of process-grammar finds root in some recent descriptions of oral and written discourse. Taking a communicative-pragmatic viewpoint, Brazil (1995) considered that language units are used in communication in order to construct meaning. He proposed a purpose-driven process grammar of speech, i.e. a description of the way in which we produce and interpret linguistic units in real time. The selected unit of analysis was not the sentence, but sequences parallel to a phonological tone unit. In line with Brazil's work, Sinclair and Mauranen (2006) attempted an articulation of writing and speech, and offered an account of the ways "speakers and listeners

chunk language in manageable units as they participate in language-mediated interaction" (Sinclair and Mauranen 2006: 40). Their approach was based on the assumption of the linearity flow principle, and chunks were seen as ways to process and sort the incoming information. The boundaries of the chunks were fixed intuitively, and it is regrettable that no real-time process data were used to check chunking. They classified chunks in two basic types: organization chunks and message chunks, each type being subdivided in several subclasses. The idea of functional classifications is particularly interesting in the context of the present study, as it provides different frames for the description of bursts. The analysis of real-time writing processes and linguistic description of bursts provides evidence for chunking boundaries and shows the way text production is configured. It also gives a glimpse of genre constraints on form and content articulation.

Our analysis is based on hybrid methods of corpus analysis supported by real-situation text production (see Biber 2009, for instance). The 5 709 bursts we analyzed were produced during a real-time and real-situation writing activity. We compared the linguistic contents of bursts of writing with segments of texts occurring at least twice in the final texts of the reports (see also Olive and Cislaru 2015). Repeated segments (RSs) are strings of at least two graphical units that occur together at least twice in a text or a corpus (Lafon and Salem 1983). Textometric software detects them by their graphical form. RSs are considered to be ready-to-speak units (which are somewhat different from collocations): in the framework of textometry and discourse analysis, they are considered as discourse routines that characterize either a language or a type of discourse (Cislaru et al. 2013; Née et al. 2014). Biber (2014: 4) argues that "[...] linguistic co-occurrence reflects underlying communicative functions". Accordingly, repeated segments are viewed as key elements of text organization in the framework of discourse analysis and corpus linguistics, as they signal discourse routines related to genre, social sphere of activity, professional domain and occupation, etc. The list of repeated segments (7 246) is corpus-driven, and we operated with raw data, without sequence pre-selection. Both bursts and repeated segments were described grammatically (in the terms of constituent analysis by using predefined grammatical categories) and semantically.

We used a Python script developed by Lardilleux (Lardilleux et al. 2013) to extract the bursts from Inputlog's log files and automatically detect the repeated segments. The script begins by calculating all bursts produced in the different versions of a report. As Inputlog's log files are saved in time-based XML format (each action on the keyboard or with the mouse is logged with its time of occurrence), sequences of characters between two pauses are identified by calculating the time (a pause) between events: each sequence of key

presses which is produced without interruption constitutes a burst of written language production. The script then calculates the different repeated segments present in a report with an algorithm that compares all the strings of characters in order to find similar ones. Similarity can be modulated and in the present study we used a 75% similarity level. Finally, the script compares and aligns the repeated segments and bursts, as shown in Figure 1. The right frame gives counts of repeated segments (columns to the left of the text) and bursts (columns to the right of the text) recorded in a complete file of drafts for one report. The outermost numbers indicate the total number of units that are represented by a given set of related repeated segments (far left of the frame) and bursts (far right of the frame). In the left frame, the upper part shows context for the selected repeated segments, while the bottom part shows the neighborhood of the selected bursts, with temporal data (the first column indicates the timing within the time log for the writing session, and the second column the length of the burst in seconds).

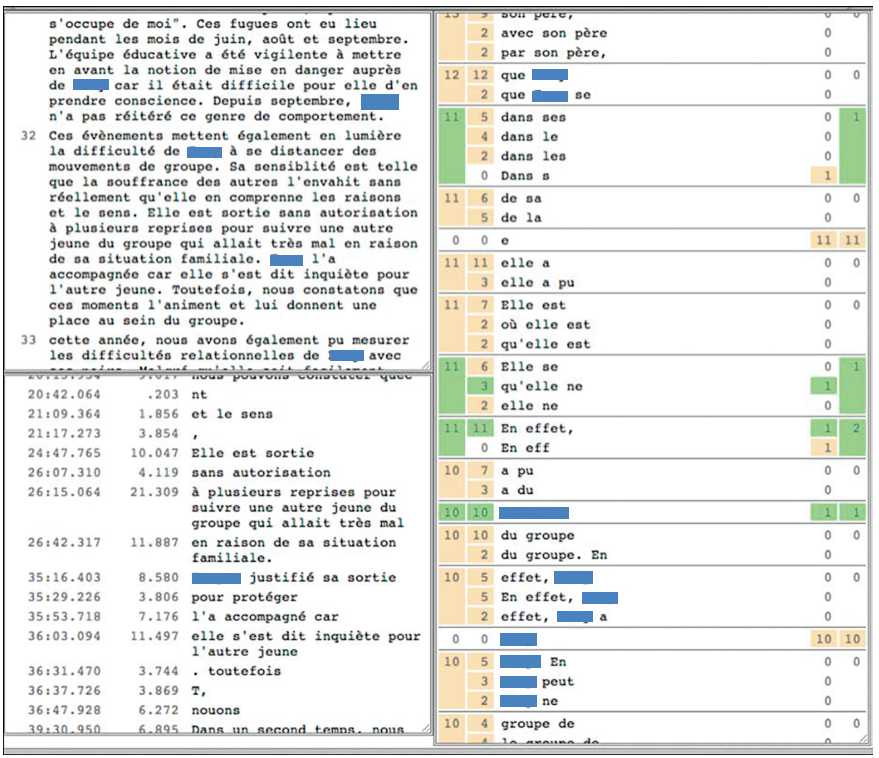


Figure 1: Alignment of bursts and repeated segments, with direct access to the text.

4 Genre constraints, routines, and spontaneous production

4.1 A comparison between repeated segments and bursts

In the framework of corpus linguistics, several studies (Sinclair 1991, 2004; Erman and Warren 2000; Kuiper 2009) have shown that all linguistic productions, oral or written, are half constituted of prefabricated sequences, following an "idiomaticity principle". Biber's corpus-driven studies on multi-word regular sequences (Biber 2009; Biber et al. 2004) also showed a high prevalence of various types of formulaic sequences in both oral and written corpora. Formulaic sequences may have multiple forms and involve different linguistic domains: the lexicon, syntax and discourse (cf. Legallois and Tutin 2013). Talking about the idiom-collocation principle, Partington (1998: 19) suggests that the use of prefabs should facilitate communication processing for the speaker as well as for the hearer. However, what seems obvious — at least at first glance — for oral communication may not hold for written communication, since in the latter process and product are clearly separated both materially and chronologically. According to Biber (2009), lexical bundles in writing, such as the construction *in the light of,* usually serve to bridge pairs of phrases, and are open-choice oriented on their right border. From a cognitive point of view, bursts may be considered to function in a similar way, with the writer having to pause in order to choose the contextually relevant development. However, to date, little is known about the linguistic structure of the sequences of text that are produced during bursts; although automatized and routine-like, they are not linguistically predictable. In fact, we found only one study (Kaufer et al. 1986, mentioned above) that has analyzed the linguistic structures of bursts, and no fine-grained linguistic description of bursts is available.

We attempted to verify the routine-status of bursts by comparing their contents to the contents of repeated segments. The repetition principle of RSs leads to the hypothesis of a routinization of discourse, as defined by Wray (2002: 9): sequences of words or other units that seem to be prefabricated, memorized and reproduced "as is" in the text, and not generated ad hoc. Along the same lines, Mayaffre (2007: 10) suggests that repeated segments of significant length are "linguistic tunnels where the creativity of the speaker/writer is reduced in favor of a kind of recitation [our translation]". Some recent studies test these strong hypotheses of routinization. Thus, Conklin and Schmitt (2008) used cloze contexts to show that prefabricated units are automatized in reading processing, for instance, and Blumenthal-Dramé (2013) showed through behavioral methods that high log-relative frequency items are better memorized than low-frequency

Table 1: General features of the real-time recorded corpus.

	Reports									
	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
Number of sessions	13	5	13	8	9	8	4	9	9	2
Number of words in the final reports	3223	7400	3144	3356	2841	2878	892	1239	2776	1372
Number of repeated segments	796	628	1432	934	983	1060	113	110	754	436
Number of bursts of writing	1092	441	373	949	868	563	493	394	474	62

items. But these studies provide no evidence about real-time writing processing: are repeated segments and other formulas "automatically" reproduced? By contrasting bursts of writing and repeated segments, we wanted to test the prefabs hypothesis, and to verify whether linguistic co-occurrence is spontaneously generated. The general features of the reports that constitute our corpus are presented in Table 1 above.

Our results indicate that cognitive routinization does not seem to overlap discourse routinization, since preliminary corpus analysis showed that around 5% of the bursts and the repeated segments are identical. These results suggest that final stabilized sequences and texts are the product of linguistic dynamics that remain fully outside usual linguistic analysis.

The fact that a great number of repeated segments do not have an equivalent burst may reflect strategies of communicative adaptability (Mey 1998; Verschueren and Brisard 2009), which fits particularly well with the fact that most of the repeated segments did not emerge in the first drafts of writing (see Cislaru et al. 2013 on possessive repeated segments). This point may suggest an overlap between the use of linguistic prefabs and the shaping of text in order to conform to social norms, in accordance with genre and pragmatic competence, as proposed by Edmonds (2013).

4.2 What relationship between frequency and spontaneity?

In order to clarify the relationship between frequency and spontaneity and its role in genre configuration, we first analyzed the bursts and repeated segments which share the same contents. Despite the low similarity ratio between the two categories, the nature of the shared sequences can provide interesting data about the relationship between text-in-progress and finished text.

The repetition principle is defining for repeated segments, and also underlying for language habitus and conventions. Thus, while repetition does not interfere with the identification of bursts of writing, it might influence their nature in accordance with language habitus schemas. The first question was to determine the place of frequency in the production of bursts of writing: Do bursts repeat during text processing of texts belonging to the same discourse genre? If some bursts appear at least twice in the corpus, are they connected to the repeated segments of the corpus in some way?

The bursts of writing that appeared more than once in our corpus are mostly simple graphemes (letters, or morphemes: *-e* for the feminine, *-s* for the plural), relational words (*de, le, dans,* etc.) or proper names (first names of the foster children). Some – rare – polylexical bursts that appeared two or three times in the same text hardly ever had a corresponding repeated segment (see Table 2). The parallel with the repeated segments is not relevant from a statistical point of view. Yet it is interesting to attempt to give a global description of repeated bursts of writing. Table 3 below groups bursts and repeated segments in series associating grammatical and semantic features.

The first column of Table 3 lists the repeated bursts of writing that have no corresponding repeated segment in the final text. The repeated bursts in Series A begin with a punctuation mark; this feature eliminates repeated segments. Series B contains three prefabricated types of constructions which denote i) social practices (*en milieu ouvert [non-formal {educational action}], les droits de visite et d'hébergement [the visiting or accommodation rights {of a parent who does not have custody}]*); ii) assessing criteria (*son manque de travail [his lack of work]*); iii) the discursive involvement of the writers (*nous observons*

Table 2: Number and percentage of bursts and their relationship to repeated segments (NB: b: Bursts; s: Repeated segments).

	Reports										
	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**	**Total**
Number of repeated bursts	44 4.0%	29 6.6%	20 5.4%	34 3.6%	27 3.1%	19 3.4%	27 5.5%	16 4.1%	20 4.2%	6 9.7%	242 4.2%
Number of identical bursts and repeated segments	30 2.7% b 3.8% s	11 2.5% b 1.8% s	11 2.9% b 0.7% s	31 3.3% b 3.3% s	34 3.9% b 3.5% s	14 2.5% b 1.3% s	11 2.2% b 9.7% s	8 2.0% b 7.3% s	10 2.1% b 1.3% s	3 4.8% b 2.1% s	163 2.9% b 2.2% s

Table 3: Recurrent bursts and bursts identical to repeated segments.

Bursts without corresponding repeated segments	**Bursts with corresponding repeated segments**	
Series A	*Series 1*	*Series 3*
, autant	. De plus,	cinq enfants
, etc.	En effet,	(12 ans)
. Aussi	à chaque fois	
. Puis	à ce sujet	*Series 4*
. toutefois	au quotidien	Madame ChXXX
d'emploi	en permanence	Monsieur GXXXX
« , un antipsychotique »		
Series B	*Series 2*	*Series 5*
Nous observons	il est	du groupe
Son manque de travail	FXXX a	de KXXX
En milieu ouvert	a été	
Les droits de visite et d'hébergement		

[we notice]). Each of these types of constructions may constitute a routine and thus underpin writing habitus and text schemas. However, the lack of the corresponding repeated segments in the final versions of the reports suggests that some of these segments, although apparently prefabricated, have been rewritten and modified in the final versions. These data tend to partly confirm the remarks made in 4.1. above.

The second column lists several categories of sequences shared both by repeated segments and repeated bursts : i) connectors (*en effet [in fact]*, *en permanence [permanently]*…); ii) proper name noun phrases (*Madame / Monsieur X*); iii) Subject+auxiliary bundles (*[Subject +] être [be]/avoir [have]*); iv) sequences directly related to the situation of the family being assessed (*cinq enfants [five children]*, *12 ans [12 years]*); v) one unit belonging to professional jargon (*du groupe [of the group]*, cf. Cislaru et al. 2013). Series 1 and 2 contain idiomatic sequences, as in i), iv) and v), or patterns, as in ii) and iii). Series 3 and 4 contain sequences learned by the writers during the educational measure, probably due to their frequency of use both in oral and written production; these sequences are based on specific patterns as well: *N years*, *N children*, etc.

According to the data above, patterns, or colligation schemas, constitute a relevant tool for the linguistic analysis of bursts of writing. We thus attempted to associate burst patterns to functions and meanings, in order to determine the way bursts of writing fit into the strategies of genre configuration.

5 Linguistic description of bursts of writing

5.1 Macro-patterns of bursts of writing in social reports

The classification of lexical strings — either RSs or bursts — demands complex criteria and a number of adjustments to the types of syntactic structures to which they are assigned. The main criterion retained is syntactic saturation, due to its analytical accessibility (see Figure 2). This means that we distinguish two large categories: saturated strings, which correspond to phrase-type constructions (noun phrases, prepositional phrases, clauses, sentences, etc.), and unsaturated strings, which correspond to syntactically irrelevant constructions and units that associate elements of two grammatical groups (phrases), such as *NP+preposition*, or that stop ahead of the boundary of a grammatical group, such as *Preposition+determiner*. The concept of unsaturated sequences is comparable to that of lexical bundles, as defined by Biber (2009). First, lexical bundles are by definition extremely common (in contrast to most idioms and many 'grammar patterns', which tend to be rare). Second, most lexical bundles are not idiomatic in meaning and not perceptually salient. For example, the meanings of bundles like *do you want to* or *I don't know what* are transparent from the individual words. And finally, lexical bundles usually do not represent a complete structural unit (Biber 2009: 283).

Table 4 below lists two series of unsaturated bursts that can easily combine in order to form sentences. It is worth noting that some of the bursts listed in the first column may be grammatically saturated, but unsaturated from a semantic and discursive point of view. For instance, due to genre constraints, the sequences

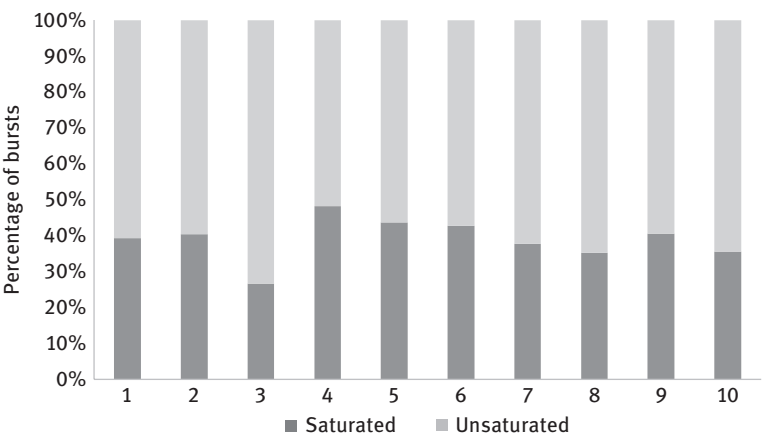


Figure 2: Percentage of saturated and unsaturated bursts in the 10 collected reports.

Table 4: Series of productive paradigms: lexical bundles and semantic *fillers*.

Unsaturated bursts	
Series 1	**Series 2**
Axel a pu dire que c'était	Authentique
Axel bénéficie d'une therapie	plus authentique
Axel ne veut pas perdre	, autant
Axel nous parait dans ces moments	beaucoup plus
Axel se prétend	Actuelle
Axel a diffrénte	ouvert
C'est un garçon	personnel
	personnel et
	distincts
	différent
	, non jugeante
	, accessible,
	'adulte
	Agressif
	Apparente
	assez inquiétantes
	collectif)
	(Axel;)

cited in the first column *Axel bénéficie d'une thérapie [Axel is being given therapy]* and *C'est un garcon [He's a boy]*, require specification: *c'est un garçon plutôt timide/turbulent/etc. [he's a shy/unruly boy]*. The second column lists various kinds of specifications.

The unsaturated bursts in Table 4 were produced autonomously; the units in Series 2 do not necessarily follow the sequences in Series 1 either chronologically or constructionally. However, the two series suggest an underlying schema of production: The bursts in Series 1 set up a construction that is unsaturated on the right side, where they leave an open slot while constraining the lexical choice syntactically, semantically and pragmatically. The bursts in the second series may fill the gap, and provide elements of description or assessment, since, in social reports, the assessment is based on specific elements of the situation. From this point of view, the bursts in the first column are rather neutral prefigured constructions, while the bursts in the second column denote context-dependent features and a more subjective stance. What seems to emerge from these data is the apparent necessity of a pause between neutral/objective sequences and subjective sequences such as appreciative modalities. Another hypothesis suggests a pause segmentation between thematic/topic sequences, which are semantically

prefigured (Series 1), and rhematic/focus sequences (Series 2), i.e. between given information and new information.

While some unsaturated constructions are markers of discontinuity that anticipate specific semantic fillers, others include patterns built around a pivotal reference point:

- Break before and after a full stop: ... *other children. She*...
- Break before/after/around coordination: *Alex shows some signs of sadness and/but* [he]...
- Break after concatenation between a saturated unit and a connector: *She decided to leave. Therefore*...
- Etc.

Such discontinuous constructions may highlight cases where the connection between facts and representations is pre-constructed, and only the discourse elements that are to be connected are selected from a list of possibilities. The writer of a social report may thus search for the appropriate formulae, first describing the individual situation of the young person they are monitoring, and then adapting the particular wording to social norms, in terms of both assessment and language choices.

5.2 Syntactic discontinuity and discourse cohesion: the case of coordinative connections

The last section of this article focuses on coordinative bursts of writing, mainly around two conjunctions, *et [and]* and *mais [but]*, which are the most frequent:

***et** la réaction*	**and** the reaction
***et** les résulta*	**and** the resul[ts]
*ces états d'agitation régulières **et** la vie quotidienne **et** ordinaire au sein d'une famille.*	these states of regular agitation **and** everyday **and** ordinary life in a family
*La présence d'animaux **et** d'espace*	the presence of animals **and** space
été chagriné ***mais** pas abattu*	been upset **but** not despondent
***mais** qu'il en était victime*	**but** that he was the victim of

It is difficult to decide whether these constructions are syntactically saturated or unsaturated. Most of them are clearly unsaturated, others contain saturated structures (phrases, clauses), but the presence of the coordinative conjunction, especially in constructions *X conjunction / conjunction X*, creates an unsaturated bond.

Table 5: Number and percentage of bursts containing *et [and]*.

	Reports										
	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**	**Total**
ET-bursts	68	50	20	79	57	50	27	18	24	0	393
	6.2%	11%	5.4%	8.3%	6.6%	8.9%	5.5%	4.6%	5.1%	–	
ET X	21	15	7	34	13	15	17	8	8	–	138
X ET	4	4	2	2	6	7	2	1	2	–	30
X ET Y	43	31	11	43	38	28	8	9	14	–	225

Table 6: Number and percentage of the different structures of bursts containing *mais [but]*.

	Type of structure			
	MAIS X	**X MAIS**	**X MAIS Y**	**Total**
Number	15	7	36	58
Percentage	25.9%	12.1%	62.1%	1%

5.2.1 Types and frequency of coordinative bursts

The coordinative conjunction *et [and]* is the most frequent: 393 bursts (6.9%) contain the conjunction *et [and]*. Compared to the 596 occurrences of *et [and]* in the final texts, about 66% of the *AND*-constructions are therefore spontaneously produced in complex polylexical bursts. Table 5 summarizes the proportion of bursts containing ET.

It is interesting to note that coordinative constructions are spontaneously produced in our corpus. This may suggest that syntactic invariants that are spontaneously produced in writing signal formal regularities of text-structure and also, probably, cognitive schemas underlying language production. Connectors condense information about discourse relations (Rossari 2000) and as such they provide processing instructions to the addressee to accurately integrate textual content by stating how to connect discourse units. For instance, the conjunction *and* marks a certain degree of compatibility between two opinions or referents (sentences or phrases), which can coexist in a text (Rousseau 2007: 33). *AND*-constructions function as colligational patterns (Hoey 2005; close to *motifs*, Longrée and Mellet 2013: 66) that regularly associate *n* units of the text and thus offer a collocational frame, able to receive a paradigm of fixed or variable units and liable to shape text structure and thus to support genre configuration and authentication.

The conjunction *et*, which is rather frequent, expresses various types of relations in French: addition, succession, chronology, opposition, consequence, etc. (cf. Riegel et al. 2004: 880; see also Charaudeau 1992: 503-504; Rousseau 2007). *Et* functions like an "archi-connector" (Bronckart and Schneuwly 1984; Schneuwly and Bronckart 1986); its additive value allows it to concatenate different types of units recursively (Rousseau 2007).

The conjunction *mais* plays an oppositive role (see Bruxelles et al. 1980). From a discursive point of view, *mais[but]* is a 'dialogical' marker, focusing on presupposed inferences of the addressee (see Mélis 2007). *Mais [but]*, which is less frequent in our corpus, occurs in 58 bursts; compared to the total number of occurrences in the final texts (68), this means that 85.3% of occurrences of *mais* are produced in complex bursts.

5.2.2 Coordinative constructions as discourse organizers

Connectors function as text organizers (Schneuwly 1997), and studying the way coordinative connections are fixed during text production may help us understand the psycholinguistic processes that underlie the syntactic relationships (Antoine 1996: 79). Recently, Sinclair and Mauranen (2006: 13) proposed to analyze coordinative conjunctions separately, and not to attach them to the clauses that they precede. Our processual data do not validate this suggestion. As shown in Tables 5 and 6, coordinative constructions may have three different structures, where the conjunction is necessarily attached to a preceding and/or following segment: *connector X, X connector, X connector Y*. As in oral conversations, where *et* is inserted in the linearity of the interaction, in order to maintain discourse continuity (Mouchon, Fayol and Gombert 1989, cited by Favart and Passerault 1999: 159), the occurrences of *et* in bursts represent cohesive relationships.

The binary constructions *X et/mais Y* are the most frequent in the corpus, representing as much as 57.25% of the total of AND-constructions, and 62.07% of the BUT-constructions. Binary constructions are a particularly interesting field of inquiry, in that they spontaneously process a preestablished relationship. According to Mélis (2007: 41), coordinative conjunctions label the addition of a focused content in an already complete schema. In binary constructions the focused content is processed simultaneously with the global scheme. Our hypothesis is that the nature and the contents of this relationship are related to genre competence.

About 81% of the 225 binary AND-constructions are syntactically rather symmetrical, relating constituents, although mostly non-saturated on their borders (Table 7).

Table 7: Constituents in *X* et *Y* bursts.

	Type of constituents					
	NP+et+NP N+et+N	**PP+et+PP**	**VP+et+VP V+et+V**	**Adj+et+Adj or equivalent**	**Clauses**	**Total**
Number	64	34	16	20	35	183
Percentage	34.7%	15.1%	7.1%	8.9%	15.6%	81.3%

Nearly 35% of these constructions coordinate nouns or noun phrases. From a grammatical point of view, in these nominal constructions *et* mainly links double-subjects and elements of complex prepositional phrases. From a denotative point of view, *et* links the names or statuses of the parents or other members of the family, or various experiences of the child able to contribute to the assessment of the situation or to state the social measure:

*Relations familiales avec sa mère **et** son parrain* (section title)
Family relationships with his mother **and** his godfather
*un désir de partager du temps avec sa sœur **et** le petit ami de celle-ci*
a desire to spend time with her sister **and** her sister's boyfriend
*la question du placement **et** l'éloignement*
the question of foster care **and** separation
*le souvenir de Mme Dos Santos **et** la maladie du grand-père*
the memory of Mme Dos Santos **and** the illness of her grand-father
*un week end sur deux **et** la moitié des vacances chez chacun de ses parents*.
Every other weekend **and** half of the holidays at each of the parents' place

Apart the coordination of names or statuses, the other cases of nominal coordination may be interpreted in terms of oriented accumulation: the difficulties or sensitive topics pile up (as in 'the memory of Mme Dos Santos **and** the illness of her grand-father', 'the question of foster care **and** separation'), as well as the steps of the social measure. The accumulation mechanism also concerns coordinated verbs and adjectives, which are generally elements of macro-constructions that relate elements of verbal phrases or attributes. The second coordinated segment often tends to specify the first one:

*En parallèle, la vie sociale de Fleur au collège s'est révélée de plus en plus complexe **et** conflictuelle. En effet,*
At the same time, Fleur's social life at secondary school appeared to be more and more complex **and** confrontational. Indeed,

Table 8: Constituents in *X mais Y* bursts.

	Type of structure					
	NP+mais+NP N+mais+N	PP+mais+PP	VP+mais+VP V+mais+V	Adj+mais+Adj or equivalent	Clauses	Total
Number	–	1	–	2	10	13
Percentage	–	2.8%	–	5.6%	27.8%	36.1%

*, la communication reste relativement sommaire **et** fragile*
, communication remains relatively basic **and** fragile

AND-coordinated verbal phrases or clauses often express a consecutive asymmetry, which is particularly in line with the evaluative purpose of the reports: *Les parents peuvent changer d'avis et demander son retour au domicile [The parents may change their mind and ask for him to return home]*.

Binary *X mais Y* bursts have a different status in discourse organization (see Table 8). Only 36% of the 36 binary BUT-constructions are syntactically rather symmetrical, relating constituents. In fact, *mais* links clauses more frequently (27.8%). There is no specific orientation in the relationship between clauses: the second clause may have either positive or negative polarity:

*[E]lle semble avoir beaucoup d'amis **mais** les relations sont très fluctuantes*
[S]he seems to have a lot of friends **but** the relationships are very unstable
Fleur souhaitait passer la totalité des vacances chez son père mais elle a pu entendre que nous sohaitions
Fleur wished to spend all the holidays at her father's place **but** she realized that we wished

However, in some particular reports, the orientation is massively negative, while in others polarities are rather balanced. In one report (D6), the negative polarity is correlated with a very high frequency of binary BUT-constructions, 14 occurrences out of 36 altogether.

In binary BUT-constructions, eleven occurrences of *mais* are associated with *aussi [also], également [equally], à la fois [at once], en même temps [meanwhile]* and thus acquire a strengthening value. Five other occurrences moderate the preceding assertion, as in *se montre plutôt indifférent mais pas attaquant [appears indifferent but not aggressive]*. *Mais* is used in these sixteen sequences to adjust the assessment with respect either to the judge or to the family.

Finally, in the context of social reports, which aim at analyzing a specific situation in the frame of social conventions and rules, a relevant distinction is the one between constructions that refer to the particular case (the situation and behavior of the child, family, etc.) on the one hand, and constructions denoting social norms, genre clichés, or evaluation keys/grids. The aim is then to identify the nature of the contents that are spontaneously produced in bursts. However, it is not always easy to distinguish between the two denotative domains, because the description of the particular situation is sometimes formulated with respect to stereotypical features. The construction *Les mises en danger ont été plus régulière et ont suscité beaucoup d'interrogation [At-risk situations have become more frequent and have raised numerous questions]* denotes the particular situation of a child using genre-specific formulations. Likewise, constructions such as *poser problème et envahir* [*create difficulties and overwhelm*] or *parcours scolaire émaillé de ruptures et donc de lacunes importantes* [*severely interrupted schooling and consequently significant gaps*] denote shared knowledge of social practices. Finally, generic social features are used to characterize and assess specific situations. Routinization, and more specifically genre routinization, appears to be spontaneously processed during writing.

6 Conclusion

We attempted in this study a longitudinal approach to genre analysis, exploiting real-time processed text. We thus were able to compare spontaneously produced units of text, the bursts of writing, to sequences found in finished texts, and pinpoint some strategies of text organization and genre adaptation. We have no evidence that the linguistic analysis of bursts allows text classifications and genre elucidation, and this was not the expected result of the present study. Rather, we wanted to underline some linguistic patterns that jointly contribute to ensuring the successfulness of the communicative purpose of the reports, so as to account for their performativity. This could offer new perspectives on the articulation of form and meaning in genre configuration.

The real-time data we analyzed showed that lexical strings that are repeated in finished written discourse are generally not produced as blocks or bundles. The study of these bursts of production indeed shows that a very limited number (5%) of the lexical bundles found in the finished text are produced as such. More often than not, the social workers spontaneously produce specific non-reiterated strings. This means that the strings usually called prefabs in the literature and considered as memorized formulae are not part of a genre or discursive stock, but

result at least partly from adaptation strategies. This suggests an overlap between the use of linguistic prefabs and the shaping of text in order to adapt language performance to social norms.

The boundaries between bursts of writing trace regularly although not systematically the frontiers between prefabricated, given, topical, neutral or objectivizing units on the one hand and new, focal, subjective, appreciative units on the other hand. This might be a genre-specific feature of social reports, which comply with the evaluative communicative purpose and thus constantly exploit generic categories and social norms to describe and assess particular cases.

Some syntactically complex patterns such as coordination are produced as bursts. These are performance patterns (actually produced units, in Hymes' theory) specific to the genre of social reports, but not exclusively, and to the type of discourse. Indeed, we examined some expository and descriptive texts that were collected in the context of another study, and found a very limited number of coordinative bursts. Coordination patterns semiotize a junction or a bifurcation in the textual flow. In our corpus of social reports, binary coordinative bursts around *et [and]* and *mais [but]* are the most frequent. Through the linearity of spontaneous production, they articulate an expected focused unit to a previous topical unit. The spontaneity of the additive and oppositive connections signals preformatted textual structures as well as relational contents.

References

Adam, Jean-Michel. 2011a. *Genres de récits. Narrativité et généricité des textes*. Louvain-la-Neuve – Paris: Academia – L'Harmattan.

Adam, Jean-Michel. 2011b [1992]. *Les textes types et prototypes. Récit, description, argumentation, explication et dialogue*. Paris: Armand Colin.

Alves Rui. A., Sao Luis Castro, Luisa Sousa, Sven Strömqvist. 2007. Typing skill and pause-execution cycles in written composition. In Mark Torrance, Luuk Van Waes, & David Galbraith (eds), *Writing and cognition, research and applications*, 55–65. Dordrecht: Elsevier Sciences Publishers.

Alves, Rui A., Marta Branco, Sao Luis Castro & Thierry Olive. 2011. Children of high transcription skill compose using bigger language bursts. In Virginia W. Berninger (ed), *Past, Present, and Future Contributions of Cognitive Writing Research to Cognitive Psychology*, 389–402. New York: Psychology Press.

Antoine, Gérald. (1996 [1958-1962]). *La coordination en français*. Paris: Editions d'Artrey.

Askehave, Inger. 1999. Communicative Purpose as Genre Determinant. *Hermes – Journal of Linguistics* 23. 13–23.

Baaijen, Veerle M., David Galbraith & Kees de Glopper. 2012. Keystroke Analysis: Reflections on Procedures and Measures. *Written Communication* 29(3). 246–277.

Bakhtine, Mikhail. 1984. *Esthétique de la création verbale*. Paris: Gallimard.

Bax, Stephen. 2011. *Discourse and Genre. Analysing Language in Context*. Houndmills: Palgrave Macmillan.

Beghtol, Clare. 2001. The concept of genre and its characteristics. *Bulletin of The American society for Information Science and Technology,* 27(2). http://www.asis.org/Bulletin/Dec-01/beghtol.html

Bhatia, Vijay K. 1993. *Analysing Genre. Language Use in Professional Settings*. London/New York: Longman.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English. Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3). 275–311.

Biber, Douglas. 2014. "The ubiquitous oral versus literate dimension: a survey of multidimensional studies". In Jeffrey Connor-Linton & Luke Wander Amoroso (eds), *Measured Language. Quantitative Approaches to Acquisition, Assessment, and Variation*, 1–19. Washington: Georgetown University Press.

Biber, Douglas, Susan Conrad & Viviana Cortes. 2004. If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3). 371–405.

Biber, Douglas, Ulla Connor & Thomas A. Upton (eds). 2007. *Discourse on the Move. Using Corpus Analysis to Describe Discourse Structure*. Amsterdam – Philadelphia: John Benjamins.

Blumenthal-Dramé, Alice. 2013. *Entrenchment in Usage-Based Theories. What Corpus Data Do and Do Not Reveal About the Mind.* Berlin: De Gruyter Mouton.

Branca-Rosoff, Sonia. 1999. Types, modes et genres : entre langue et discours. *Langage et Société* 87. 5–23.

Branca-Rosoff, Sonia, & Valérie Torre. 1993. Observer et aider : l'écrit des assistantes sociales dans les Demandes d'intervention. *Recherches sur le français parlé* 12. 115–135.

Brazil, David. 1995. *A Grammar of Speech*. Oxford: Oxford University Press.

Bronckart, Jean-Paul & Bernard Schneuwly. 1984. La production des organisateurs textuels chez l'enfant. In Michel Moscato & Gilberte Pieraut-Le Bonniec (eds), *Le Langage : construction et actualisation,* 165–178. Rouen: Presses Universitaire de Rouen.

Bruxelles, Sylvie, Anne-Marie Diller, Oswald Ducrot, Eric Fouquier, Jean Gouazé, & Anna Rémis. 1980. Mais occupe-toi d'Amélie. In Oswald Ducrot et al. (eds), *Les mots du discours*, 93–130. Paris: Editions de Minuit.

Charaudeau, Patrick. 1992. *Grammaire du sens et de l'expression*. Paris: Hachette.

Chenoweth, Anne & John R. Hayes. 2001. Fluency in writing. *Written Communication* 18. 80–98.

Chenoweth, Anne & John R. Hayes. 2003. The inner voice in writing. *Written Communication* 20. 99–118.

Chenu, Florence, François Pellegrino, Harrient Jisa & Michel Fayol. 2014. Interword and intraword pause threshold in the writing of texts by children and adolescents: A methodological approach. *Frontiers in Psychology 5*. 182.

Cislaru, Georgeta, Frédéric Pugnière-Saavedra & Frédérique Sitri (éds). 2008. *Analyse du discours et demande sociale : le cas des écrits de signalement. Les Carnets du Cediscor* 10. Paris : Presses Sorbonne nouvelle.

Cislaru, Georgeta, Frédérique Sitri & Frédéric Pugnière-Saavedra. 2013. Figement et configuration textuelle : les segments de discours répétés dans les rapports éducatifs. In Catherine Bolly, and Liesbeth Degand (eds), *Across the Line of Speech and Writing Variation. Corpora and Language in Use,* 165–183. Louvain-la-Neuve: Presses universitaires de Louvain.

Conklin, Kathy & Norbert Schmitt. 2008. Formulaic sequences: are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics* 29. 72–89.

Delcambre, Pierre. 1997. *Écriture et communications de travail: Pratiques d'écriture des éducateurs spécialisés*. Lille : Presses du Septentrion.

Edmonds, Amanda. 2013. Une approche psycholinguistique des phénomènes phraséologiques: le cas des expressions conventionnelles. *Langages* 189. 121–138.

Erman, Britt & Beatrice Warren. 2000. The idiom-principle and the open-choice principle. *Text* 20. 29–62.

Favart, Monik & Jean-Michel Passerault. 1999. Aspects textuels du fonctionnement et du développement des connecteurs: approche en production. *L'Année Psychologique 99*. 149–173.

Flower, Linda. 1979. Writer-based prose: a cognitive basis for problems in writing. *College English* 41(1). 19–37.

Grésillon, Almuth & Jean-Louis Lebrave. 2008. Linguistique et génétique des textes : un décalogue. *Le français moderne [Special issue Tendances actuelles en linguistique française]*. 37–49.

Hayes, John R. & Anne N. Chenoweth 2006. Is working memory involved in the transcribing and editing of texts? *Written Communication* 23. 135–149.

Hoey, Michael. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Huyette, Michel & Philippe Desloges. 2009. *Guide de la production judiciaire de l'enfant*. Paris : Dunod.

Hymes, Dell. 1984. *Vers la compétence de communication*. ENS Saint-Cloud: Hatier-Credif.

Kaufer, David S., John R. Hayes & Linda Flower. 1986. Composing written sentences. *Research in the Teaching of English* 20. 121–140.

Kuiper, Koenraad. 2009. *Formulaic Genres*. New York: Palgrave Macmillan.

Lafon, Pierre & André Salem. 1983. L'inventaire des segments répétés d'un texte. *Mots* 6. 161–177.

Lardilleux, Adrien, Serge Fleury & Georgeta Cislaru. 2013. Allongos: Longitudinal Alignment for the Genetic Study of Writers' Drafts. *Computational Linguistics and Intelligent Text Processing*, *Springer LNCS* 7817. 537–548.

Legallois, Dominique & Agnès Tutin (eds). 2013. Vers une extension du domaine de la phraséologie [special issue]. *Langages* 189.

Leijten, Marielle & Luuk Van Waes. 2006. Inputlog: New Perspectives on the Logging of On-Line Writing. In Gert Rijlaarsdam (Series ed.) and Kirk Sullivan & Eva Lindgren (Vol. eds), *Studies in Writing, Volume 18 – Computer Keystroke Logging and Writing: Methods and applications*, 73–94. Oxford: Elsevier.

Léglise, Isabelle. 2004. *Pratiques, langues et discours dans le travail social.* Paris: L'Harmattan.

Longrée, Dominique & Sylvie Mellet. 2013. Le motif : une unité phraséologique englobante ? Etendre le champ de la phraséologie de la langue au discours. *Langages* 189. 65–79.

Maingueneau, Dominique. 2004. Retour sur une catégorie : le genre. In Jean-Michel Adam, Jean-Blaise Grize and Magid Ali Bouacha (eds), *Texte et discours : catégories pour l'analyse*, 107–118. Dijon: Editions Universitaires de Dijon.

Maingueneau, Dominique. 2014 (rééd.), *Discours et analyse du discours*. Paris: Armand Colin.

Mayaffre, Damon. 2007. L'analyse de données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan. Retour sur les travaux actuels de topographie/topologie textuelle (partie I). *Lexicometrica* 9. http://lexicometrica.univ-paris3.fr/numspeciaux/special9/mayaffre.pdf. (2 March 2012).

Mélis, Gérard. 2007. La coordination inter-propositionnelle. In André Rousseau, Louis Begioni, Nigel Quayle & Daniel Roulland (eds), *La coordination*, 141–150. Rennes: Presses universitaires de Rennes.

Mey, Jacob. 1998. Adaptability. In Jacob Mey (ed.), *Concise Encyclopedia of Pragmatics*, 5–7. Oxford: Elsevier.

Miller, Carolyn. 1984. Genre as social action. *Quarterly Journal of Speech* 70. 151–167.

Mouchon, Serge, Michel Fayol & Jean-Emile Gombert 1989. L'utilisation des connecteurs dans les rappels de récits chez les enfants de 5 à 8 ans, *L'Année Psychologique* 89. 513–529.

Née Emilie, Frédérique Sitri & Marie Veniard. 2014. Pour une approche des routines discursives dans les écrits professionnels. Communication au *Congrès Mondial de Linguistique Française*, 19–23 juillet, Berlin.

Olive, Thierry & Georgeta Cislaru. 2015. Linguistic forms at the process-product interface: Analyzing the linguistic content of bursts of production. In Georgeta Cislaru (ed.), *Writing(s) at the Crossroads: the Process-Product Interface*, 99–123. Amsterdam – Philadelphia: John Benjamins.

Partington, Alan. 1998. *Patterns and Meanings*. Amsterdam – Philadelphia: John Benjamins.

Plane, Sylvie, Thierry Olive & Denis Alamargot (eds). 2010. Traitement des contraintes de la production d'écrits: Aspects linguistiques et psycholinguistiques [special issue]. *Langages* 177.

Pugnière-Saavedra, Frédéric. 2008. Quelques régularités des écrits du signalement. *Les Carnets du Cediscor* 10. 21–36.

Rastier, François. 1989. *Sens et textualité*. Paris : Hachette.

Riegel, Martin, Jean-Christophe Pellat & René Rioul. 2004 [1994]. *Grammaire méthodique du français*. Paris : Presses universitaires de France.

Rossari, Corinne. 2000. *Connecteurs et relations de discours : des liens entre cognition et signification*. Nancy: Presses Universitaires de Nancy.

Rousseau, André. 2007. La coordination : approche méthodologique, critique et raisonnée des questions essentielles. In André Rousseau, Louis Begioni, Nigel Quayle & Daniel Roulland (eds), *La coordination,* 18–57. Rennes: Presses universitaires de Rennes.

Scardamalia, Marlene & Carl Bereiter. 1991. Literate expertise. In K. Anders Ericsson & Jacqui Smith (eds), *Toward a general theory of expertise: Prospects and limits,* 172–194. Cambridge, MA: Cambridge University Press.

Schneuwly, Bernard. 1997. Textual organizers and text types. Ontogenetic aspects on writing. In Jean Costermans & Michel Fayol (eds), *Processing Interclausal Relationships: Studies in the Production and Comprehension of Text*, 245–263. Mahwah, NJ: Lawrence Erlbaum.

Schneuwly, Bernard & Jean-Pierre Bronckart. 1986. Connexion et cohésion dans quatre types de textes d'enfants. *Cahiers de linguistique française* 7. 279–294.

Sinclair, John. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, John. 2004. *Trust the Text. Language, corpus and discourse*. London – New York: Routledge.

Sinclair, John & Anna Mauranen. 2006. *Linear Unit Grammar: Integrating Speech and Writing*. Amsterdam –Philadelphia: John Benjamins.

Sitri, Frédérique. 2013. *Il peut exprimer ses difficultés* : une lecture "évènementielle" de *pouvoir* dans des rapports de travailleurs sociaux. In Sophie Moirand, Sandrine Reboul-Touré, Danielle Londei & Licia Reggiani (eds), *Dire l'événement. Langage, mémoire, société*, 73–83. Paris, Presses de la Sorbonne nouvelle.

Spelman Miller, Kriss & Kirk P. H. Sullivan. 2006. Keystroke logging: An introduction. In Kirk P. H. Sullivan & Eva Lindgren (eds.), *Computer Keystroke Logging and Writing: Methods and Applications*, 19. Amsterdam: Elsevier.

Swales, John. 1990. *Genre Analysis*. Cambridge, MA: Cambridge University Press.

Verschueren, Jeff & Franck Brisard. 2009. Adaptability. In Jeff Verschueren & Jan-Ola Östman (eds), *Key Notions for Pragmatics*, 28–47. Amsterdam – Philadelphia: John Benjamins.

Wengelin, Åsa. 2014. *Temps* et *pause* dans l'écriture au clavier. In Christophe Leblay & Gilles Caporossi (eds), *Temps de l'écriture. Enregistrements et représentations,* 97–124. Paris: L'Harmattan.

Widdowson, Henry G. 1979. *The Description of Scientific Language*. *Explorations in Applied Linguistics*. Oxford: Oxford University Press.

Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Index

https://doi.org/10.1515/9783110595864-011