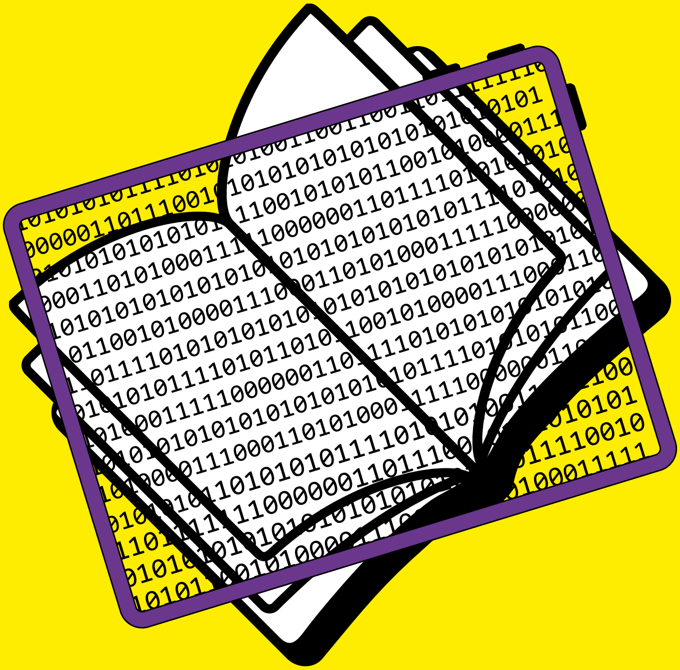


Digital Stylistics in Romance Studies and Beyond

Edited by
Robert Hesselbach
José Calvo Tello
Ulrike Henny-Krahmer
Christof Schöch
Daniel Schlör

HEIDELBERG
UNIVERSITY PUBLISHING

Digital Stylistics in Romance Studies and Beyond

Robert Hesselbach, José Calvo Tello, Ulrike Henny-Krahmer,
Christof Schöch, Daniel Schlör (Eds.)

Digital Stylistics in Romance Studies and Beyond

HEIDELBERG
UNIVERSITY PUBLISHING

ORCID®
Robert Hesselbach https://orcid.org/0000-0001-9758-8290
José Calvo Tello https://orcid.org/0000-0002-1129-5604
Ulrike Henny-Krahmer https://orcid.org/0000-0003-2852-065X
Christof Schöch https://orcid.org/0000-0002-4557-2753
Daniel Schlör https://orcid.org/0009-0001-6983-3719

The conference documented by this publication was organized by the early-career research group “Computational Literary Genre Stylistics” (CLiGS) funded by the German Ministry for Education and Research (BMBF) from April 2014 to March 2022 (funding identifiers 01UG1408 and 01UG1508). The present open access publication was enabled by funding from the Chair for Digital Humanities at Trier University.

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

Published by Heidelberg University Publishing (heiUP), 2024

Heidelberg University/Heidelberg University Library
Heidelberg University Publishing (heiUP)
Grabengasse 1, 69117 Heidelberg, Germany
https://heiup.uni-heidelberg.de

The electronic open access version of this work is permanently available on Heidelberg University Publishing’s website: https://heiup.uni-heidelberg.de
urn: urn:nbn:de:bsz:16-heiup-book-1157-2
doi: https://doi.org/10.17885/heiup.1157

Cover illustration: https://www.vecteezy.com/free-vector/open-book

ISBN 978-3-96822-201-1 (Hardcover)
ISBN 978-3-96822-200-4 (PDF)

Table of Contents

Introduction

Robert Hesselbach, José Calvo Tello,
Ulrike Henny-Krahmer, Christof Schöch, Daniel Schlör

1. A New Subdiscipline: *Digital Stylistics*

The increasing access to massive amounts of digital texts in the last decades has allowed different disciplines to emerge or flourish. Disciplines from the humanities (linguistics, literature, history, and philosophy, among others) and computer science (natural language processing, computational linguistics, information retrieval) have created and made use of a new generation of resources, for example corpora, websites, digitized collections, annotation tools, and methods of analysis. Although the disciplines share many interests and methods, there are also specificities about their research objects and the research questions they tackle.

Stylistics is one of the traditional disciplines in the humanities that has benefited from these new resources. The main focus of this discipline is investigating the linguistic style of texts, and literary texts are the object of analysis more frequently than in many other linguistic subdisciplines. In this way, for decades, stylistics has been an interdisciplinary field where linguists and literary scholars have engaged in fruitful discussions.

In the past years, several projects and working groups have coined the concept of *digital stylistics* differently. Digital stylistics combines two previous research fields: digital humanities and stylistics. Digital stylistics is more specific than digital humanities, probably because the researchers needed to use a narrower label to define their community, research objects, and interests since the digital humanities community as a whole has become considerably larger and more diverse in the past decade. However, digital stylistics remains an interdisciplinary field, similar to both digital humanities and stylistics themselves, and it helps to challenge, confirm, or correct existing research paradigms using digital methods.

Moreover, in the past years, several disciplines related to computer science (data mining, machine learning, computational linguistics, artificial intelligence) have influenced and shaped several new subdisciplines in the digital humanities. This influence is visible in many articles in this book, in which methods from computer science are applied to humanities data. The explorative perspective has been widely applied in the past years, for example with the creation of character networks or the use of topic modeling, clustering techniques, or word embeddings. In many cases, this explorative

Hesselbach, Robert, et al.: Introduction, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 1–13.
DOI: https://doi.org/10.17885/heiup.1157.c19361

use is combined with a proper evaluation of the options and parameters that the methods offer. A further possibility is the proposal of new theoretical models, which can be expressed through computational means to structure specific phenomena from the humanities, or the development of new computational methods to investigate specific humanities research questions that could not be answered with previous methods.

Regarding the research object, categories of texts are frequently described by metadata and linguistic features. With categories of texts, we are referring to groups of texts defined, for example, by a common author, the genre, the year of composition or publication, the complexity of the text, or the degree of canonization or literariness. The linguistic features can be extracted from different linguistic levels: phonological, graphemic, morphological, lexical, semantic, syntactic, and pragmatic. In addition, content-related features can be taken into account, for example, the place or time span of the plot. This kind of study of textual categories or groups of texts through linguistic and other types of textual features can be seen in many chapters of this volume. To mention two examples, Laura Hernández-Lorenzo uses lexical information to study two textual categories: authorship and periodization. On his side, Andreas van Cranenburgh looks at specific morphosyntactic units in Dutch to better understand their impact on the perception of literariness.

2. The Founding of *Digital Stylistics* and Its Recent Development

Another evidence for the interdisciplinarity of digital stylistics can be seen in its pioneers, who started analyzing stylistic features with computational means in the 1980s. One of the most frequently cited authors in articles about the style and linguistic particularities of corpora (not necessarily literary ones) is Douglas Biber. His pioneering work has shaped not only the understanding of corpus linguistics but also many other subdisciplines of digital humanities which mainly work with texts. In their article "Exploring Fictional Styles along Universal Dimensions of Register Variation," Douglas Biber and Jesse Egbert lay out the principles of multi-dimensional analysis and present the results of such an analysis for English across discourse domains. The authors can show how two dimensions, the fundamental opposition between clausal versus phrasal discourse and the opposition between narrative versus non-narrative discourse, play an important role in fictional literature.

On the side of literary texts, John Burrows' work was an early precursor of the application of computational methods to literature. Digital humanities and, more specifically, the areas of stylometry and digital stylistics are in great debt to this pioneering scholar. Not only did he employ many methods unknown to the humanities to this point (such as principal component analysis), but he also developed specific new methods himself (Zeta, Delta, Iota) for a range of different research questions.

Since the 1990s, the application of computational and quantitative methods to texts to analyze stylistic features has developed in several directions. A lot of stylistic analysis has been done in corpus linguistics, a subdiscipline that is rooted in linguistics, even though literary texts are also analyzed from time to time. For many years, literary studies did not develop their own subdiscipline for the analysis of corpora, their own *corpus literary studies*, a label that could have been possible but is not a reality in the academic panorama. From the 2000s onwards, several new labels were coined, and one gained momentum during the following years: *digital humanities*, referring to a very broad area of which computational stylistic studies are only one specific part. However, this is not the only label that appears: *corpus stylistics*, *digital stylistics*, and more recently, *computational humanities* and *computational literary studies*. The latter terms focus on the application of computational methods for analysis rather than the representation and presentation of data. The adjective *computational* also emphasizes the importance of quantitative methods. These new labels for subfields are a sign that the very general interdisciplinary field of digital humanities is branching out again.

In order to situate the topic of this publication, *digital stylistics*, in the wider context of different (sub)disciplines related to the digital and the humanities, we ran a series of queries in a library catalog to quantify and visualize its results for the different labels of disciplines. This catalog is called the K10plus database and constitutes the largest database of library records in Germany. It contains information about monographs (both printed and digital) and journals, but not about chapters, journal papers, or conference contributions.

The queries retrieve the number of publications that contain the labels *humanities computing*, *corpus linguistics*, *digital humanities*, *digital philology*, *corpus stylistics*, *computational humanities*, *computer philology*, *stylometry*, *corpus literary studies*, *cultural analytics*, *digital stylistics*, and *textometry* in their title. To cover the terms in the two most predominant languages of the K10plus, for the searches, the terms were used in both German and English, and the results for the translations were summed up. The search is defined to also find publications that contain all the tokens of the label in any order. That means a book with the title *A Corpus Study of Literary Works* would be found for the label *corpus literary studies* since the title contains the three words from the label. We also ran exact searches for each label, giving a similar distribution of results but with around a fifth of the number of hits for each label, and many labels with no exact hits.

As Figure 1 shows, *corpus linguistics* and *digital humanities* are the terms that appear most frequently in the titles of the publications, in more than a thousand monographs. *humanities computing* follows these, with more than a hundred monographs.

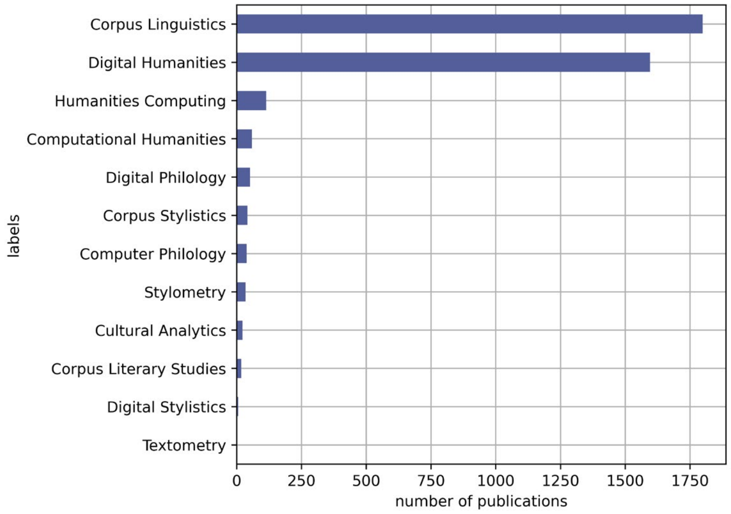


Fig. 1 Frequency of discipline labels in the titles of publications in the database K10plus (Hesselbach, Calvo Tello, Henny-Krahmer, Schöch, Schlör, CC BY).

Next, a group of labels with around forty publications is formed by *digital philology*, *corpus stylistics*, *computational humanities*, and *computer philology*. The following group has around twenty publications populated by labels such as *stylometry*, *corpus literary studies*, and *cultural analytics*. Finally, *digital stylistics* and *textometry* have fewer than ten publications. The results show that the term *digital stylistics*, which is part of the title of the present volume, is not the most widespread one. It addresses a quite specific field of study and community. At the same time, by history, meaning, and research culture, it is related in different ways to all the other terms that were queried and with which the articles published here can be associated, as well, to different degrees.

Given the data presented in Figure 1, one may ask: How has the frequency of these labels developed over time? When did *digital humanities* and *corpus linguistics* gain momentum? Is their frequency still increasing, or have they already reached their peak? To answer this, we ran queries in the same database extracting the frequency of each label from the last 40 years. For better visualization, only the labels with more than 20 publications in total are considered in this second figure.

Figure 2 shows how the number of publications with the label *corpus linguistics* in their title was increasing solidly until the year 2015, with over 100 publications per year. Since then, it seems that the number of titles has been rather stable, with around 90 monographs.

The label *digital humanities* appeared in the late 1990s and showed a steep development in its frequency a decade later. In 2010, only 20 books had this label in their title, but the number increased rapidly during the following years. It appears to peak in the year 2016 with 185 publications and its frequency seems to have stabilized since then with around 160 monographs. The rest of the labels show differences in their chronological frequency. An interesting group of labels is *humanities computing* and *computational humanities*. While the first is one of the most frequent labels up to 2000, the second shows an increase in the number of publications since the 2010s (Biemann et al. 2014), especially since 2020. This return to terms emphasizing the computational in contrast to the digital was predicted by Hockey (2012).

Regarding the labels with fewer publications, *stylometry* shows a similar chronological distribution, with some publications in all analyzed decades. *Corpus stylistics* seems to appear during the decade of 2000, with a few publications each year since then. *Computer philology* starts in the 1990s and keeps appearing in the title of some publications from then on. The labels *corpus stylistics*, *digital philology*, and *corpus literary studies* appear mainly in the twenty-first century. Finally, *cultural analytics* only appeared in the last years of the 2010s. An interesting outcome is that no label appears to have been abandoned completely by the community, as all the analyzed labels keep being used for the titles.

What are the differences between all these subdisciplines? All these labels combine two elements: On the one side, a reference to their main methodological framework:

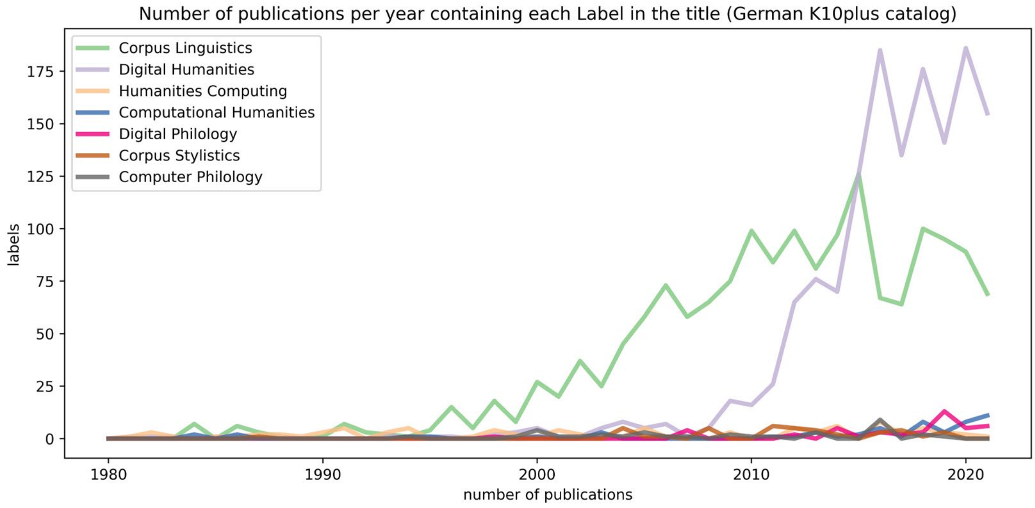


Fig. 2 Frequency of labels in the titles per year in the database K10plus (Hesselbach, Calvo Tello, Henny-Krahmer, Schöch, Schlör, CC BY).

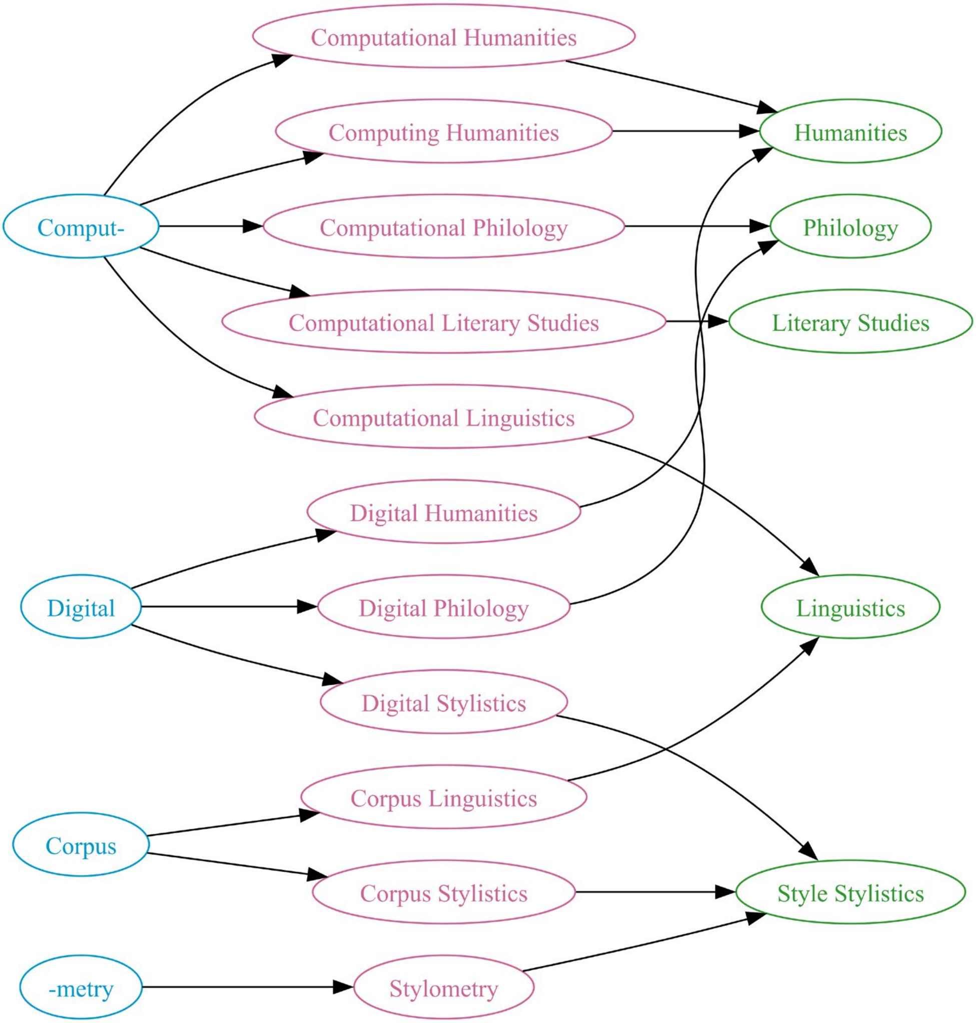


Fig. 3 Combination of the elements of the labels of the subdisciplines (Hesselbach, Calvo Tello, Henny-Krahmer, Schöch, Schlör, CC BY).

corpus, digital, the root *comput-*, or the Greek-based suffix *-metry*. On the other side, there is always a link to a particular discipline from the humanities: linguistics, philology, humanities, literary studies, or stylistics.

By combining both elements, the subdisciplines mark their affinity but also their differences to the other fields. *Computational humanities*, for example, is connected to *digital humanities* and to *computational linguistics*. In that form, the label could be read as a message: our area of study is similar to the digital humanities, however, our methods are more similar to computational linguistics. As this volume shows in its different chapters, digital stylistics shares interests with corpus stylistics and stylometry, applying diverse methods from digital humanities. However, the research field is narrower than that of digital humanities in general. On the other hand, the repertoire of methods differs from the traditional ones in stylometry or corpus stylistics.

We doubt that all these (sub)disciplines and research fields can be understood as mutually exclusive classes. In many cases, they show a great deal of overlap, with scholars publishing or being active in journals, associations, and projects belonging to several of these areas.

3. *Digital Stylistics* in Romance Studies

Traditionally, the majority of the publications in which literature is analyzed with computational means have focused on English texts. The two pioneers mentioned before, for example, have worked more frequently with English texts than with texts in other languages. Some reasons for this fact are the lack of digitized material, state-of-the-art tools, or financial support for projects concerned with other languages. A further challenge is that, in many cases, the only possible method to analyze multilingual corpora is to first split the texts into monolingual corpora and then run parallel analyses. However, in the last decade, an increasing interest can be noted in working with languages other than English or working with several languages. Some examples of this interest are the versions of the tutorial platform *Programming Historian* in Spanish, French, and Portuguese, the *COST Action Distant Reading* for European Literary History with the *European Literary Text Collection* (ELTeC), the network *Multilingual DH* (http://multilingualdh.org/en/) or the special interest group of the Alliance of Digital Humanities Organizations, *Global Outlook::Digital Humanities*. As we will describe later, the *CLiGS* (*Computational Literary Genre Stylistics*) project at the University of Würzburg (Germany) and this volume are part of this increasing interest in analyzing languages other than English.

In the center of this volume are the Romance languages, a group of languages that can be seen as being in the middle range when looking at the available resources.

Often, there are fewer tools and digital linguistic resources, such as dictionaries or annotated data, for these languages than for English. In other cases, available resources have been derived or translated from English, leading to gaps and a cultural bias in them. In many cases, access to large collections of digital texts is more limited in Romance languages than, for example, in German, when resources such as *TextGrid* or the *Deutsches Textarchiv* are considered. However, some Romance languages have a privileged situation in comparison to others, as many articles in this volume do focus on analyzing corpora in Spanish, French, and Italian.

There is also a specificity of the study of Romance languages in the German-speaking area. In other countries the study programs and faculties are split into the national traditions (Spanish Studies, French Studies, etc.), while these are generally grouped together in Romance languages and literatures departments in the German-speaking area. However, in recent years, many digital humanities endeavors have been defined using national or linguistic frontiers. In the last decade, a series of associations have emerged in the Romance languages-speaking countries, such as the *Associazione per l'Informatica Umanistica e la Cultura Digitale*, the *Red de Humanidades Digitales*, *Humanistica*, *Humanidades Digitales Hispánicas*, the *Canadian Society for Digital Humanities–Société canadienne des humanités numériques*, just to mention a few from the Americas and Europe.

In the German Romance Studies Association (*Deutscher Romanistikverband*), a working group called Digitale Romanistik on digital humanities in Romance Studies has been active since 2014. During the main conference of Romance studies (*Romanistentag*) in 2021, a specific section dedicated to digital humanities took place for the first time, organized by Digitale Romanistik. Furthermore, in the last year, some interdisciplinary projects with a Romance studies background and a focus on computational methods have been funded, such as the *PhraseoRom* project (Fesenmeier and Novakova 2020), a cooperation between the Université Grenoble Alpes and the universities of Osnabrück, Erlangen-Nürnberg, and Bonn, the above-mentioned project *CLiGS* at the University of Würzburg, or the Mining and Modeling Text project at the Trier Center for Digital Humanities.

4. About this Anthology

The present publication is the outcome of the conference Digital Stylistics in Romance Studies and Beyond, which took place in 2019 at the University of Würzburg from February 27 to March 2. The aim of the conference was to foster research in digital stylistics in the Romance languages and literatures, and to provide an opportunity for international researchers in the field to share their work in order to strengthen the community of literary scholars and linguists working with these languages. However, the amendment “and Beyond” aimed to keep the conference open for other languages and literatures as well. As digital stylistics has a strong methodological focus and is empirically oriented, the cross-language perspective and exchange of results between different application areas are important to advance the field. This is the reason why in this book one can find contributions on French, Italian, and Spanish as well as other studies on German, English, and Dutch.

The conference was organized by the Early-Career Research Group *Computational Literary Genre Stylistics* (*CLiGS*), a project that was based at the Department for Digital Literary Studies at the University of Würzburg and funded by the German Ministry for Education and Research (BMBF) from April 2014 to March 2020. The overall aim of the CLiGS project was to create a convergence between recent methods for the quantitative analysis of literature, on the one hand, and fundamental research questions from genre theory and stylistics, on the other. Several subprojects were realized, focusing primarily on French classical theater as well as Spanish and Spanish-American novels from the nineteenth to the twentieth century (see Schöch 2017; Calvo Tello 2021; Henny-Krahmer 2023). The focus of the *CLiGS* project on Romance literatures and the investigation of literary genres with quantitative and stylistic methods formed the basis for the conference call and it also influenced the kinds of contributions that are published in this volume. All the research results presented in the articles of this book are based on the empirical, computational analysis of literary corpora chosen to analyze and compare either major genres or subgenres of a particular major genre (poetry, drama, prose).

By focusing on genres and subgenres for Spanish, **José Calvo Tello** analyzes in his article “Classification of Genres through 500 Years of Spanish Literature in CORDE” the development of numerous genres over several centuries with the diachronic corpus CORDE (*Corpus diacrónico del Español*) of the Spanish Royal Academy (*Real Academia Española*, RAE). Several classification tests analyze the key factors for each category, finding the length of the text to be a good predictor. A historical analysis furthermore suggests a certain stability of the genres over time.

The next section, devoted to the computational analysis of poetry, begins with a contribution by **Laura Hernández-Lorenzo** (Seville). The author works with Golden Age Spanish poetry and Fernando de Herrera’s poems to analyze Herrera’s role in the stylistic evolution from Renaissance to Baroque and to verify if the posthumous edition of his poetry, *Versos* (1619), is more Baroque, as some critics have suggested. Results point to the transitional role of Herrera’s work in general, with the detection of a more Baroque component in the *Versos* edition. Subsequently, **Nanette Rißler-Pipka** (Göttingen, now Bonn) deals in her contribution “Cross-Language Stylometry: Picasso’s Writings in Spanish and French” with the poetry of one of the most famous Spanish painters, Pablo Picasso. The hypothesis that Picasso’s writings and poetry are characterized by a unique style is tested with a Spanish and French corpus, and in comparison to contemporary writers, the difference and distinctiveness can be shown by cross-language analyses. After articles on French and Spanish, **Jan Rohden** (Bonn) concludes the poetry chapter with a contribution on Italian. In his article "Digital Approaches to Poetic Style: A Quantitative Stylistic Analysis of Italian Petrarchism," Rohden raises the question of to what extent digital methods can provide new impulses for research on Petrarchism. A quantitative stylometric analysis of a corpus of Italian love poetry is conducted to identify stylistically distinctive elements of Petrarchism.

The following two chapters focus on the analysis of Spanish and French drama, respectively, and starts with a contribution on the Spanish theater of the Golden Age by **Álvaro Cuéllar** (Vienna, now Barcelona). In his article "Stylometry and Spanish Golden Age Theater: An Evaluation of Authorship Attribution in a Control Group of One Hundred Undisputed Plays," the author presents a study on 100 Spanish Golden Age theater plays whose authorship is undisputed in order to evaluate which algorithms and which text length are effective for authorship attribution. In his study on French drama ("Repetitive Research: Spitzer and Racine"), **Christof Schöch** (Trier) attempts to retrace Leo Spitzer's (1887–1960) famous stylistic reading of the tragedies of French seventeenth-century author Jean Racine (1639–1699) using digital text collections and computational methods of analysis, not only revealing new insights into Racine's and the classical period's style but also serving to highlight the respective strengths and limitations of established and computational approaches to stylistic analysis.

The following articles all explore narrative literature, with a focus on the nineteenth century at the beginning of this section. In her article "Family Resemblance in Genre Stylistics: A Case Study with Nineteenth-Century Spanish-American Novels," **Ulrike Henny-Krahmer** (Rostock) applies the concept of family resemblance in a digital genre stylistics analysis of subgenres of nineteenth-century Spanish-American historical and sentimental novels, with a formal implementation of the concept and quantitative evaluation. Besides the use of digital methods as an approach to soft categorization, this analysis shows that the concept of family resemblance itself undergoes change. **Julian Schröter** (Würzburg, now Munich) discusses how procedures of computational and literary genre stylistics can be built and implemented in order to reconstruct the ways in which genres undergo historical change, finding that genre stylistics should, in general, be based on aesthetic interest. In his study on German literature ("Machine Learning as a Measure of the Conceptual Looseness of Disordered Genres: Studies on German *Novellen*"), Schröter outlines the specific historical situation of the German *Novelle* as well as the basis of an aesthetic historiography of this disordered genre. He suggests machine learning tasks be integrated into a psychological framework interpreting accuracy scores as a measure of the semantic looseness of the concept of a specific genre within historical, literary communities.

Turning to twentieth-century prose, **Douglas Biber** (Arizona) and **Jesse Egbert** (Arizona) present an overview of results of multi-dimensional analysis of English across discourse domains. In their article "Exploring Fictional Styles along Universal Dimensions of Register Variation," the authors show how the two universal dimensions, the fundamental opposition between clausal versus phrasal discourse and the opposition between narrative versus non-narrative discourse, are of great importance in fictional literature. **Andreas van Cranenburgh** (Groningen) analyzes strong and weak pronouns as a stylistic marker of literariness. In his study "Dutch Strong and Weak Pronouns as a Stylistic Marker of Literariness," he investigates the case of Dutch with a corpus of literary novels and presents quantitative as well as qualitative judgments. The results suggest that style is a prominent factor in the strong/weak pronoun distinction and that a high proportion of strong pronouns is associated with literary prestige and Dutch authorship. **Robert Hesselbach** (Erlangen-Nürnberg) explores in his contribution ("Investigating the Relation between Syntactic Complexity and Subgenre Distinction: A Case Study on Two Contemporary French Authors") the ways in which the analysis of the syntactic complexity of a narrative text's sentences can help distinguish between different literary genres for contemporary French novels (1979–2002). Based on an analysis of both qualitative as well as quantitative features, the author argues that syntactic complexity has very little influence on genre distinction and that the degree of syntactic complexity is more likely to appear as an author-related characteristic. The section concludes with a study ("Digital Stylistic Analysis in PhraseoRom: Methodological and Epistemological Issues in a Multidisciplinary Project") by **Clémence Jacquot** (Montpellier), **Ilaria Vidotto** (Grenoble), and **Laetitia Gonon** (Grenoble). The three authors analyze a large annotated corpus of novels (in French, English, and German) from the twentieth and twenty-first centuries. A stylistic annotation methodology of this corpus is proposed, which links the phraseological analysis of a large literary corpus together with stylistic issues concerning its formal and literary implications, through the concept of the motif.

The last two contributions in this anthology are concerned with Italian and Spanish literature of the twenty-first century, with **Katharina Dziuk Lameira** (Kassel) focusing on modern Spanish novels. In her article "Complexity and Style of Modern Spanish Literary Texts," the author discusses the question of whether text complexity can be seen as a dimension of authorial style and the relevance of linguistic features for the description of both aspects for modern Spanish literary texts. First analyses show which features and parameters could be suitable for the description of some of the novels analyzed. **Michele A. Cortelazzo** (Padova), **George K. Mikros** (Qatar), and **Arjuna Tuzzi** (Padova) conclude this anthology with their article "Applying General Impostors Method to the Ferrante Case." By focusing on the contemporary Italian author Elena Ferrante, the authors address Ferrante's authorship investigation as a verification problem to analyze whether the real author behind Ferrante's pseudonym is among the candidates considered in previous studies for a corpus of 150 novels written by 40 authors and a non-literary corpus of 113 texts signed by 14 different entities. In the literary corpus, Starnone emerged as the most likely author of Ferrante's novels; however, Starnone was not the only possible author since in many non-literary texts, Raja, Martone, as well as the E/O publishing house staff and publishers, seem to have made authorial contributions.

As all the articles in this book are concerned with the stylistic analysis of specific linguistic and literary research data, the research results presented here will only be fully transparent and reproducible if the underlying data is accessible. Therefore, the editors of this volume encouraged the authors to publish their research data in an open format, with an open license allowing for re-use and in a specific, identifiable version corresponding to the state that the data had at the time of writing the articles. To facilitate the publication of the research data, a community was created on Zenodo, a service provided by CERN for the long-term archiving and versioning of data, describing it with appropriate metadata and making it citable through digital object identifiers (DOIs) (Nielsen 2013). A Zenodo community serves to bundle different but related data publications. Several articles in this volume are accompanied by data or code records archived on Zenodo. Research data related to the other articles are, in part, offered elsewhere. The volume itself is also published in Open Access form.

Both the conference and the publication would not have been possible without institutional and personal support. We would therefore like to thank first and foremost the German Federal Ministry of Education and Research (*Bundesministerium für Bildung und Forschung*, BMBF) for its generous funding, which made the *CLiGS* project and thus also the Digital Stylistics in Romance Studies and Beyond conference possible in the first place. Furthermore, we have always been supported by Fotis Jannidis, professor for digital literary studies at the University of Würzburg, and his whole team and, therefore, we would like to express our special thanks to Prof. Fotis Jannidis, Dr. Steffen Pielström, and Barbara Gersitz. We would also like to thank our two student assistants, Lea Kunz and Constanze Ludewig, who ensured that the conference ran very smoothly. With this publication, we conclude the *CLiGS* project in a very satisfactory way and hope that the individual contributions will generate further research in the field of digital stylistics.

ORCID®

Robert Hesselbach https://orcid.org/0000-0001-9758-8290
José Calvo Tello https://orcid.org/0000-0002-1129-5604
Ulrike Henny-Krahmer https://orcid.org/0000-0003-2852-065X
Christof Schöch https://orcid.org/0000-0002-4557-2753
Daniel Schlör https://orcid.org/0009-0001-6983-3719

References

Biemann, Chris, Gregory R. Crane, Christiane D. Fellbaum, and Alexander Mehler. 2014. "Computational Humanities – Bridging the Gap Between Computer Science and Digital Humanities." *Dagstuhl Reports* 7: 80–111. https://drops.dagstuhl.de/opus/volltexte/2014/4792/.

Calvo Tello, José. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld: transcript. https://doi.org/10.14361/9783839459256.

Fesenmeier, Ludwig, and Iva Novakova, eds. 2020. *Phraséologie et stylistique de la langue littéraire approches interdisciplinaires* [Phraseology and Stylistics of Literary Language: Interdisciplinary Approaches]. Berlin, Bern, Bruxelles: Peter Lang.

Henny-Krahmer, Ulrike. 2023. *Genre Analysis and Corpus Design: Nineteenth-Century Spanish-American Novels (1830–1910)*. Universität Würzburg. https://nbn-resolving.org/urn:nbn:de:bvb:20-opus-319992.

Hockey, Susan. 2012. "Digital Humanities in the Age of the Internet: Reaching Out to Other Communities." In *Collaborative Research in the Digital Humanities: A Volume in Honour of Harold Short, on the Occasion of his 65th Birthday and his Retirement, September 2010*, edited by Marilyn Deegan and Willard McCarthy, 81–92. Farnham: Ashgate.

Nielsen, Lars Holm. 2013. "ZENODO – An Innovative Service for Sharing All Research Outputs." Presented at the Joint OpenAIRE/LIEBER Workshop, Ghent, Belgium, May 28, 2013. https://doi.org/10.5281/zenodo.6815.

Schöch, Christof. 2017. "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Classification of Genres through 500 Years of Spanish Literature in CORDE

José Calvo Tello

Abstract In this work I analyze the development of numerous genres in almost five hundred years of Spanish literature. For that, I use a large diachronic corpus composed by the *Real Academia Española*. After an introductory section, the dataset is described, focusing on some important aspects of the distribution and balance of the categories. Next, several classification tests are applied in order to find out which parameters lead to the highest scores, and what the results for each category are. The variance of these results is then explored using linear regression, resulting in the length of the text being a good predictor for the classification results. Finally, the historical evolution is analyzed, showing that the classification results for genres neither get better nor worse over time, but remain stable.

Keywords genre, classification, Spanish literature, corpora

1. Introduction

Several researchers have applied computational methods for genre analysis during the last 30 years. In general, works studying genres can be divided into two groups: the ones with a greater historical interest, looking closely at the development of a few categories (Raible 1980; Schrott 2015; Underwood 2016; 2019; Schröter 2019), and the ones analyzing a larger number of categories with a cross-sectional and comparative interest, trying to classify the different genres (Kessler, Numberg, and Schütze 1997; Stamatatos, Fakotakis, and Kokkinakis 2000; Santini 2011; Jockers 2013; Schöch 2013; Underwood 2014; Hettinger et al. 2016; Henny-Krahmer et al. 2018; Jannidis, Konle, and Leinen 2019). This difference with respect to the perspective of the research needs to be defined in an early stage of the work since it affects the data collection of texts and labels. Until now, only a few studies (like Biber and Conrad 2009) have tried to

Calvo Tello, José: Classification of Genres through 500 Years of Spanish Literature in CORDE, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 15–36. DOI: https://doi.org/10.17885/heiup.1157.c19362

combine both perspectives. However, the costs of this combination in terms of time and money easily exceeds the possibilities of many research projects: it requires collecting thousands of texts written in many centuries and labeled with dozens of categories.

Nevertheless, the collaboration between projects from different disciplines, such as digital humanities and corpus linguistics, can open up new possibilities. The access to data from institutions that have been gathering texts and metadata for years or even decades is especially interesting. For Spanish, the *Real Academia Española* (RAE) has played a central role in the composition of large corpora (Sánchez Sánchez and Domínguez Cintas 2007).

In this contribution, I combine the historical and the comparative perspective on genre analysis. The research question that I pursue is the following: how do various genres develop through the last five centuries of Spanish literature? Although certain works have looked closely at the development of very specific genres such as science fiction (Underwood 2016) or *historiette* (Raible 1980, 341), it is uncertain whether the status or influence of genres have changed over time (Todorov 1976). On the one hand, it is possible that genres had a greater influence in previous centuries in which the normative aspect of genres was perceived to be stronger. If this is the case, the classification results should deteriorate over time. On the other hand, the publishing sector has gone through an increasing professionalization starting in the nineteenth century that could have caused that texts belonging to a category becoming more clearly described in linguistic terms. If so, the classification results should improve over time. Of course, a third outcome is also possible: that the classification results of genres stay stable as time goes by.

In the next section, I will first describe the dataset used. Then, the results of the classification will be shown together with the influence of several classification parameters (algorithms, transformation, number and characteristics of tokens). In the third main section, several characteristics of the categories will be tested to explain the variance of the classification results. Finally, I classify the genres in each decade, looking at large historical patterns. To observe further details and the code used, the accompanying Jupyter Notebooks can be accessed online.[1]

2. Description of the Dataset: CORDE

The diachronic corpus CORDE (*Corpus Diacrónico del Español*) is one of the two corpora that the RAE launched in 2002, along with the contemporary *Corpus de Referencia del Español Actual* (CREA). Originally, the CREA was supposed to span the

[1] https://github.com/cligs/projects2020/blob/master/stylisitcs_500_years_CORDE/code/Explanation%20of%20the%20results.ipynb

most recent 25 years, while CORDE would append every year the texts that became older than 25 years, considered already as historical (Sánchez Sánchez and Domínguez Cintas 2007). In this manner, every year would cause the modification of the boundaries of both corpora. However, the development of a new corpus, *Corpus del Español del Siglo XXI* (CORPES XXI), changed these plans. This new corpus covers the twenty-first century and adds new texts every year. For this reason, the development of the two original corpora of the RAE was frozen.

CORDE contains more than 34,000 texts and around 300 million tokens in its final version. Its material comes from all Spanish-speaking countries, although the majority are from Spain (74 percent of the texts). The corpus contains metadata about genres and topics, and literary works make up a considerable part of the corpus. The dates of production are between the eighth century (with only a few instances) and 1974 (Sánchez Sánchez and Domínguez Cintas 2007). The corpus aims to be a representative sample of the language for research purposes (Sánchez Sánchez and Domínguez Cintas 2007, 143; Rojo Sánchez 2010). Until now, the users have access to websites of CORDE or CREA that allow queries using several filters, such queries in the text, author, title of work, medium, etc. Its acceptance as a standard tool by Hispanic scholars is high (Kabatek and Pusch 2011), although it has also received critique by several researchers who have highlighted the poor philological quality of the medieval section (see a discussion in Rodríguez Molina and Octavio de Toledo y Huerta 2017).

Until recently, researchers could not have access to the full text version of these corpora. The reason given for this were copyright issues relating to the editions of the texts. However, in the last years the RAE has published more data extracted from the corpora and signaled its availability for research requests. In the case of my doctoral thesis (Calvo Tello 2021), I requested the frequencies of the tokens per document. This information cannot be protected by the status of the edition anymore since it does not contain the original text; it only contains facts about the analyzed object (more specifically, how frequently each token appears in each text). Along with this statistical data, several metadata fields were available as well.

To explore the historical development of the genres and topics I use the main label that the RAE assigns to each text. Table 1 shows the original labels and their translation into English.

To my knowledge, these labels do not follow any previous classification system or taxonomy of genres. Many of them could be grouped together in broader groups (such as verse, prose, expositive texts about a topic, etc.). However, it is necessary to recognize that finding a set of labels which can be used of a such a long span of time for such a wide community is an almost impossible task. These labels may be imperfect, but constitute a reasonable solution.

How is the distribution of these categories over time? Figure 1 shows the number of texts in each category over centuries.

Table 1 Labels in Spanish and English

Label in Spanish	Label in English
Artes y espectáculos	Arts and entertainment
Ciencias aplicadas	Applied sciences
Ciencias exactas, físicas y naturales	Exact, physical and natural sciences
Ciencias sociales y humanidades	Social sciences and humanities
Derecho	Law
Historia y documentos	History and documents
Prensa	Press
Prosa	Prose
Prosa didáctica	Didactic prose
Prosa dramática extensa	Extensive dramatic prose
Prosa narrativa breve	Short narrative prose
Prosa narrativa extensa	Extensive narrative prose
Religión	Religion
Sociedad	Society
Verso dramático breve	Short dramatic verse
Verso dramático extenso	Extended dramatic verse
Verso lírico culto	Cultured lyric verse
Verso lírico tradicional	Traditional lyric verse
Verso narrativo culto	Cultured narrative verse
Verso narrativo tradicional	Traditional narrative verse

The bars show a very irregular pattern: the eighth and ninth centuries have very few texts. From the tenth to the twelfth century the number of texts does not exceed one thousand instances. The rest of the centuries can be divided into two groups: centuries with more than 4,000 texts (the fifteenth, sixteenth, and twentieth century) and the rest, with nearly 3,000 instances. Neither the number of texts per century nor the distribution of genres and topics over centuries are balanced. In order to gain a theoretical perspective about corpora it could be desirable to obtain a similar number of texts for different variables such as genre or century (Schöch 2017), but this collides with the historic reality when looked at more closely. For example, it is not possible to balance drama in verse and prose over centuries: drama was written mainly either in verse or prose depending on the historical period. Likewise, it is unrealistic to expect journalistic texts during the Middle Ages or the sixteenth and seventeenth century. Only a few categories show a stable number of texts over more than three centuries and only one category offers a minimal level of stability over the 12 centuries: legal texts.

The number of texts in each category is one possible way of exploring the corpus. However, perhaps the categories have very different typical lengths in words. For

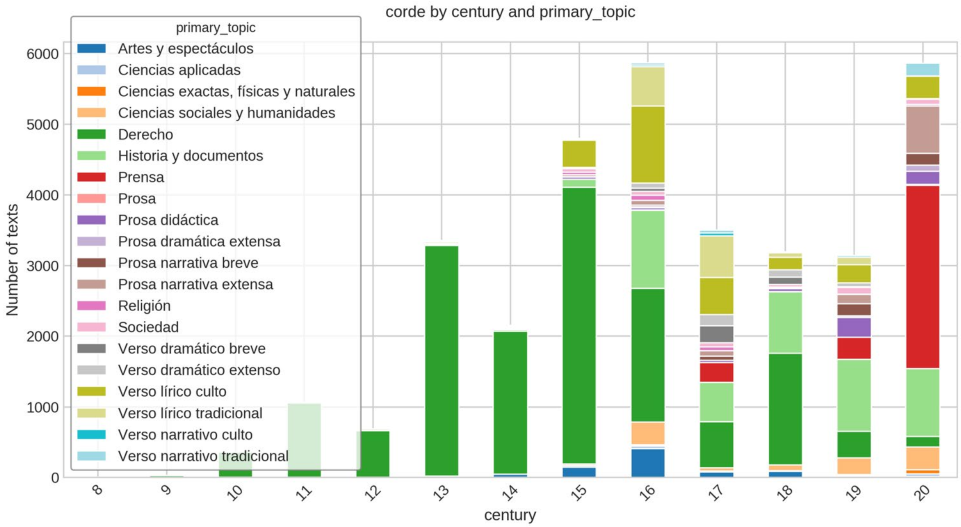


Fig. 1 Distribution of genres in number of texts by century (Calvo Tello, CC BY).

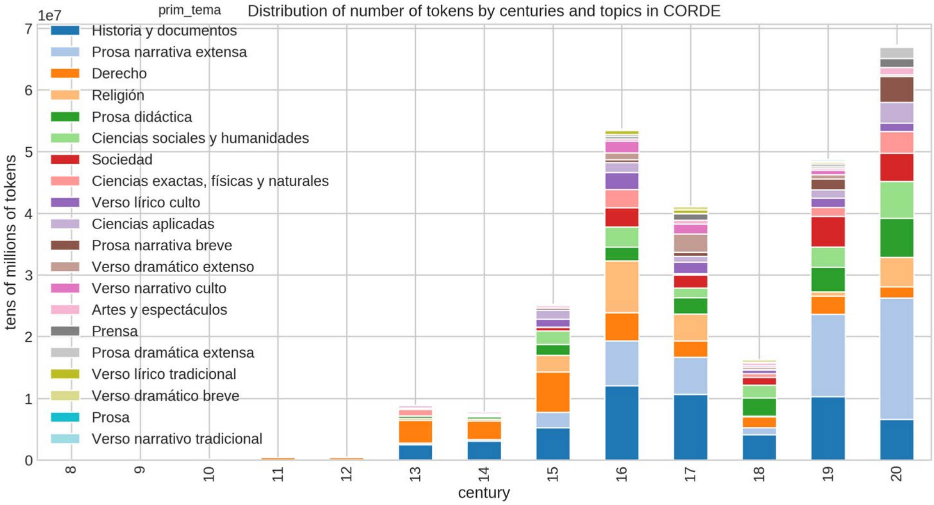


Fig. 2 Distribution of genres in number of tokens by century (Calvo Tello, CC BY).

example, journalistic texts are expected to be shorter than drama or prose fiction. For this reason, the Figure 2 shows the number of tokens populating the category in each century instead of the number of texts.

The overall picture does change notably. Until the twelfth century the pattern is mostly flat, increasing to around 10 million tokens in the thirteenth and fourteenth century. The rest of the centuries are populated by either around 20 million tokens (fifteenth and eighteenth century), or more than 40 million. Not only the total number of instances in each century has changed from Figure 1 to Figure 2, but also the distribution of each category. For example, legal texts predominate vastly in Figure 1, while this is much nuanced in Figure 2 when tokens are counted. A similar effect occurs with journalistic texts (red in Figure 1), whose proportion is reduced in Figure 2 (dark gray) because these texts are typically short. On the contrary, lengthy narrative prose (dominated by novels) has a larger proportion in Figure 2 than in Figure 1 because novels tend to be very long texts.

In any case, both bar plots show that the distribution of centuries and categories are not balanced, and the reason for that is to a certain point obvious: numerous categories do not exist in many of the centuries. Similar effects are observable when other variables of the corpus are taken into account, such as the country where the text was originated. Therefore, some countries have notably more texts in some centuries than in others, and some genres were more profusely written in some regions than in others (more details are to be found in the Jupyter Notebooks).

Instead of trying to artificially balance the corpus (Schöch 2017), I use it as it is and will use a subsampling approach in the analyses. In any case, several variables will be considered for explaining the classification results. However, the entire corpus is not considered: Due to the previous philological critique and the great deviation in terms of texts in the medieval section, only the texts from 1500 onward are analyzed. Additionally, many texts shorter than 100 tokens are also eliminated. With these restrictions, the remaining version of the CORDE contains 18,709 texts and over 192 million tokens. More details can be observed in the Jupyter Notebooks accompanying this publication.

3. Results and Parameters Evaluation

When genre classification is pursued, the researcher needs to define the variables of their data (language, categories, period) but also about the methods applied and which features exactly are accessed by the algorithm. Previous research has closely analyzed the effect of one or two of these so-called parameters (Kessler, Numberg, and Schütze 1997; Stamatatos, Fakotakis, and Kokkinakis 2000; Cerviño Beresi et al. 2004; Berninger, Kim, and Ross 2008; Santini 2011; Allison et al. 2011; Schöch 2013; Jockers 2013;

Underwood 2014; Hettinger et al. 2016; Calvo Tello 2019; Henny-Krahmer et al. 2018; Jannidis, Konle, and Leinen 2019). In this publication, I want to observe the effect of several methodological options. More specifically, I consider the following parameters:

- Categories: the 24 main categories defined by the CORDE
- Classifiers: ridge regression, support vector machines (SVC), logistic regression (LR), decision trees (DT), and random forest (RF)
- Transformation of the token frequencies: logarithmic transformation, z-scores, tf-idf, and binary
- Number of features: 100, 1,000, 2,000, 3,000, 4,000, 5,000
- Punctuation: kept or deleted

All these parameters have been tackled in the previous studies on text classification in the humanities or computer science cited above. How high the results of the classification will be for each possible combination of these five aspects is unknown. For instance, logistic regression when using 5,000 tokens without punctuation and transformed as tf-idf might be the best parameters for the classification of journalistic texts. For other categories, the highest results could be obtained with other parameters. These five aspects are tested in what computer science calls a *grid search* (Müller and Guido 2016, 262–64), that is trying every possible combination of the values of the parameters.

For the classification, a series of steps are taken to control the tests. First, for each category the corpus is split into two subcorpora of equal size, both containing the same number of texts. For example, the category *traditional narrative verse* contains 452 texts in the corpus. For its classification, the same number of texts from other categories are randomly sampled. These two subcorpora represent the instances belonging positively to the category, and those which do not. This produces a binary multi-label classification for each genre. This process is repeated five times to control the effect of the random samples of instances. The maximum number of instances in both subcorpora is settled to 1,000 texts to save computational costs. Finally, ten-fold cross validation is applied and the F1-score is measured in every test. A mean F1-score of this double loop is calculated, which constitutes here the data points of the following visualizations. The combination of all listed parameters above through the double loop produces a total of 276,000 iterations of classifications.

First, I would like to explore the results of the different categories annotated by CORDE. In Figure 3, the 50 highest scores for each category are summarized in box plots.

The red line at the bottom represents the baseline if an algorithm would ascribe to all instances the same category (0.50 since the corpus has been undersampled to create

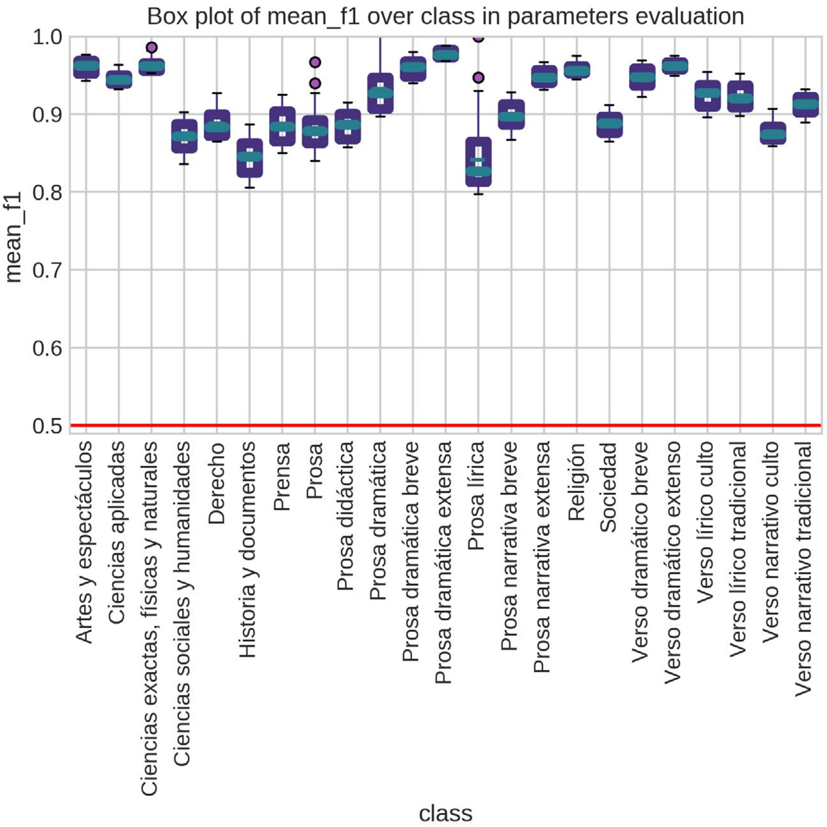


Fig. 3 Results of the classification process by categories (Calvo Tello, CC BY).

subcorpora of equal size). The results show that the classification results are clearly above this baseline. The median of the box plots are between 0.82 F1-score and 0.97 (*Artes y espectáculos, Ciencias exactas, Prosa dramática extensa*, or *Verso dramático extenso*) which are almost perfect results. The lowest scores are achieved by *Prosa lírica* (*Lyrical prose*), a hybrid category that can be easily accepted as difficult to differentiate from other categories. However, the next worst results are for *Historia y documentos* (*History and documents*), a category that I expected to show distinctive linguistic patterns that would differentiate it from the rest. Besides how high the results are, they also show certain variance, with some categories with all their 50 top results very close to each other (*Ciencias exactas, Prosa narrativa extensa*) while others encompass a variance of 10 percent (*Prosa, Prosa lírica, Prosa dramática*). This will be closely analyzed in the following section.

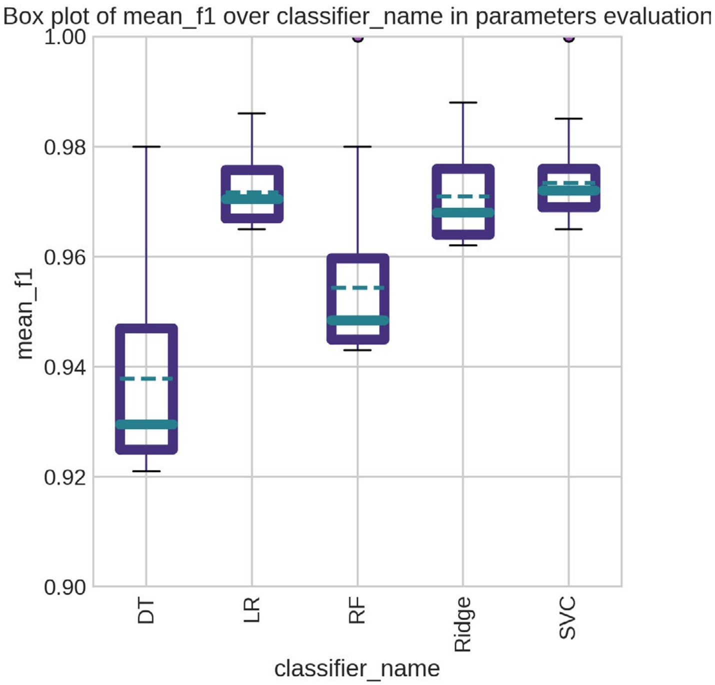


Fig. 4 Results of the classification process by classifiers (Calvo Tello, CC BY).

Later on, I show the top results of several parameters. Since the baseline of 0.5 F1-score is so clearly below the results, the next plots will not show it anymore. To maintain consistent figures, enable comparison, and gain in detail visibility of the results, the vertical axis will show from now on only the results between 0.90 and 1.00 F1-score. As in the previous figure, each box plot contains the top 50 results of each analyzed possibility. The results for the several classifiers are shown in Figure 4.

The box plots show two clear groups: decision trees and random forests with lower results, and logistic regression, ridge and support vector machines with similarly high results. A pairwise t-test between all these scores shows that the differences between these two groups are consistently significant (p-value < 0.001), while the differences between the algorithms with better performance are not significant (more details in the Jupyter Notebooks).

The second analyzed parameter is the transformation of the frequencies of tokens. The original data delivered by the RAE are first put in relation to the length of the document, obtaining the relative frequency of each token in each text. After that,

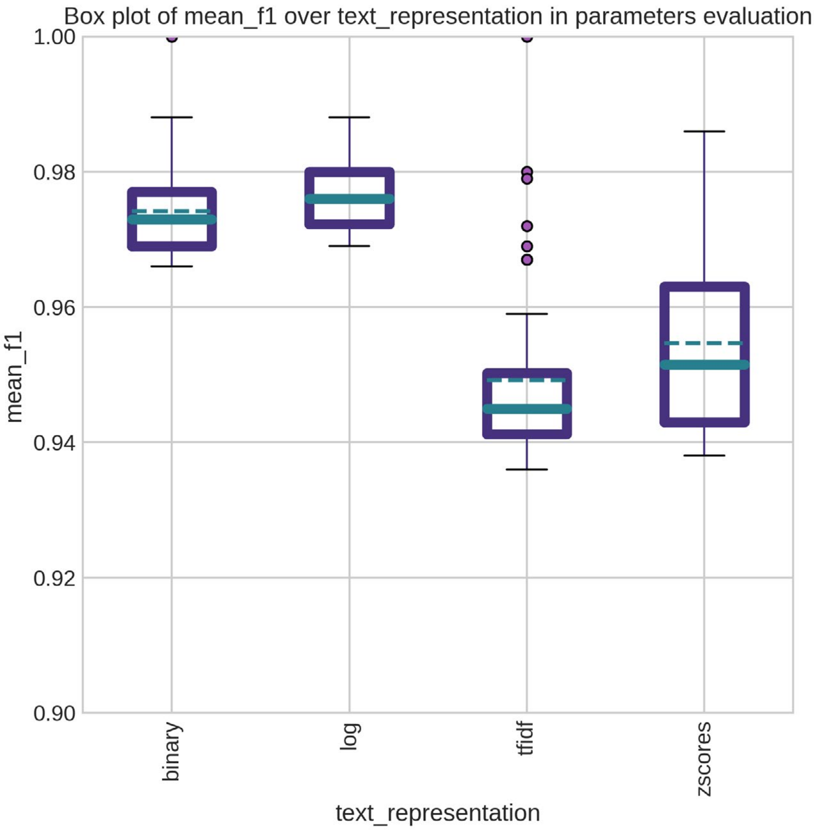


Fig. 5 Results of the classification process by transformations (Calvo Tello, CC BY).

these frequencies are typically transformed in other ways. In many cases this decision is associated with the field that the researcher feels closest to: z-scores in stylometry, tf-idf or binary frequency in computer science, or logarithm in statistics. Due to the corpus size, it was not feasible to run the algorithms using relative frequency.

Two groups of transformations can be observed in Figure 5: tf-idf and z-scores with lower results, and binary and logarithmic transformation with higher results. When these are compared pairwise, the two groups show statistical difference (p-value < 0.001), but the differences are not significant within the groups. In other words, logarithmic transformation does not yield statistically significant higher results than binary frequency (more details in the Jupyter Notebooks). Although these results are similar to what I have observed for my thesis (Calvo Tello 2021), binary frequency leads to

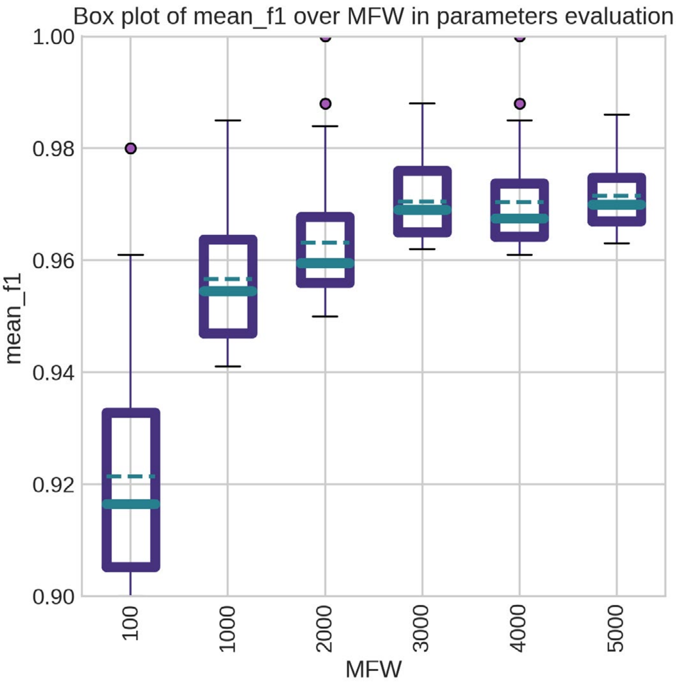


Fig. 6 Results of the classification process by number of features (Calvo Tello, CC BY).

notably higher scores in this case. This might be caused either by a greater number of texts or because the texts cover a much longer span of time. Further comparison in other corpora and languages would be of interest.

The next parameter is the number of features (frequently mentioned as most frequent words or MFW) passed to the algorithms. The results for this are shown in Figure 6.

The results improve up to 3,000 features (with statistical significance in every step, p-value < 0.01). From this point, the box plots overlap and there is no statistical improvement. These results are consistent with previous analyses that have also shown optimal results with around 3,000 features (Underwood 2014; Hettinger et al. 2016; Calvo Tello 2021).

The last analyzed parameter deals with the question whether the typographical tokens are good predictors for these categories or not. Although typographical characters have shown a positive effect in the classification results in other corpora (Calvo Tello 2021), in this case the results are nearly identical. The reason for this can be again the fact that a long period is being analyzed. Even when the medieval section has been ignored for the analysis, the typographical conventions before the eighteenth century are notably different to the latter period. A lack of homogeneity in the philological transcription might diminish the positive effect of typography in this corpus. Further details about this parameter can be observed in the Jupyter Notebooks.

4. Explanation of the Classification Variance

Every publication about genre classification observes that there is a certain variance in the results: Some genres yield notably higher results than others. Although it might not strike anyone as particularly surprising that erotic novels can be classified with higher accuracy than social novels, the exact causes for that are still not fully explained. Previous research has put these differences in relation to the specificity of the features (Jockers 2013; Calvo Tello 2021), or to the disagreement in the labels of human institutions (Calvo Tello 2018). In this section, I want to observe whether the composition of the categories in the corpus might be influencing the results. For this purpose, I run a series of regression analyses comparing, for each category, two numerical variables: the F1-scores and a second variable that hypothetically should predict the classification results.

For example, a question that has been raised is whether genres constituted more robust textual categories in the past, but their importance has diminished over time (Todorov 1976). Especially for the period of modernity at the beginning of the twentieth century, several voices stated the decreasing influence of genres (Bradbury 1978; Calinescu 1987; Romero López 1997; Mainer, Alvar, and Navarro 1997; Buckley 2008; Longhurst 2008). To operationalize this hypothesis, I calculate the median year of publication of all the texts of each category. The median year of novels in the corpus is around 1900, while dramatic verse is typically earlier than 1700. These values are the independent variable, plotted in Figure 7 as the horizontal axis, while the vertical axis represents the mean accuracy of the classification results, measured in F1-scores.

The scatter plot shows that there is in fact a slightly decreasing tendency: The categories that were written in the earliest centuries of the analyzed corpus tend to yield higher scores than the ones published in latter ones. However, the results show a large deviation to the regression line. A regression analysis shows a p-value of 0.23 (in a Wald test, as the SciPy library applies for linear regression analysis). In other words: although there is a certain tendency to find lower classification results in latter centuries, it is not statistically significant.

One of the possible reasons for this is the fact that the median year of publication has reduced the thousands of data points of each publication into a single numerical value for each category. This reduction might have removed too much information.

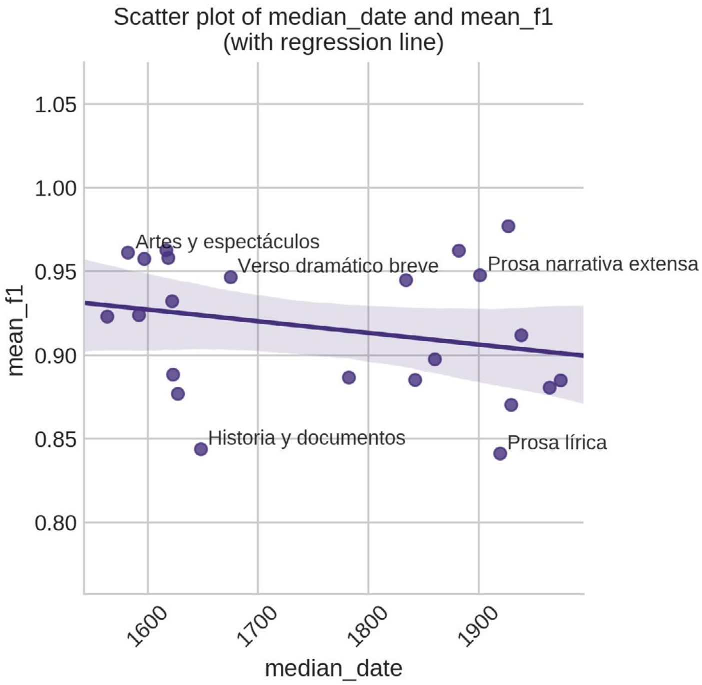


Fig. 7 Scatter plot with the categories by their median date and their classification results (Calvo Tello, CC BY).

Each category can also be captured by its chronological dispersion in the corpus. While journalistic texts are placed mainly in only two centuries, other categories cover the entire spectrum of centuries (*exact science, short narrative prose, legal texts*). To capture this, I have also calculated the standard deviation of the dates of publication, and use it as horizontal axis in Figure 8.

In this case the results show a much clearer tendency: The categories with greater standard deviation in their publication dates tend to show also higher classification results, like the category on the right of the scatter plot. On the contrary, *press* or *lyrical prose* show a narrower span of years of publications and this correlates with lower classification results. This tendency is statistically significant (p-value = 0.02).

However, this dispersion of the categories in the corpus might be hiding other lurking variables that correlate with the temporal dispersion. Perhaps the categories with a large standard deviation of years of publication are also the categories with more instances. More data per category might be beneficial to the classifier. To operationalize this, I measure the number of texts per category and apply it as an independent

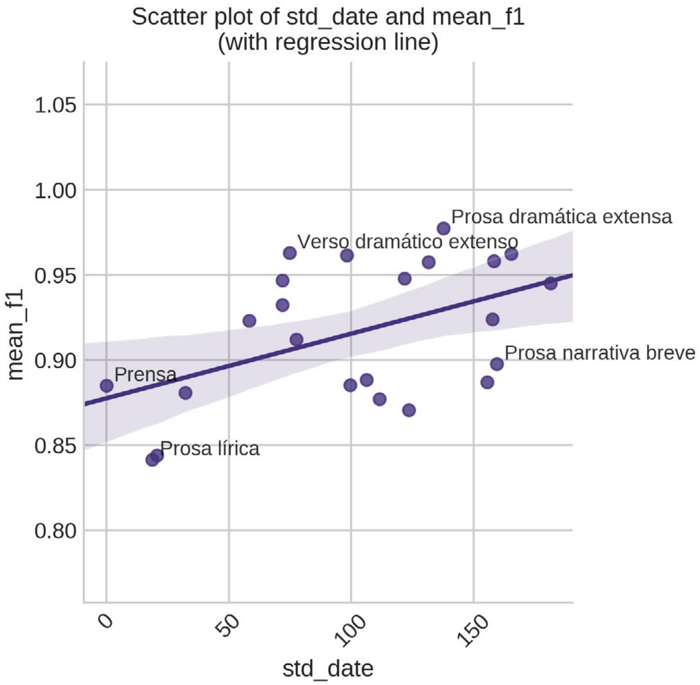


Fig. 8 Scatter plot with the categories by their standard deviation date and their classification results (Calvo Tello, CC BY).

variable. A regression analysis shows no correlation between both variables (p-value = 0.83). In this corpus, the size of the category cannot be used to predict the classification results (the scatter plot and further details can be observed in the Jupyter Notebooks).

The effect of a further variable is derived from Figure 3, which shows that the long versions of some categories tend to yield higher scores than the short ones. For example, *long narrative prose* achieves higher results than its short counterpart. This might be pointing toward a correlation between the length of the documents and the classification results: The more words a text has, the more data the classifier has to predict its category correctly. In the following figure, each category is represented by the median length in tokens of their texts. The position of the short and lengthy versions of several categories (*dramatic prose, dramatic verse, narrative prose*) can be read in labels of Figure 9.

It can be observed that in all cases, the long version of a category leads to higher results than the short one. For example *Prosa narrativa breve* (*short narrative prose*) has an F1-score around 0.90, while the *Prosa narrativa extensa* (*long narrative prose*) has an

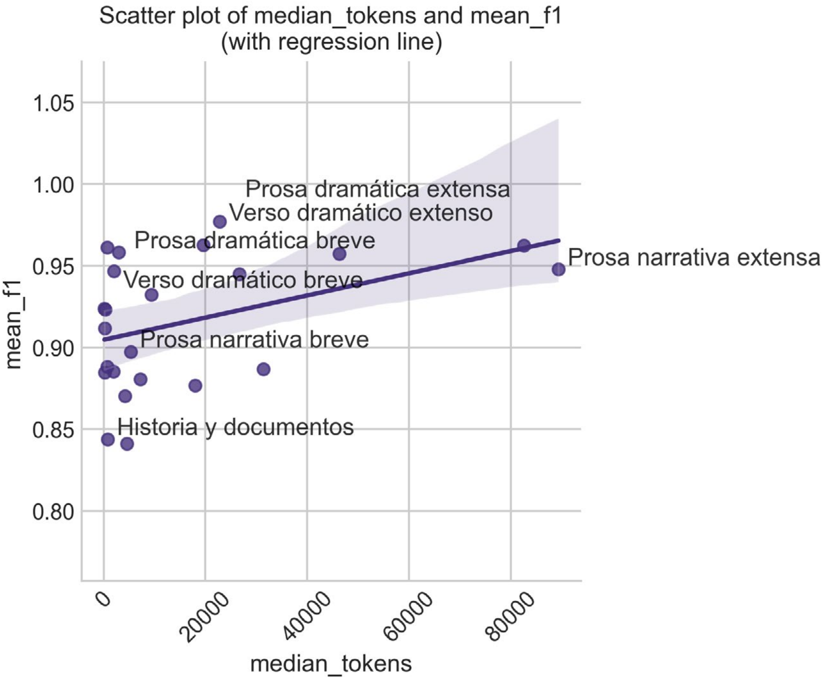


Fig. 9 Scatter plot with the categories by their median length in tokens and their classification results (Calvo Tello, CC BY).

F1-score of 0.95. A regression analysis gives a positive slope (p-value < 0.05), meaning that each additional token contributes to a slightly better classification result. This could also be the explanation for the low results of the category *Historia y documentos* (*history and documents)*: the texts pertaining to this category are among the shortest of the entire corpus.

5. Genre Classification by Decades: Historical Development of the Results

The most important question in this article is what is the general pattern of these categories over the five last centuries. Do the classification results become higher or lower over time? As mentioned earlier, there are good arguments for both possibilities.

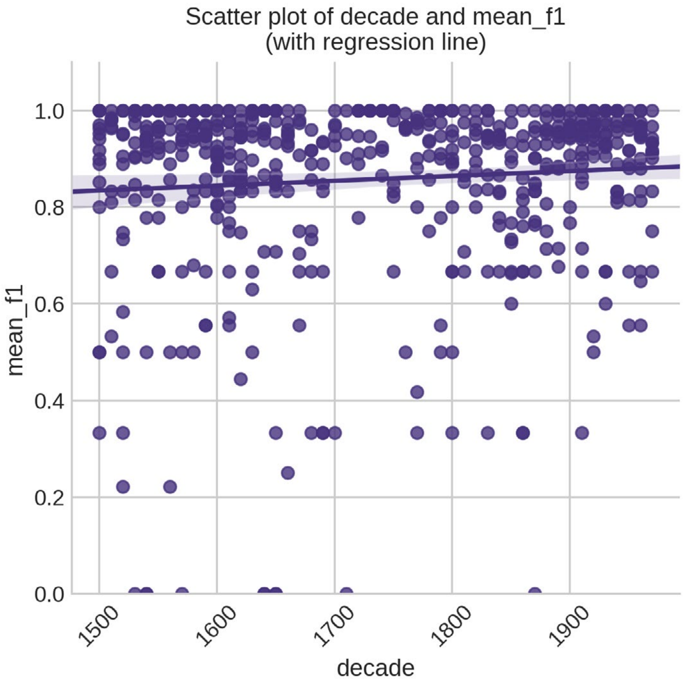


Fig. 10 Scatter plot with the categories by their decade and their classification results (Calvo Tello, CC BY).

To observe the classification results from a historical perspective, I split the corpus in the 47 decades spanning between 1500 and 1970. In each decade I repeat the classification process explained in the previous section, analyzing all the categories. In Figure 10, each data point is one category in each decade.

The overall tendency is positive: the classification results tend to improve over time. This seems to support the second hypothesis that the genres have gone through a process of professionalization over time and therefore genres can be classified in later centuries more accurately using linguistic features. However, this improvement is not statistically significant (p-value = 0.066). In other words, genres can be classified better in later centuries, but that the hypothesis that this tendency is just produced by the random variability of the data cannot be rejected.

That is the general pattern, but how does each genre evolve over time? To answer this question, I have evaluated which categories do present a statistical tendency in their classification results. Only three of them show p-values under 0.05. Two of them improve their classification results: *Ciencias exactas, físicas y naturales* (*Exact,*

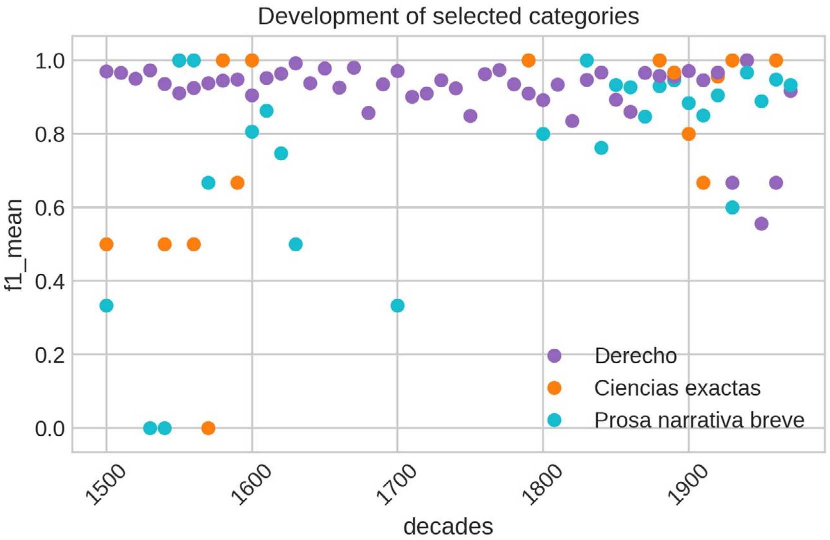


Fig. 11 Scatter plot with three categories for which the classification results show a statistical pattern (Calvo Tello, CC BY).

physical and natural sciences, p-value = 0.03) and *Prosa narrativa breve* (*Short narrative prose*, p-value = 0.002). In comparison, the classification of the category *Derecho* (*Law*, p-value = 0.007) tends to get lower results over time. The tendencies of these three categories are shown in Figure 11.

Besides these categories, the rest lead to very similar results over the entire period: Genres cannot be classified neither better nor worse as time goes by.

Figure 10 shows a positive tendency, although it is not statistically significant. Nevertheless, I want to explore whether this trend can be explained by spurious variables. In the previous section, I have observed that the length of the texts correlate with the classification results (Figure 9). The observed tendency in Figure 10 could be partially explained if texts become lengthier as time goes by. To explore this, Figure 12 represents each text as a data point, with its year of publication in the horizontal axis and its length in the vertical one.

Although there is a large variance in all periods, there is certain trend for longer texts in the later centuries than in the earlier ones. This can be evaluated through a regression analysis that confirms that texts tend to become longer, more specifically, 10 tokens longer each year on average (p-value < 0.001).

If the classification of these categories leads to higher results with longer texts and texts tend to become longer over time, the logical consequence is that classification

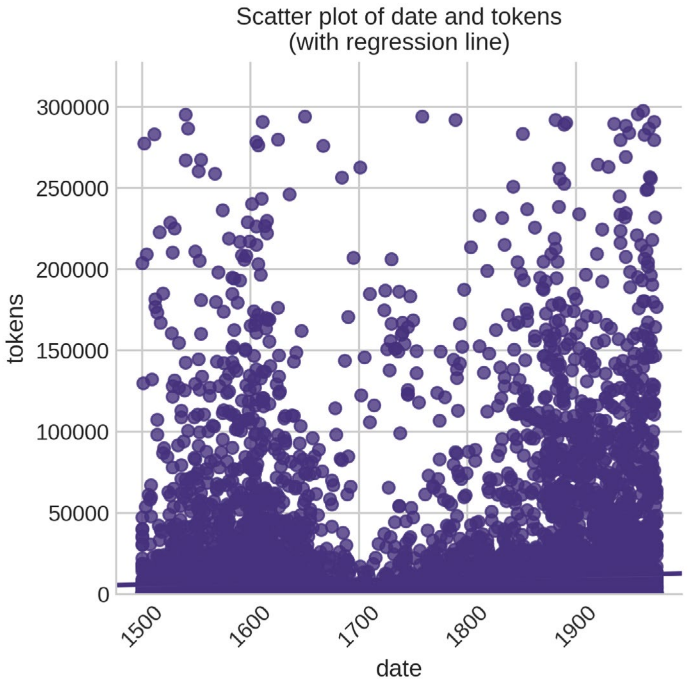


Fig. 12 Scatter plot with the texts by their year and their length in tokens (Calvo Tello, CC BY).

would tend to work better in the later centuries. This is what it was observed in Figure 10. To control for text length, I sample from the until now analyzed subcorpus, taking only the texts in the middle range of length. More specifically, I calculate the 25th and the 75th percentile of the tokens and retain only the texts within this middle range. That produces a corpus of 9,366 texts that are taken as the basis for a new classification test as described above.

The results of the classification in Figure 13 when the length of the texts is controlled shows exactly the opposite direction from the one observed in Figure 10. In this case, lower scores are obtained over time. However, as in the previous case, this tendency is not statistically significant (p-value = 0.10). The length of the texts seems to have an impact on the results: The direction of linear regression changes, and the p-value increases, meaning that the probability of the pattern being caused by randomness is higher. Overall, the results reject both hypotheses about the development of these categories over time: genres cannot be classified either better or worse over time; rather, the classification accuracy remains stable over time. This outcome contradicts

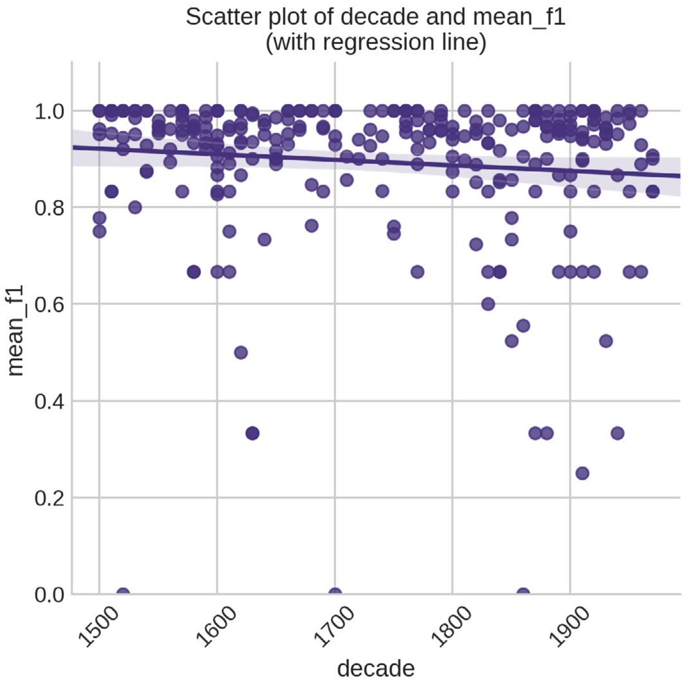


Fig. 13 Scatter plot with the categories controlling the length, by their decade and their classification results (Calvo Tello, CC BY).

the expectations from many literary scholars who, as cited before, assume that genres had greater influence, at least previous to the nineteenth century.

6. Conclusion

Have genres gained or lost importance over time? Did these labels describe the linguistic content of the texts better in earlier than in later centuries? To answer this leading question, I have used one of the largest diachronic corpora for Spanish composed by the RAE, which recently has opened up new possibilities of accessing the frequencies of tokens and metadata of their corpora. Larger datasets can reconcile the two perspectives on genre analysis: the cross-sectional and the historical one. Moreover, this research is an example about how corpora launched some decades ago can be currently combined with new methodologies such as machine learning.

The first main conclusion of this contribution is derived from the description of the dataset: The categories are neither balanced over time nor by other dimensions such as geographical origin. Although this can be seen as a defect in the composition of the corpus, a closer look reveals the impossibility of balancing genres such as dramatic texts (in prose or verse) or journalistic texts over periods of several centuries. The data show that some textual categories were produced more profusely during specific periods.

The second conclusion is that the multi-label classification yields very high results for all genres, some of them being very close to perfect scores. Moreover, the parameter analysis using grid search indicates that logistic regression (both in its classic version and as ridge regression) and support vector machines achieve better results than decision trees or random forest. Relating to the features, around 3,000 of them transformed logarithmically or binary tend to acquire the best results.

The third conclusion points toward two variables that can predict the classification results. In this corpus, the length of the texts of each category and the chronological span of the published texts can predict to a certain degree the classification results. In other words, if a genre tends to be populated by long texts, or if a genre was published over a long period of time, its classification results tend to be higher.

The final conclusion about the historical development shows that classification of genres remains stable over centuries. The results do not indicate either an increase or a decrease in the classification results. In other words, the majority of the genres neither gain nor lose linguistic distinctiveness. However, a closer look shows that two categories (technical scientific texts and short narrative prose) have a tendency to lead to higher classification scores in later centuries, while legal texts show the opposite trend.

This research offers a Jupyter Notebook companion which can be consulted to obtain deeper details than what is reported in this article.[2] Although these conclusions are relevant, it would be necessary to observe similar questions in several datasets. It would be of particular interest to analyze whether other sources of the labels of genres (information from the cover, literary scholars' opinions, etc.) lead to similarly stable patterns over centuries. Moreover, other languages might show dissimilar tendencies. Interdisciplinarity and collaboration between academic and national traditions can lead us to a deeper understanding of how cultures have changed over centuries through their texts.

ORCID®

José Calvo Tello https://orcid.org/0000-0002-1129-5604

[2] https://github.com/cligs/projects2020/blob/master/stylisitcs_500_years_CORDE/code/Explanation%20of%20the%20results.ipynb

References

Allison, Sarah, Ryan Heuser, Matthew L. Jockers, Franco Moretti, and Michael Witmore. 2011. *Quantitative Formalism: An Experiment* (Stanford Literary Lab, Pamphlet 1). Stanford: Stanford Literary Lab.

Berninger, Vera, Yunhyong Kim, and Seamus Ross. 2008. "Building a Document Genre Corpus: A Profile of the KRYS I Corpus." In *BCS-IRSG Workshop on Corpus Profiling*. London: BCS Learning and Development Ltd.

Biber, Douglas, and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge, UK; New York: Cambridge University Press.

Bradbury, Malcolm, ed. 1978. *Modernism: 1890–1930*. Pelican Guides to European Literature. Hassocks: Harvester Pr.

Buckley, Ramón. 2008. "Tales from the Avant-Garde." In *A Companion to the Twentieth-Century Spanish Novel*, edited by Marta E. Altisent, 1. publ., 45–59. Woodbridge: Tamesis.

Calinescu, Matei. 1987. *Five Faces of Modernity: Modernism, Avant-Garde, Decadence, Kitsch, Postmodernism.* Durham: Duke University Press.

Calvo Tello, José. 2018. "Genre Classification in Spanish Novels: A Hard Task for Humans and Machines?" In *Data in Digital Humanities*. Galway: EADH.

Calvo Tello, José. 2019. "Gattungserkennung über 500 Jahre." In *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts*, 292–94. Frankfurt, Mainz: DHd.

Calvo Tello, José. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld: transcript.

Cerviño Beresi, U., José Juan García Adeva, R. A. Calvo, and Hermenegildo Alejandro Ceccatto. 2004. "Automatic Classification of New Articles in Spanish." In *X Congreso Argentino de Ciencias de la Computación*, n. p. http://sedici.unlp.edu.ar/handle/10915/22551

Henny-Krahmer, Ulrike, Katrin Betz, Daniel Schlör, and Andreas Hotho. 2018. "Alternative Gattungstheorien: Das Prototypenmodell am Beispiel hispanoamerikanischer Romane." In *DHd 2018 Digital Humanities: Kritik der digitalen Vernunft. Konferenzabstracts*, 105–12. Köln: DHd.

Hettinger, Lena, Isabella Reger, Fotis Jannidis, and Andreas Hotho. 2016. "Classification of Literary Subgenres." In *DHd 2016 Digital Humanities. Konferenzabstracts*, 154–58. Leipzig: Universität Leipzig.

Jannidis, Fotis, Leonard Konle, and Peter Leinen. 2019. "Makroanalytische Untersuchung von Heftromanen." In *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts*, 167–73. Frankfurt am Main/Mainz: DHd. https://zenodo.org/record/2600812.

Jockers, Matthew L. 2013. *Macroanalysis – Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.

Kabatek, Johannes, and Claus D. Pusch. 2011. *Spanische Sprachwissenschaft: eine Einführung.* Tübingen: Narr.

Kessler, Brett, Geoffrey Numberg, and Hinrich Schütze. 1997. "Automatic Detection of Text Genre." In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 32–38. Stroudsburg, PA: Association for Computational Linguistics.

Longhurst, Carlos Alex. 2008. "The Early Twentieth-Century Novel." In *A Companion to the Twentieth-Century Spanish Novel*, edited by Marta E. Altisent, 30–44. Woodbridge: Tamesis.

Mainer, José-Carlos, Carlos Alvar, and Rosa Navarro. 1997. *Breve historia de la literatura española*. Madrid: Alianza.

Müller, Andreas C., and Sarah Guido. 2016. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. Beijing: O'Reilly.

Raible, Wolfgang. 1980. "Was sind Gattungen?" *Poetica* 12: 320–49.

Rodríguez Molina, Javier, and Álvaro Sebastián Octavio de Toledo y Huerta. 2017. "La imprescindible distinción entre texto y testimonio: el CORDE y los criterios de fiabilidad lingüística." *Scriptum digital: revista de corpus diacrònics i edició digital en llengües iberoromàniques* 6: 5–68.

Rojo Sánchez, Guillermo. 2010. "Sobre codificación y explotación de corpus textuales: Otra comparación del Corpus del español con el CORDE y el CREA." *Lingüística* 24: 11–50.

Romero López, María Dolores. 1997. "Hispanic Modernismo in the Context of European Symbolism – Towards a Comparative Dekon-Struction." *Orbis Litterarum* 52 (3): 194–210. https://doi.org/10.1111/j.1600-0730.1997.tb01978.x.

Sánchez Sánchez, Mercedes, and Carlos Domínguez Cintas. 2007. "El banco de datos de la RAE: CREA y CORDE." *Per Abbat: boletín filológico de actualización académica y didáctica* 2: 137–48.

Santini, Marina. 2011. *Automatic Identification of Genre in Web Pages: A New Perspective*. Saarbrücken: Lambert Academic Publishing.

Schöch, Christof. 2013. "Fine-Tuning Our Stylometric Tools: Investigating Authorship and Genre in French Classical Theater." In *Digital Humanities 2013: Conference Abstracts*. Lincoln: UNL: 383–6.

Schöch, Christof. 2017. "Quantitative Analyse." In *Digital Humanities: Eine Einführung*, edited by Fotis Jannidis, Hubertus Kohle, and Malte Rehbein, 279–98. Stuttgart: Metzler.

Schröter, Julian. 2019. "Gattungsgeschichte und ihr Gattungsbegriff am Beispiel der Novellen." *Journal of Literary Theory* 13 (2): 227–57.

Schrott, Angela. 2015. "Kategorien diskurstraditionellen Wissens als Grundlage einer kulturbezogenen Sprachwissenschaft." In *Diskurse, Texte, Traditionen: Modelle und Fachkulturen in der Diskussion*, edited by Franz Lebsanft and Angela Schrott, 115–46. Göttingen: V&R unipress.

Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis. 2000. "Automatic Text Categorization in Terms of Genre and Author." *Computational Linguistics* 26 (4): 471–95.

Todorov, Tzvetan. 1976. "The Origin of Genres." *New Literary History* 8 (1): 159–70.

Underwood, Ted. 2014. "Understanding Genre in a Collection of a Million Volumes, Interim Report." https://doi.org/10.6084/m9.figshare.1281251.v1.

Underwood, Ted. 2016. "The Life-Cycle of Genres." *Journal of Cultural Analytics* 2 (2). https://doi.org/10.22148/16.005.

Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change*. Chicago: The University of Chicago Press.

Digital Stylistics Applied to Golden Age Spanish Poetry

Is Fernando de Herrera Really a Transitional Poet between Renaissance and Baroque?

Laura Hernández-Lorenzo

Abstract This paper applies Digital Stylistics methods to Golden Age Spanish poetry, one of the most important literary periods of Spanish literature, and to Fernando de Herrera's poems, who has been considered a transitional writer between the Renaissance style of Garcilaso de la Vega and the Baroque of Luis de Góngora. The aim of this study is to analyze Herrera's role in the stylistic evolution from Renaissance to Baroque and to verify if the posthumous edition of his poetry, *Versos* (1619), is more Baroque, as some critics have suggested. For this purpose, a stylometry technique (Zeta), different features (words and PoS n-grams) and parameters (PoS bigrams and trigrams) have been used. Results point to the transitional role of Herrera's work in general, with the detection of a more Baroque component in *Versos* edition through some of the analyses.

Keywords stylometry, poetry, Baroque

1. Introduction

1.1 Golden Age as a Period in Spanish Poetry

The Golden Age—*Siglo de Oro* in Spanish—is a very well established and distinguished period in the history of Spanish literature. According to Spanish literature scholars, it covers literary works from the beginning of the sixteenth century to the end of the seventeenth century, including works from Renaissance and Baroque literature in Spain.[1]

[1] For more information on the emergence of this concept, see Rozas' chapter (Rozas 1983).

Hernández-Lorenzo, Laura: Digital Stylistics Applied to Golden Age Spanish Poetry. Is Fernando de Herrera Really a Transitional Poet between Renaissance and Baroque?, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 37–52.
DOI: https://doi.org/10.17885/heiup.1157.c19365

This period is so well established that it counts a large number of studies, academic associations, such as the *International Association Siglo de Oro* (AISO),[2] the *Society for Renaissance and Baroque Hispanic Poetry* (SRBHP)[3] or the *International Association of Spanish Drama of the Golden Age* (AITENSO);[4] research groups, thematic conferences and journals, etc.[5]

From a traditional point of view, literary scholars have distinguished two main and opposites periods—each of them having a very different style from the other—in Golden Age Spanish poetry: the Renaissance and the Baroque. In the sense of style evolution in this period and how the change from the first to the second style is produced, there are some relevant studies, especially the book by Begoña López Bueno (2000), and another one she coordinated (2006).

1.2 Stylometry Applied to Golden Age Spanish Poetry

It is well known that, although stylometric methods have been widely applied to English texts, research in other languages is not as extensive. In the case of Spanish texts, they suffer from few resources and repositories. Since it is necessary to have machine-readable texts available in order to enable computational literary studies, Spanish scholars have focused on first creating the required resources. As a consequence, digital humanities projects with Spanish texts mostly revolved around digital editions, digital libraries, and databases, while stylometric studies were not very frequent.

However, there has been a notable increase in stylometric research in recent years. In this sense, the most controversial anonymous Spanish works have already been analyzed with these techniques (de la Rosa and Suárez 2016; Rißler-Pipka 2016a; 2016b; Blasco 2016), as well as plays attributed to Cervantes (Calvo Tello and Cerezo Soler 2018), and other authors (García-Reidy 2019; Ulla Lorenzo, Martínez Carro, and Calvo Tello 2021), Spanish novels of the Silver Age (Calvo Tello 2019; 2021), Middle Ages works, and contemporary narrative (Fradejas Rueda 2016; 2019), among others.

[2] Website of the *International Association Siglo de Oro* (AISO): https://aiso-asociacion.org/ (accessed: 31/10/2019).

[3] Website of the *Society for Renaissance and Baroque Hispanic Poetry*: https://srbhp.org/ (accessed: 31/10/2019).

[4] Website of the *International Association of Spanish Drama of the Golden Age*: https://aitenso.net/ (accessed: 7/11/2019).

[5] Some examples for journals are *Studia Aurea* (https://studiaaurea.com/), *Edad de Oro* (https://revistas.uam.es/edadoro), *Calíope* (https://srbhp.org/caliope/), or *Hipogrifo. Revista de literatura y cultura del Siglo de Oro* (https://www.revistahipogrifo.com/index.php/hipogrifo) (accessed: 12/10/2022).

In the case of Golden Age Spanish poetry, there have been some applications of stylometry and quantitative analysis. It is worth mentioning the studies on metrical and semantic aspects in the ADSO project (Navarro-Colorado 2017; 2018), the automatic detection of enjambment in the POSDATA project (Ruiz Fabo et al. 2017), studies on the mythical fable and on Luis de Góngora's style (Rojas Castro 2017; 2018; Lescasse 2019; Rißler-Pipka 2019), an evaluation study on the use of metrical annotation in several poetic corpora and, among them, in Golden Age Spanish poetry (Plecháč, Bobenhausen, and Hammerich 2018), a paper and a dissertation on the authorship of a poetic work attributed to Fernando de Herrera (Hernández Lorenzo 2019a; 2020), a study on size constraints and strength of the stylometric signal applied to a corpus of Golden Age Spanish poets (Hernández Lorenzo 2019b), and a paper on stylistic change on Golden Age Spanish poetry through network analysis (Hernández-Lorenzo 2022). It is easily noticed that all this research is fairly recent, and definitely, there is still much work to be done.

2. Our Research Case: Fernando de Herrera as a Transitional Poet

Fernando de Herrera (1534–1597), also known as *The Divine*, is considered as one of the most important writers in Golden Age Spanish poetry. According to respected Spanish literature scholars, his poetry represents or acts as a bridge between Renaissance and Baroque writing (Vilanova 1951; Valbuena Prat 1960; Ruestes 1986; López Bueno 2000).

In addition, this view of Fernando de Herrera as a transitional poet between these two styles is essentially related to the authorship problem suffered by his poetic works, which is known as the *textual drama*. In life, Herrera published only one edition of his poetry, titled *Algunas obras* (1582), and known as *H*. It includes a selection of his poetic texts prepared by himself. However, some years after his death, the painter Francisco Pacheco (1564–1644)—teacher and father-in-law of the great Spanish painter Diego Velázquez (1599–1660)—published a new edition of Herrera's poems, titled *Versos de Fernando de Herrera* (1619) and known as *P*. It included new poems and different versions of the ones published in *Algunas obras*.

The debate started when the famous writer Francisco de Quevedo (1580–1645) noted numerous and significant differences between the two editions. Since then, and especially since the start of the twentieth century, academics and experts on Herrera's poetry have discussed the style and authenticity of *Versos* for decades without reaching an agreement. On the one hand, erudite Italian academics (Battaglia 1954; Macrí

1959; 1972; Pepe Sarno 1981; 1982; 1986; 1998) defended Herrera's full authorship and the probability of an evolution of Herrera's style toward Baroque and the poetry of Góngora. On the other hand, recognized Spanish philologists (Blecua 1958; 1975; Kossoff 1957a; 1957b; 1965; 1966; Cuevas 1985) were against Herrera's full authorship of *Versos* and suspected another writer's hand on the text, presumably Pacheco's, who was also a poet as well as an admirer of Herrera's writing.

The application of stylometric methods to this authorial problem supports the authenticity of *Versos* (Hernández Lorenzo 2019a). This is not the focus of the present study, though. In spite of the authorship outcome, the relevant point is that different positions on the controversy regarding the full authorship of the posthumous edition are the result of different views on Herrera's role in Spanish poetry. More specifically, those scholars supporting Herrera's full authorship defend that his style became closer to the Baroque aesthetic, whereas the ones rejecting *Versos* authenticity do not believe in this stylistic change and conceive *The Divine* as a Renaissance poet. Notwithstanding the different opinions on *Versos'* authenticity, scholars from both sides of the debate seem to reach an agreement on the more Baroque component of this edition's poems. This would be the most important difference between the poems published in life and those published after Herrera's death. Salvatore Battaglia, José Manuel Blecua and Cristóbal Cuevas particularly stressed this distinctive Baroque component of *Versos* (Battaglia 1954, 87; Blecua 1958, 391; Cuevas 1985, 99). For a more detailed account on Fernando de Herrera as a transitional writer from literary scholars' points of view, see Hernández-Lorenzo 2022.

2.1 Aims of This Study

To this point, we have reviewed some concepts of key importance for our study. On the one hand, the interest in the period known as Golden Age Spanish poetry as well as the state-of-the-art of stylometric applications to its texts and on the other hand, the case of Fernando de Herrera as a transitional poet in this period, related to the controversy about the posthumous edition of his poetry.

This provides us with a solid ground, contextualization and research questions for our stylometric study on Fernando de Herrera's poetic texts from the point of view of its role in the fascinating and rich period of Spanish poetry of the Golden Age. Hence, this study aims to answer the following questions:

- If we consider the first printed edition H, is Fernando de Herrera really a transitional poet between Spanish Renaissance and Baroque?
- What is the role of P in this? Is it as transitional as Herrera's earlier edition? Or is it more Baroque, as some critics have suggested?

3. Stylometric Analysis

In this section, a stylometric analysis is presented to answer the previous questions. It is divided into three main parts: firstly, the description of the dataset used in the experiment—with some necessary preliminary steps and information on corpus building; secondly, a brief explanation of the methodology applied in this study; and thirdly, the discussion of the results obtained.

3.1 Dataset

In order to carry out a quantitative textual project on Spanish texts, the first obstacle that the researcher must face is the unavailability of textual resources and the consequent necessity to digitize some of the textual material on their own. Herrera's entire poetic works were—the undoubted and doubted—not available in a suitable digital format. Thus, they were digitized through the application of optical character recognition (OCR) software to the most authorized annotated edition (Herrera 1975). The resulting text was revised and OCR mistakes were corrected manually. However, a parallel version with modernized standard orthography was created, with the aim to enable some of the computational analyses that cannot be done on the version with historical orthography. On the basis of the original digitalized version, the modernized one was prepared through manual and computer-assisted adaptation of the text to standard Spanish orthography.[6] This way, it was ensured that spelling differences between the two Herrera's editions were not an issue.[7]

In addition, more Spanish texts of the same period were needed in order to answer the previous research questions, serving as a contextualization of Herrera's role and place in the poetic transition from Renaissance to Baroque. For this purpose, some authors were highlighted as referents for these styles. On the one hand, Garcilaso de la Vega (1501–1536), Juan Boscán (1490–1542) and Juan de Almeida (1542–1572) are considered important referents for Renaissance style by literary critics; on the other hand, Luis de Góngora (1561–1627) and Francisco de Quevedo (1580–1645) are considered referents for Baroque style by literature scholars. Their texts were extracted from the Corpus of Spanish Golden Age Sonnets, also known as

[6] In order to speed out the process, some systematic changes, such as diacritics removal, were carried out using the "search and replace" function in the text editor used (UltraEdit). Microsoft Word's Autocorrect feature was also of great assistance in this process. The final version was revised through a careful reading of the text.

[7] For example, we find *del* in *H* and *d'el* in *P*, and they both stand for the same word ('of the').

ADSO, (Navarro-Colorado, Ribes Lafoz, and Sánchez 2016). However, this corpus was used with some changes:

Firstly, the ADSO corpus provides some of Herrera's sonnets. In spite of this, in this study our own digitized texts of Herrera's works were used—selecting only sonnets—, since the ADSO texts were not as complete and, most importantly, combined the doubted and undoubted compositions, regardless of the particularities and authorial problems presented by the posthumous poems. In the case of *P*, the analysis was restricted to its unique poems, that is, the new ones, which were not published in *H*. This sub-corpus was called *P2*.[8]

Secondly, the ADSO texts were contrasted with the digitized texts of Herrera and Pacheco in order to ensure that there were no orthographic differences which could interfere in the analysis.[9] This was done automatically through the extraction of keywords and the application of the oppose function in the *stylo* package in R (Eder, Rybicki, and Kestemont 2016).

As sonnets are very short, the poems written by each of the authors were collected together in one or two files per author, depending on the size of the resulting text. In terms of corpus balance, all authors are from Spain and male. It is well balanced in terms of period, since it contains only Golden Age poets, starting from the early Renaissance and the beginning of the sixteenth century, embodied by Garcilaso de la Vega, to the Baroque climax depicted by Góngora and Quevedo's poems; and literary genre, as it only includes poetic texts, and even from only one sub-genre, as all are sonnets. Also, if one considers that the main goal of this study is to establish Herrera's role, who is a Sevillian male writer, the resulting corpus, albeit not perfect, seems adequate enough to answer the research questions.

3.2 Methodology

As mentioned previously, stylometric techniques are used with the aim of answering the research questions. The next step was, then, to choose which was the most accurate for our purpose. Despite being less popular than Delta (Burrows 2002), Zeta (Burrows 2007; Craig and Kinney 2009) has revealed itself as a powerful tool for comparing and contrasting corpora with the aim of finding which words are unusually more frequent

[8] With the aim of avoiding possible biases created by poems that appear in both editions, the analyses are restricted to the new poems of *P*.

[9] Spelling and orthographic differences are one of the features that emerge when comparing corpora. Using the oppose function in *stylo*, one of the preferred words for Herrera and Pacheco's texts was *solo* ('only'), whereas one of the avoided words—and preferred by ADSO corpus—was *sólo*. When the orthography of this word was standardized to *solo*, both *solo* and *sólo* disappeared from the preferred and avoided wordlists.

in one corpus than others (Schöch 2017; 2018). As the goal of this paper is to mark out the role of Herrera's editions in the change of style from the Renaissance to the Baroque in Spanish poetry, Zeta will assist us in answering this question in two ways: firstly, determining which are the main features that typical poets of these periods possess, and, secondly, using these features as a reference point to establish the place of both *H* and *P2*.

The texts are analyzed across different features, in order to test consistency of results: words and part-of-speech (PoS) tags. The analysis of word tokens is one of the most effectives in stylometry to this date, whereas the examination of PoS n-grams is of great interest for the periods under study, as literature scholars have claimed important syntax differences between Renaissance and Baroque Spanish poetry. For this reason, a parallel version of the corpus was annotated with PoS) information using the software Freeling (Padró 2011; Padró and Stanilovsky 2012). In order to keep only the information of the morphological category, Freeling tags were shortened to the two first characters (i.e. NC, that is, common noun, instead of NCMS, common noun in masculine singular form). PoS bigrams and trigrams were chosen for this first approach, with the possibility of using other n-grams in future studies. PoS bigrams and trigrams were generated for each file using *stylo* and a script in R (see Appendix). Finally, all the analyses were carried out using the *stylo* package in R (Eder, Rybicki, and Kestemont 2016).

3.3 Results

This section presents the results obtained with Zeta. As mentioned previously, all the analyses are restricted to the typical authors for each style, plus the case study, consisting of Herrera's editions. Poems from other Golden Age Spanish poets are not used since their role in the change of style could not be as clearly established as with the typical authors. In addition, since the interest here is in the more Renaissance or Baroque component of the studied collections, and not in the authorial signal, using more than one typical author for each style may be beneficial. Poems by the classic Renaissance authors, as selected above, were collected in a first dataset (primary set), while poems by typical Baroque authors were gathered in another one (secondary set). This way, the two sets are contrasted for the more characteristic words of each of them and to extract Renaissance and Baroque markers, applying Craig and Kinney's idea of authorial markers (Craig and Kinney 2009) to different stylistic trends. Lastly, Herrera's editions *H* and *P* were placed in a third set (test set) in order to see if, according to the selected features and resulting markers, each of them was closer to the Renaissance poets or to the Baroque ones. For this purpose, a two-dimensional plot was produced for every kind of analysis, carried out through the application of Craig's Zeta on 1,000 text slices.[10]

The first experiment uses the untagged corpus. In the resulting graph (see Figure 1), Renaissance poets form a pentagonal shape at the top left, while Baroque ones form another shape at the bottom right. *H* and *P2* appear at the center of the graph, between the Renaissance and the Baroque groups, but, most interestingly, some of their markers enter the Baroque shape created by Baroque markers and some are close to them. This seems to suggest that both editions are indeed in a middle ground between Renaissance and Baroque, albeit closer to the latter. In this sense, *P2* markers go in greater quantity and more deeply into the Baroque markers zone, implying a more Baroque component of *P2* poems.

Regarding the underlying features for this analysis, a plot confronting the list of *preferred* words by Renaissance against the Baroque and vice versa is at the Appendix (Figure 4), completed with tables which include English translations of top *preferred* 15 words for Renaissance (Table 1 in Appendix) and Baroque (Table 2 in Appendix). Words *preferred* by Renaissance are focused on unrequited love and how the poet experiences it through his pain, thoughts and feelings (*me*, *I feel*, *feelings*, *torment*, *thoughts*, *sad*, *love*, *hope*, *die*), whereas Baroque *preferred* words are far more open to the outside world. In this sense, we find terms related with metaphors and images to describe the woman's beauty (*gold*, *stars*), terms related with religion (*sin*, *piety*, *pilgrim*), political power (*king*, *Spain*, *tyrant*) and *tempus fugit* (*smoke*, *ashes*).

Besides that, special significance has been given to syntax as a distinguishable feature between these two stylistic trends, and even some Herrera's experts pointed out that in syntactical terms, *Versos* would be much closer to Baroque than *Algunas obras*. For this reason, the PoS tagged version of the corpus was also analyzed through the same method and parameters.

Firstly, the PoS bigram corpus was analyzed. As a result, a list of PoS-tag bigrams *preferred* by the primary set authors and *avoided* by the secondary set authors was produced, besides another list, in this case, of bigrams *avoided* by the primary set authors and *preferred* by secondary set authors. Thus, including typical Renaissance authors in the primary set and Baroque authors in the secondary set, the first list contains the PoS bigrams *preferred* by Renaissance authors and *avoided* by Baroque ones—in other words, it consists of the most characteristic Renaissance bigrams—, whereas the second list incorporates the PoS bigrams *preferred* by Baroque authors and *avoided* by the Renaissance—featuring the most characteristic Baroque bigrams. The top 15 PoS bigrams of each list can be found at the Appendix at Tables 3 (Renaissance) and 4 (Baroque), completed with examples of use. The top PoS bigrams of Renaissance and Baroque are fairly different. When comparing them, Renaissance seems to favor simpler structures:

[10] As poems are very short texts, this small value was selected in order to enable the analysis.

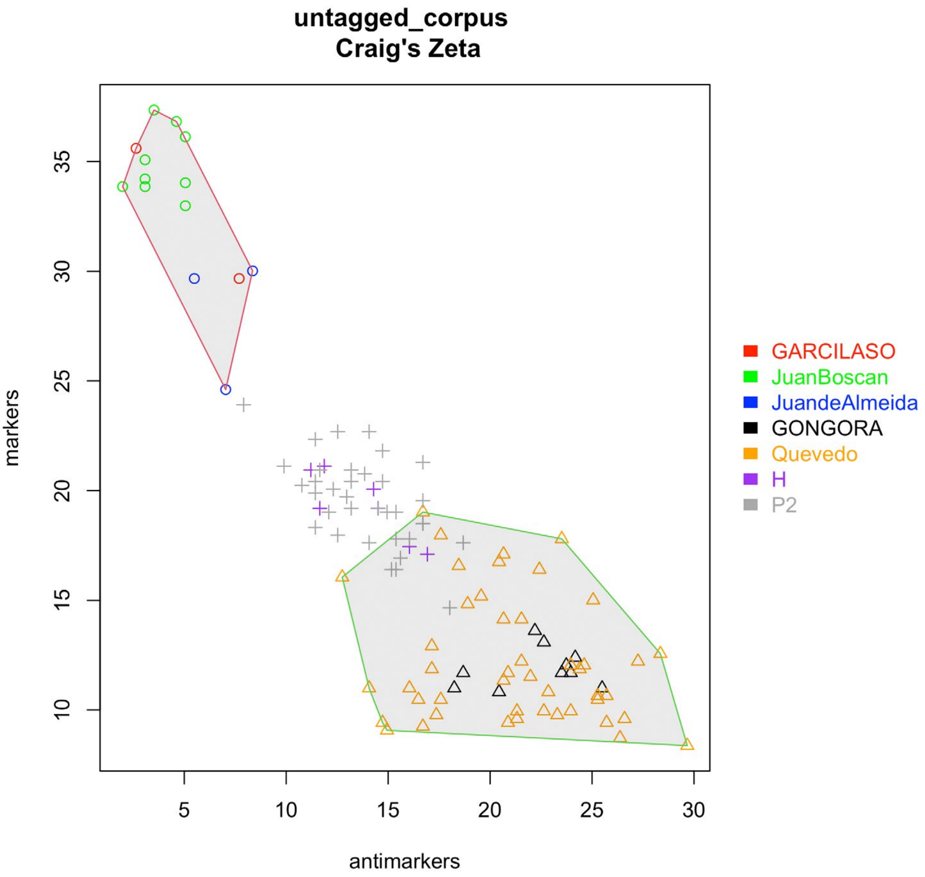


Fig. 1 *Oppose* (Craig's Zeta) using typical Renaissance authors (Garcilaso, Boscán and Juan de Almeida) in the primary set, typical Baroque authors (Quevedo and Góngora) in the secondary set, as well as *H* and *P2* in the test set (Hernández-Lorenzo, CC BY).

i.e., subordinating conjunction + demonstrative pronoun. In this sense, Baroque bigrams seem to suggest a more complex syntax: i.e. subordinating conjunction + adjective. It is also remarkable that most *preferred* Baroque bigrams include proper names, suggesting that these are more frequent in typical Baroque authors: i.e., proper noun + adjective, which comes out as the most characteristic Baroque PoS bigram structure.

Apart from the analysis of concrete features, *H* and *P2* were included in the test set, as in the previous analysis with the untagged corpus, and a visualization plot was generated (see Figure 2).

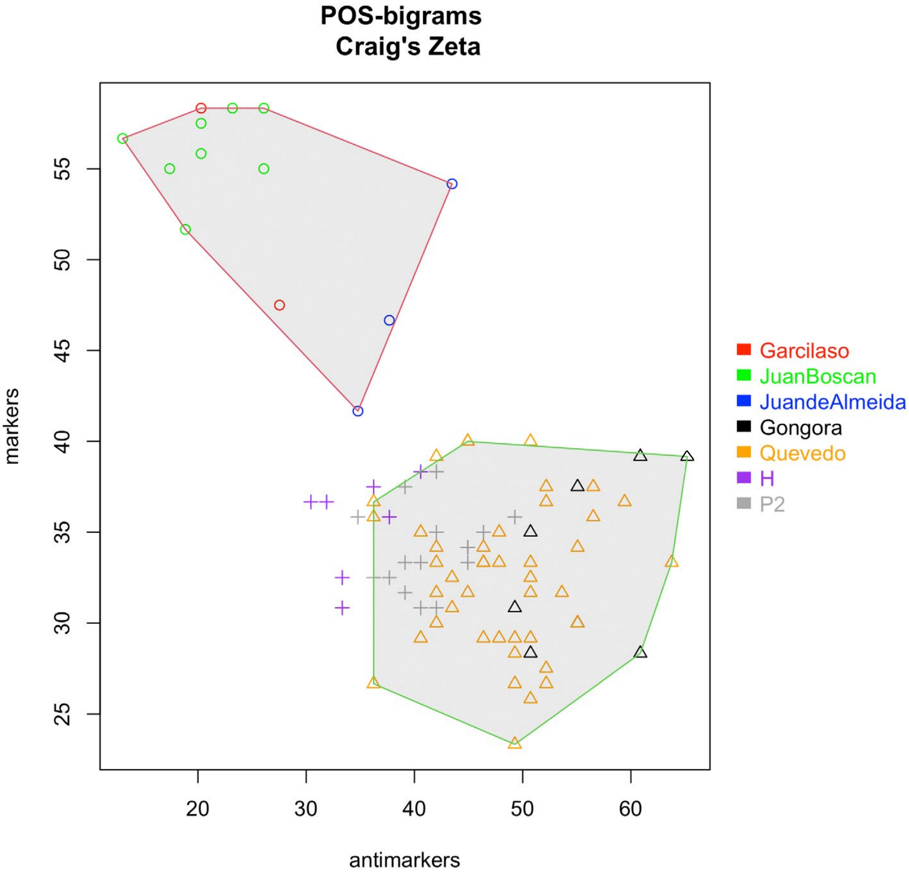


Fig. 2 Zeta on PoS bigrams (Hernández-Lorenzo, CC BY).

The results obtained are of great interest. When considering the PoS bigrams annotation, Herrera's editions *H* and *P* are not only stylistically close to Baroque, as we found out previously through lexical information (compare with Figure 1), but also syntactically. As a matter of fact, when their markers go into the Baroque zone, only a few of *H* get into the Baroque shape, whereas most of *P* markers appear there. Thus, in terms of PoS bigrams, *P* comes out as more Baroque than the edition published by Herrera.

After the previous analysis with PoS bigrams, the PoS trigram corpus version was analyzed. Once again, a list with *preferred* words by typical Renaissance authors (see top 15 at Table 5 in Appendix), and another one with the *preferred* ones by characteristic Baroque authors (see top 15 at Table 6 in Appendix) were produced. Top Renaissance

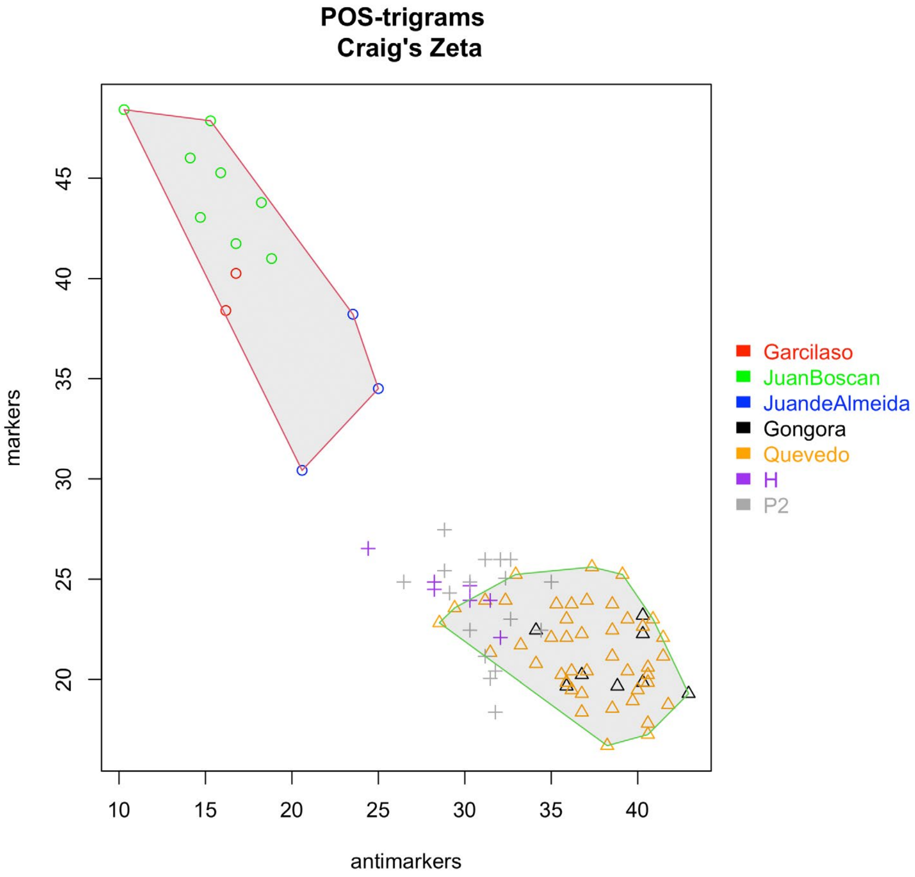


Fig. 3 Zeta on PoS trigrams (Hernández-Lorenzo, CC BY).

ones feel more natural and harmonious for a Spanish speaker: i.e., preposition + personal pronoun + personal pronoun. As for Baroque trigrams, as with bigrams, proper nouns are extremely frequent: i.e., main verb + article + proper noun, which is the preferred trigram by Baroque authors. They also include much more complex and artificial syntactic constructions than Renaissance ones: i.e., common noun + common noun + adjective.

Besides the analysis of concrete features, *H* and *P2* were added to the test set in order to analyze their closeness to Renaissance and Baroque markers, as done before with the untagged and the PoS-tag bigrams corpus, producing the following plot (see Figure 3).

As shown by Figure 3, once more the Renaissance authors and texts formed a shape on the top left of the graph, whereas the Baroque authors formed another one at the bottom right, smaller than in the previous one generated with bigrams (compare with Figure 2). *H* and *P* markers appear at the center of the graph between both clusters, as in previous analyses, but some of them go into the Baroque shape. The situation is perhaps not as clear as in the case of PoS bigrams, but again *P* markers increase and move toward the Baroque shape. Therefore, the Zeta result obtained with the PoS trigrams agrees with the outcome obtained through the comparison of lexical features and PoS bigrams.

4. Conclusions and Future Work

This paper presents a first exploratory analysis, applying methods of digital stylistics and stylometry to questions of periodization and style evolution in Golden Age Spanish poetry through the case of Fernando de Herrera's poems. A brief introduction on Golden Age Spanish poetry was presented, as well as the discussion about Herrera's transitional role between the Renaissance and the Baroque. With the aim of placing his poetic printed editions in the Golden Age period, some preliminary analyses have been conducted after carefully preparing the corpus of his poems and of typical Renaissance and Baroque Spanish poets. Using Zeta and across different features (words and PoS n-grams) and parameters (PoS bigrams and trigrams), the results seem to confirm the transitional role of the poems included in the editions published by Herrera, although both editions would be closer to Baroque typical style, as depicted in Góngora and Quevedo. Regarding the unique poems of the edition published after Herrera's death, contained in *P2* sub-corpus, results unanimously point to even a stronger Baroque component in them against *H* poems.

Naturally, this is a first approach to this question and future work would be needed before drawing definite conclusions. In this sense, further exploration of these first attempts and results is required, as well as a more comprehensive study and interpretation of the decisive features for Renaissance and Baroque obtained with Zeta and their relation to Herrera's position in the change of period and style. Nevertheless, the present paper serves as an example of applications of stylometric techniques to the exploration of style evolution in poetry across different literary movements, and especially on the change from the Renaissance to the Baroque.

Finally, the results obtained in this study on Herrera's poems show that it is worthwhile to apply techniques from digital humanities and stylometry to literary historical questions in general, and studies of poetic and Spanish texts in particular, and that they open up fascinating venues of research.

Appendix

Complementary materials to this study, such as the corpus, tables with *preferred* words / PoS-tag n-grams of each period, and code used, can be found at this GitHub repository: https://github.com/lamusadecima/Digital-Stylistics-Applied-to-Golden-Age

Acknowledgments

The idea for this project came up during my research stay at Julius-Maximilians University in Würzburg (Germany) in fall 2018. I thank Fotis Jannidis and José Calvo for their helpful comments and suggestions, as well as the anonymous reviewers for their valuable insights to improve this paper.

Funding

As previously stated, this project benefited from a research stay at the University of Würzburg, supervised by Fotis Jannidis and supported by an Erasmus Traineeship Fellowship. Last stages of this study were carried out within the project *Hacia la Institucionalización Literaria: Polémicas y Debates Historiográficos* (1500–1844) (RTI2018-095664-B-C22), funded by the Spanish Ministry of Science (*Ministerio de Ciencia, Innovación y Universidades*).

ORCID®

Laura Hernández-Lorenzo https://orcid.org/0000-0003-3489-2193

References

Battaglia, Salvatore. 1954. "Per Il Testo Di Fernando de Herrera." *Filologia Romanza* I: 51–88.

Blasco, Javier. 2016. "Avellaneda Desde La Estilometría." In *Cervantes. Los Viajes y Los Días*, edited by Pedro Ruiz Pérez, 97–116. Madrid: Prosa Barroca y SIAL Ediciones.

Blecua, José Manuel. 1958. "De Nuevo Sobre Los Textos Poéticos de Herrera." *Boletín de La Real Academia Española* XXXVIII: 377–408.

Blecua, José Manuel. 1975. "Introducción." In *Obra Poética*, by Fernando de Herrera, 11–78. Madrid: Boletín de la Real Academia Española.

Burrows, John. 2002. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17 (3): 267–87.

Burrows, John. 2007. "All the Way Through: Testing for Authorship in Different Frequency Strata." *Literary and Linguistic Computing* 22 (1): 27–47.

Calvo Tello, José. 2019. "Delta Inside Valle-Inclán: Stylometric Classification of Periods and Groups of His Novels." In *Theorien von Autorschaft und Stil in Bewegung: Stilistik und Stilometrie in der Romania*, edited by Nanette Rißler-Pipka, 151–64. München: AVM edition.

Calvo Tello, José. 2021. The Novel in the Spanish Silver Age. A Digital Analysis of Genre Using Machine Learning. Bielefeld: transcript.

Calvo Tello, José, and Juan Cerezo Soler. 2018. "La Conquista de Jerusalén ¿ de Cervantes ? Análisis Estilométrico Sobre Autoría En El Teatro Del Siglo de Oro Español." *Digital Humanities Quarterly* 12 (1): 1–10.

Craig, Hugh, and Arthur F. Kinney, eds. 2009. *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Cuevas García, Cristóbal. 1985. "La Cuestión Textual." In *Poesía Castellana Original Completa*, edited by Cristóbal Cuevas García, 87–99. Madrid: Cátedra.

Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. "Stylometry with R: A Package for Computational Text Analysis." *R Journal* 8 (1): 107–21.

Fradejas Rueda, José Manuel. 2016. "El Análisis Estilométrico Aplicado a La Literatura Española: Las Novelas Policiacas Históricas." *Caracteres. Estudios Culturales y Críticos de La Esfera Digital* 5 (2): 196–246.

Fradejas Rueda, José Manuel. 2019. "Estilometría y La Edad Media Castellana." In *Theorien von Autorschaft und Stil in Bewegung: Stilistik und Stilometrie in der Romania*, edited by Nanette Rißler-Pipka, 49–74. München: AVM edition.

García-Reidy, Alejandro. 2019. "Deconstructing the Authorship of *Siempre Ayuda La Verdad*: A Play by Lope de Vega?" *Neophilologus* 103: 493–510.

Hernández Lorenzo, Laura. 2019a. "Fernando de Herrera y La Autoría de Versos. Un Primer Acercamiento al Drama Textual Desde La Estilometría." In *Theorien von Autorschaft und Stil in Bewegung: Stilistik und Stilometrie in der Romania*, edited by Nanette Rißler-Pipka, 75–90. München: AVM edition.

Hernández Lorenzo, Laura. 2019b. "Poesía Áurea, Estilometría y Fiabilidad: Métodos Supervisados de Atribución de Autoría Atendiendo al Tamaño de Las Muestras." *Caracteres. Estudios Culturales y Críticos de La Esfera Digital* 8 (1): 189–228.

Hernández Lorenzo, Laura. 2020. *Los Textos Poéticos de Fernando de Herrera: Aproximaciones Desde La Estilística de Corpus y La Estilometría*. Sevilla: Universidad de Sevilla. https://idus.us.es/handle/11441/93465.

Hernández Lorenzo, Laura. 2022. "Stylistic Change in Early Modern Spanish Poetry Through Network Analysis (with an Especial Focus on Fernando de Herrera's Role)." *Neophilologus* 106: 397–417.

Herrera, Fernando de. 1975. *Obra Poética*, edited by José Manuel Blecua. Madrid: Boletín de la Real Academia Española.

Kossoff, David. 1957a. "Algo Más Sobre *Largo-Luengo* En Herrera." *Revista de Filología Española* XLI: 401–10.

Kossoff, David. 1957b. "Algunas Variantes de Versos de Herrera." *Nueva Revista de Filología Hispánica* XI: 57–63.

Kossoff, David. 1965. "Another Herrera Autograph: Two Variant Sonnets." *Hispanic Review* 33 (3): 318–25.

Kossoff, David. 1966. *Vocabulario de La Obra Poética de Herrera*. Madrid: RAE.

Lescasse, Marie-Églantinne. 2019. "Góngora Hors Norme? Étude Stylométrique d'un Motif Gongorin." In *Theorien von Autorschaft und Stil in Bewegung: Stilistik und Stilometrie in der Romania*, edited by Nanette Rißler-Pipka, 91–116. München: AVM edition.

López Bueno, Begoña. 2000. *La Poética Cultista de Herrera a Góngora*. Sevilla: Alfar.

López Bueno, Begoña, ed. 2006. *La Renovación Poética Del Renacimiento al Barroco*. Madrid: Síntesis.

Macrí, Oreste. 1959. "Autenticidad y Estructura de La Edición Póstuma de 'Versos' de Herrera." *Filologia Romanza* VI: 1–26 and 151–84.

Macrí, Oreste. 1972. *Fernando de Herrera*. Madrid: Gredos.

Navarro-Colorado, Borja. 2017. "A Metrical Scansion System for Fixed-Metre Spanish Poetry." *Digital Scholarship in the Humanities* 33 (1): 112–27.

Navarro-Colorado, Borja. 2018. "On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry." *Frontiers in Digital Humanities* 5. https://www.frontiersin.org/article/10.3389/fdigh.2018.00015.

Navarro-Colorado, Borja, María Ribes Lafoz, and Noelia Sánchez. 2016. "Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation." *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference (LREC 2016),* 4360–64.

Padró, Lluís. 2011. "Analizadores Multilingües en Freeling." *Linguamática* 3: 13–20.

Padró, Lluís, and Evgeny Stanilovsky. 2012. "FreeLing 3.0: Towards Wider Multilinguality." *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2012)*, 2473–2479.

Pepe Sarno, Inoria. 1981. "Bianco Il Ghiaccio, Non Il Velo: Ritocchi e Metamorfosi in Un Sonetto Di Herrera." *Strumenti Critici* XLVI: 458–71.

Pepe Sarno, Inoria. 1982. "Se Non Herrera, Chi? Varianti e Metamorfosi Nei Sonetti Di Fernando de Herrera. I Parte." *Studi Ispanici*: 33–69.

Pepe Sarno, Inoria. 1986. "Se Non Herrera, Chi? Variante e Metamorfosi Nei Sonetti Di Fernando de Herrera. V Parte." *Studi Ispanici*: 21–59.

Pepe Sarno, Inoria. 1998. "La Negazione Del Colore: 'Oh, Soberbia y Cruel En Tu Belleza'". In *Signoria Di Parole. Studi Offerti a Mario Di Pinto*, edited by G. Calabrò, 393–402. Napoli: Liguori Editore.

Plecháč, Petr, Klemens Bobenhausen, and Benjamin Hammerich. 2018. "Versification and Authorship Attribution. A Pilot Study on Czech, German, Spanish, and English Poetry." *Studia Metrica e Poetica* 5 (2): 29–54.

Rißler-Pipka, Nanette. 2016a. "Digital Humanities und die Romanische Literaturwissenschaft – Der Autorschaftsstreit um den *Lazarillo de Tormes*." *Romanische Forschungen* 128 (3): 316–42.

Rißler-Pipka, Nanette. 2016b. "Avellaneda y los problemas de la identificación del autor. Propuestas para una investigación con nuevas herramientas digitales." In *El otro Don Quijote: La continuación de Fernández de Avellaneda y sus efectos*, edited by Hanno Ehrlicher, 27–51. Augsburg: Institut für Spanien-, Portugal- und Lateinamerika-Studien (ISLA).

Rißler-Pipka, Nanette. 2019. "In Search of a New Language. Measuring Style of Góngora and Picasso." In *Theorien von Autorschaft und Stil in Bewegung: Stilistik und Stilometrie in der Romania*, edited by Nanette Rißler-Pipka, 117–150. München: AVM edition.

Rojas Castro, Antonio. 2017. "Luis de Góngora y La Fábula Mitológica Del Siglo de Oro: Clasificación de Textos y Análisis Léxico Con Métodos Informáticos." *Studia Aurea* 11: 111–42.

Rojas Castro, Antonio. 2018. "¿Cuántos 'Góngoras' Podemos Leer? Un análisis contrastivo de la poesía de Luis de Góngora." E-Spania 29: 1–19. https://doi.org/10.4000/e-spania.27448.

Rosa, Javier de la, and Juan Luis Suárez. 2016. "The Life of Lazarillo de Tormes and of His Machine Learning Adversities. Non-traditional authorship attribution techniques in the context of the Lazarillo." *Lemir* 20: 373–438.

Rozas, Juan Manuel. 1983. "'Siglo de Oro': historia y mito." In *Historia y Crítica de La Literatura Española 3. Siglos de Oro, Barroco*, edited by Aurora Egido, 64–67. Barcelona: Crítica.

Ruestes, María Teresa. 1986. "Introducción." In *Poesía*, by Fernando de Herrera, ix–lxiv. Barcelona: Planeta.

Ruiz Fabo, Pablo, Clara Isabel Martínez Cantón, Thierry Poibeau, and Elena González-Blanco. 2017. "Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets." *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 27–32.

Schöch, Christof. 2017. "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Schöch, Christof. 2018. "Zeta für die kontrastive Analyse literarischer Texte – Theorie, Implementierung, Fallstudie." In *Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven*, edited by Toni Bernhart, Sandra Richter, Marcus Lepper, Marcus Willand, and Andrea Albrecht, 77–94. Berlin, Boston: De Gruyter.

Ulla Lorenzo, Alejandra, Elena Martínez Carro, and José Calvo Tello. 2021. "Las Comedias de Dudosa Atribución de Agustín Moreto: Nuevas Perspectivas Estilométricas." *Neophilologus* 105: 57–73.

Valbuena Prat, Ángel. 1960. *Historia de La Literatura Española*. Vol. I. Barcelona: Gustavo Gili.

Vilanova, Antonio. 1951. "Fernando de Herrera." In *Historia general de las literaturas hispánicas II: Pre-renacimiento y renacimiento*, directed by Guillermo Díaz-Plaja, 689–751. Barcelona: Barna.

Cross-Language Stylometry
Picasso's Writings in Spanish and French

Nanette Rißler-Pipka

Abstract For multilingual corpora—and in this particular case for the Spanish and French writings of Pablo Picasso—we do not have an acceptable method for quantitative literary analyses. This paper discusses the existing possibilities to solve the problem of cross-language stylometry by comparing the results of single-language stylometric methods but does not present a new method. Further on, the hypothesis that Picasso's writings and poetry are characterized by a unique style is tested with a Spanish and French corpus, focusing on both the semantic and syntactic similarities and differences. In comparison to contemporary writers, the difference and distinctiveness can be shown by cross-language analyses.

Keywords Picasso, stylometry, part-of-speech, Romance literature

1. Introduction

The example of one of the most known artists who wrote half of his experimental texts in his native language, Spanish, and half of them in the language of his exile country, French, helps to illustrate one of the most urging problems for digital stylistics in Romance studies: cross-language analysis. If we don't want to use translations (and this is out of question for literary Romance studies), the only way to compare a multilingual text collection is to separate the texts by language and compare them to texts of contemporary writers of the same language. The research in stylometry done in this field so far is testing the functionality of different Deltas with different languages (Jannidis et al. 2015; Rybicki and Eder 2011). This is very important to know before trying stylometry with the same parameters (Deltas) on corpora in different languages. Though, it does not help treating the problem of stylistics for authors writing in two (or more) languages. Juola and Mikros started to think about "cross-linguistic stylometric features" (2016) and analyzed a corpus of tweets by bilingual Twitter users (Spanish and

Rißler-Pipka, Nanette: Cross-Language Stylometry. Picasso's Writings in Spanish and French, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 53–80. DOI: https://doi.org/10.17885/heiup.1157.c19366

English). The results, though, cannot easily be adopted on the question of literary style: They suggest “cross-linguistic similarities” (Juola and Mikros 2016) based on features like length of tweet and word-length. One might argue that this is due to the media Twitter that forces users to be short in every language. The similarities in French and Spanish texts by Picasso are probably proven by different style markers which are to be detected initially by close reading.

Since Picasso started to publish part of his writings (mostly his two plays) in the late 1930s and 1940s, readers and critics tried to catalogize or to cluster them into literary history (Leiris, 1966; Éluard, 1934; Sabartés, 2017). His exceptional style was compared to James Joyce, Paul Éluard, Guillaume Apollinaire, Federico García Lorca, Rafael Alberti, Góngora, Mallarmé and other authors known for innovative avant-garde-like style (Heydenreich 1979; Fernández Molina 1988; Béhar 1993; Goddard 2006; Michaël 2008; Rißler-Pipka 2015). Several papers have been published about stylometry on James Joyce (O’Sullivan, Eder, and Rybicki 2018; O’Sullivan 2014; Clement 2013), but fewer about the authors of avant-garde literature in Romance languages (Calvo Tello 2019). While using close reading as a method of stylistics it is possible to compare texts by examples of phrases, motives, etc. If necessary and well-argued this comparison may combine several languages, periods and genres, e.g. to speak about related works like the plays by Alfred Jarry and Picasso’s prose poem on Guernica or about the baroque mannerist poems by Góngora and Picasso’s twentieth-century avant-garde poetry (concrete examples in Rißler-Pipka 2015, 369–71. (Jarry) and 261–63 (Góngora)). Digital stylistics needs, out of mere technical and algorithmic reasons, a common ground for comparison: same language, same period, same genre. On the one hand that seems to be a disadvantage: if you want to compare the ideas Picasso took from Mallarmé to write his Spanish texts, the intertextual links are hidden in the difference of languages. Certainly, some influences are measurable (like the abandonment of punctuation which Picasso indeed borrowed from Mallarmé) and some are only traceable in close reading, because they represent the re-use of broader ideas (like the surrealist combination of disparate elements). On the other hand, stylometry has the advantage to free itself from the ever cited canon and to compare far more texts in a quantitative way (even if we know that selection process is still to be considered). Previous work (Rißler-Pipka 2019a; 2019b) showed rough stylistic resemblance between the French writings of Picasso and Raymond Roussel (*Nouvelles impressions d’Afrique* [1901]) and Ramón Gómez de la Serna for the Spanish writings. Roussel was named by Michaël (2012, 166–67) as possible model but both authors would not necessarily have been on the list of candidates for influences before the quantitative analysis (shown in Rißler-Pipka 2019a; 2019b). However, a direct comparison to the two-language corpus of Picasso’s writings alone is still missing and will be discussed in this paper.

2. Hypothesis

Picasso wrote over 200 Spanish and over 300 French prose poems. His very particular style of writing is hard to describe and even harder to catalog as part of literary history. The mere label of *prose poem* can be doubted, but we are missing the accurately fitting description. I do not completely agree with the concept provided by Enrique Mallen (Mallen 2009; 2012; Mallen and Meneses 2019, for a discussion on that point see Rißler-Pipka 2015, 40–42) that Picasso repeats his cubist painting in his writing. Nevertheless, comparing Picasso's writing with the technique of montage (also cited as "rhizomatique" by Michaël 2012, 163) is very tempting. Similar to geometrical forms in cubism, Picasso uses words or grammatical elements in poetry to build new images out of the very same things. By close reading we come to the hypothesis that in both languages, Spanish and French, we find the same technique, pattern, and system. If we follow up this hypothesis, we have to compare the prose poems not only regarding their content and semantics, but also regarding the syntax, word distribution, and other stylistic markers. At this point cross-language stylometry comes up.

Still, using stylometry as a method means to compare quantitatively a set of texts based on style markers (features) like word frequencies, etc. and it is not possible to compare texts of different languages in the same experiment. All efforts done in this direction end up with translations which necessarily influence the results and make them less robust (Heydel and Rybicki 2012). There are more premises than language for authorship contribution with stylometry (e. g. genre, period, etc.), but even experimenting with genre- and/or period-mixes while checking the results for consistency, methods like stylometry with R (Eder et al. 2013) do not accept two different languages. We can demonstrate the language difference also with an example in Spanish and French: The number of words for expressing "I think", in Spanish *pienso* and French *je pense*, are different because in Spanish no personal pronoun is necessary and the form of the verb expresses the case. The only way to compare the very same author in two languages is to be aware of these differences when using tools like the R package *stylo*, with the same parameters to compare the results in the end. The evaluation and interpretation of the results have to take this into consideration.

When Rybicki and Eder wrote about "cross-language authorial fingerprint" (2011) and Eder about "Delta across languages" (2011) they tested style markers for different languages to check if the most frequent words (MFW) are the best feature for authorship attribution—even for other languages than English. They prove that for French MFW still are good style markers. However, this is not true for every genre: English novels get higher rates than English epic poetry. If we agree on comparing the results of stylometry on the French and Spanish corpus we should be very careful because of the given difference of number of words. If *yo* ('I') in Spanish is only used to underline an expression (e.g., *yo, también* 'Me too') but usually not necessarily in combination of

Table 1 Example for grammatical difference in Spanish and French

Spanish	POS Spanish	French	POS French
hago ('I do')	"**o**" = 1st pers. sing. = POS: vmi	**Je** *fais* ('I do')	Personal pronoun (subject) + verb (1st pers. sing.) = POS: pro + v

Dame el papel / *Damelo* ('Give me the paper' / 'Give it to me')	"**a**" = 3rd pers. sing. (Imperative);
	"me" = personal pronoun;
	"lo" = direct object = POS: vmm + pp + pp
		Donne-moi le papier / *Donne-le moi* ('Give me the paper' / 'Give it to me')	Verb (3rd pers. sing.);
			personal pronoun;
			direct object = POS: v + pro + pro

verb and personal pronoun this may influence the number of words. In consequence, the MFW count in a French and Spanish corpus will differ enormously. Furthermore, in Spanish object pronouns (e.g., *lo*, *la* 'it') are attached to the verb (see Table 1). This makes it even more difficult to compare the results of a method that counts and calculates the distribution of MFW in a corpus. For stylometry we keep MFW as dominant style marker because they proved to be the feature that needs less presumptions than others (Evert et al. 2017; Byszuk and Eder 2019).

In order to know how these differences will affect the results and the comparability of the corpora we have to test it with different methods like frequency of MFW, but also content words or parts of speech (POS) sequences. For choosing one of these methods, it is necessary to look at the kind of differences and similarities we would like to prove or test in Picasso's writings.

From 1935 on, Picasso wrote many small and some larger pieces of prose poems and two French plays during the time of German occupation in Paris 1942–44 (Table 2).

Astonishingly, the total sum of words in his Spanish and French writings over the whole period of about 30 years is quite similar. Despite the grammatical differences between Spanish and French (as shown above), the sums of words are very close. In this calculation we did not count the plays and we need to point out that Picasso only wrote in French in 1938, 1941–54, and in 1956 but wrote one of his longest and most known fragment of poetry and drama, *El Entierro del Conde de Orgaz*, in 1957–59.

Picasso had moved to Paris, the European cultural capital at this period, in 1901 to pursue his career. He stayed, became very quickly a famous artist of the avant-garde and was forced to stay when the civil war in Spain began in 1936. Picasso, himself, thought much about communication across languages. His rather well-known prose poem on translation proves this aspect:

> Si je pense dans une langue et que j'écris "le chien court derrière le lièvre dans le bois" et veux le traduire dans une autre, je dois dire "la table en bois blanc

Table 2 Overview of the complete writings of Pablo Picasso in Spanish and French

SPANISH			FRENCH		
Date	Texts	Words	Date	Texts	Words
1935	82	25,377	1935	52	9,454
1936	35	6,896	1936	98	13,233
1937	11	1,581	1937	46	8,853
1938	0	0	1938	29	6,284
1939	2	61	1939	12	3,180
1940	61	19,376	1940-41	44	7,035
1941-54	0	0	1942-44	17 (+2 plays)	2,860 (+18,080)
1955	2	174	1945-46	0	0
1956	0	0	1947-61	48	5,619
1957-59	1	5,966	-	-	-
1958-68	6	1,374	-	-	-
Total	**200**	**60,805**	**Total**	**346 (+2)**	**56,518 (+ plays=74,598)**

> enfonce ses pattes dans le sable et meurt presque de peur de se savoir si sotte" (Picasso, October 28, 1935).
>
> "If I think in a language and write: 'The dog runs after the rabbit in the wood' and want to translate it in another language, I have to say 'The table of white wood marks its paws in the sand and nearly dies of fear knowing himself that silly'" (translation by N. R.-P.).

Here, he sticks to French, but still tells us something about translation without translating one language to another—how is this possible? There are traces of the first sentence in the second one: *chien* 'dog'—*pattes* 'paws'; *bois* 'wood'—*bois* 'wood'; *court* 'runs'—*enfonce* 'marks') and some grammatical POS are repeated like the articles and the prepositions (*le, la, dans* 'it', 'in'). Both sentences have an abstract poetic meaning which is produced by using rather concrete expressions in weird combination.[1] However, we can draw at least two possible readings out of the translation-example by Picasso:

1. Translating from one language to another means, particularly in the case of poetry, new sense, new style, all in all it means a new creation.

[1] It might be interesting to test the distribution of concrete and abstract words in Picasso's writings following the example of the ongoing study by Ryan Heuser (2020).

2. Translation produces new meanings, new poetry which might only hint to the *original* without representing the original.

Based on the results of a close-reading analysis of the two corpora in French and Spanish we cannot say that Picasso just translates what he is thinking from one language to another but that he is experimenting with the same thought and vocabulary in each language (Michaël 2008; Rißler-Pipka 2015, 383). The result is probably as ill-suited for comparison as the two sentences Picasso gave as an example for translation (see above). The only chance to find out is by testing his writings and comparing the corpora regarding the following hypotheses based on previous close-reading analyses:

1. Picasso uses the same system of deconstructing language in both languages, Spanish and French.
2. It is not translation from one language to another that influences his style or makes it inventory—but the transformation of ordinary language into poetic *Picasso-language*.
3. The effect of chaos and semantic non-sense is produced by a recognizable system of grammatical repetition.

3. Analyses

After formulating hypotheses during close reading, we want to operationalize them in a quantitative computational analysis. Beforehand, we don't know if exemplary findings like extraordinary grammatical constructions or vocabulary will hold up to a quantitative testing. The hypotheses indicate the features that should be analyzed. For the first hypothesis (1) Picasso uses *the same system* in both languages, the *system* needs to be described more precisely. In Picasso's case this can be an obsessively repeated set of *vocabulary* which is easily comparable between two roman languages like Spanish and French. Stylometry on the basis of MFW illuminates the inner structure of authorial style and permits a comparison of different authors. This would support also hypothesis (2) saying that *Picasso-language* is detectable by this method. For the third hypothesis (3) the grammatical repetition can be detected by POS tagging and the comparison of POS sequences (n-grams). The latter is only one possibility out of several but is precisely drawn out of the observation in close reading, arguing that Picasso prefers some POS sequences, like prepositions (*de*, *que*) and noun (Rißler-Pipka, 2015, 270; Rißler-Pipka 2019a and b).

3.1 Vocabulary

To give a first overview of the French and Spanish corpus of Picasso's writings, I created a word cloud of the most frequent content words using Voyant Tools (Sinclair and Geoffrey 2016) and applying a self-defined stop-word list. The effect is even stronger because Picasso does not write narrative texts but prose poems. Readers familiar with the texts recognize the typical motifs which supports the hypothesis that Picasso uses mostly a fixed set of vocabulary (Michaël 2008; Rißler-Pipka 2015, 383).

Figures 1 and 2 are not very surprising if you know some of Picasso's texts and works of art: they are dominated by different words describing color, light, night and body parts. The famous bullfight topic is hidden because *toro* ('bull') is much less used than *tarde* ('afternoon'). From the previous close reading and the historical/cultural context we know they belong together because the bullfights were always scheduled in the afternoon. Comparing both word clouds' direct translations in French and Spanish

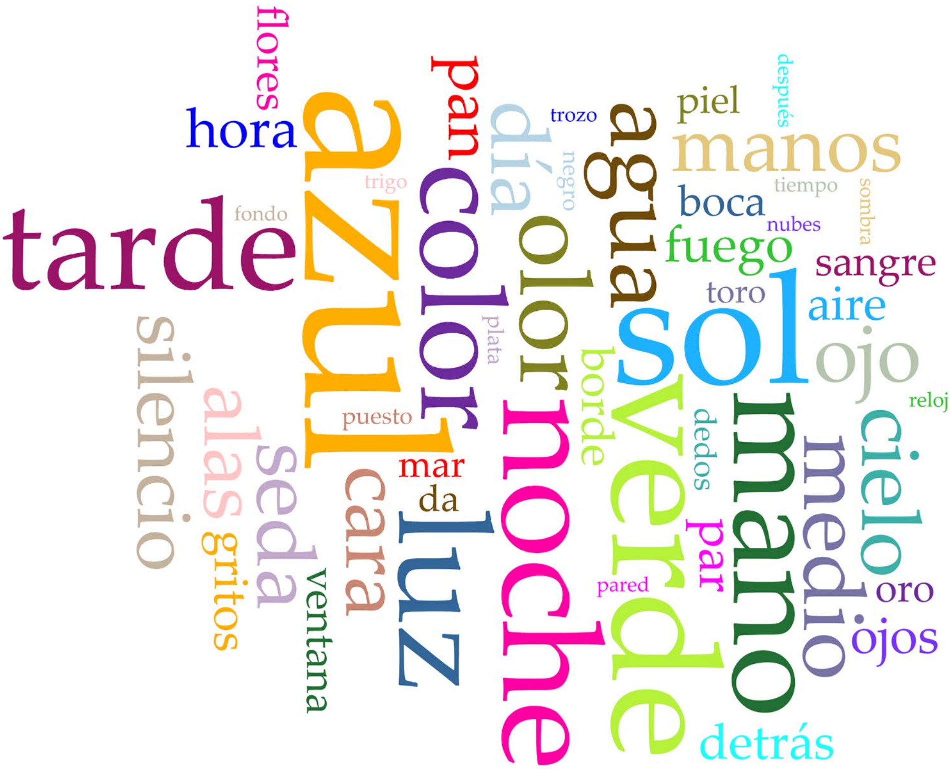


Fig. 1 Word Cloud for the Spanish texts (Rißler-Pipka, CC BY).

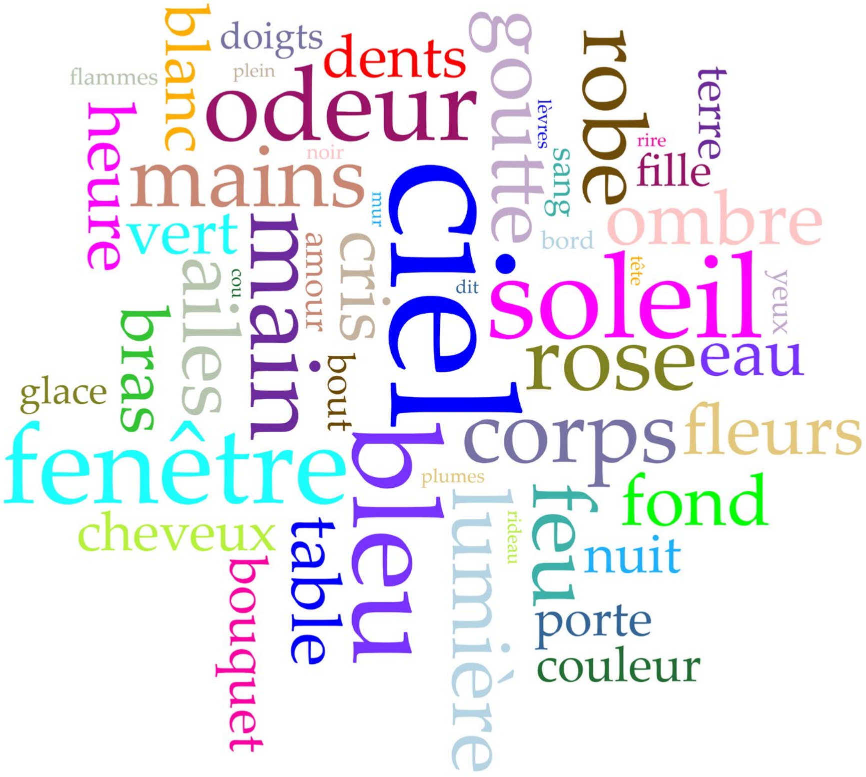


Fig. 2 Word Cloud for the French texts (Rißler-Pipka, CC BY).

are striking, for example for the colors (*bleu*/*azul* 'blue' *vert*/*verde* 'green'), for body parts (*main*/*mano;* 'hand'), for sensations (*odeur*/*olor* 'smell'; *cris*/*gritos* 'cries') and motifs (*feu*/*fuego* 'fire'; *ailes*/*alas* 'wings'). However, there are some differences in frequency we can show when Picasso speaks about the same things in Spanish and French prose poems. For example, French *ciel* ('sky') is used more often than Spanish *cielo* ('sky'), but *azul* ('blue') more often than *bleu* ('blue') because they belong together and are part of the very same semantic field. In direct comparison via translation or comparison of the bare numbers of most frequent content words this difference would not become visible (for specific semantic concepts in both languages see Meneses and Mallen 2019).

3.2 Stylometry

We want to look deeper into stylometry to be able to compare Picasso internally: what happens when he switches from one language to another? What are the main findings when comparing Picasso and his contemporaries in Spanish and French literature (1890–1960; poems or prose poems in avant-garde literature). For previous work I created a corpus of French and Spanish literature for the period of Picasso's life and writing (Rißler-Pipka 2019 a and b). For this study I used a reduced text collection of a group of authors known to be stylistically nearer to Picasso. The prose poems by Picasso were also sliced chronologically into equal sized parts to make them comparable to the length of common anthologies of poetry. The two plays *Le désir attrapé par la queue* (1941) and *Les quatre petites filles* (1948) were left out.

— Spanish corpus: 31 documents, sample size: 3,000–8,000, ∑ 157,863 tokens
— French corpus: 20 documents, sample size: 6,000–30,000, ∑ 290,131 tokens

According to the sample size and previous experiments and recommendation the chosen parameters for the initial analyses are: 2,500 MFW and the Wurzburg cosine delta (see Jannidis et al. 2015). The composition of the corpora is definitely not completely well balanced and the selection should be more precisely discussed. In the French corpus, Picasso's writings (apart of the plays) are split into seven more or less equal sized parts. The only other author who needed splitting up was Apollinaire: the well-known poetry anthologies *Calligrammes* and *Alcools* were split up in two and combined with his two erotic novels to have the prose genre represented as well. Éluard is represented in the French corpus with two texts (*Poèmes 1913–1940* and *La capitale de la douleur*). Unfortunately, for each of the other authors (Laurémont, Lourys, Segalèn, Mallarmé, Toulet) I found only one text per author and neither of these was long enough to be split up. More texts were either not available in a digital form or not suitable for comparison. For the Spanish corpus, the disproportionate number of texts by Picasso is even more striking because the sample size needed to be smaller in order to make it comparable to the other texts of the collection. Therefore, the Spanish corpus consists of ten texts by Picasso, six texts by Machado, four texts by Lorca, two by Alberti and two by Gómez de la Serna. Furthermore, seven authors are only represented with one single text.

One difficulty was being aware of the genre mixture in the avant-garde literature which might have negative influence toward the author signal. Another problem was the selection and accessibility of digital texts available for this period. However, for the rather detailed question if Picasso is using the same technique in two different languages, the representativeness of the corpus is not as important as for more advanced research questions.

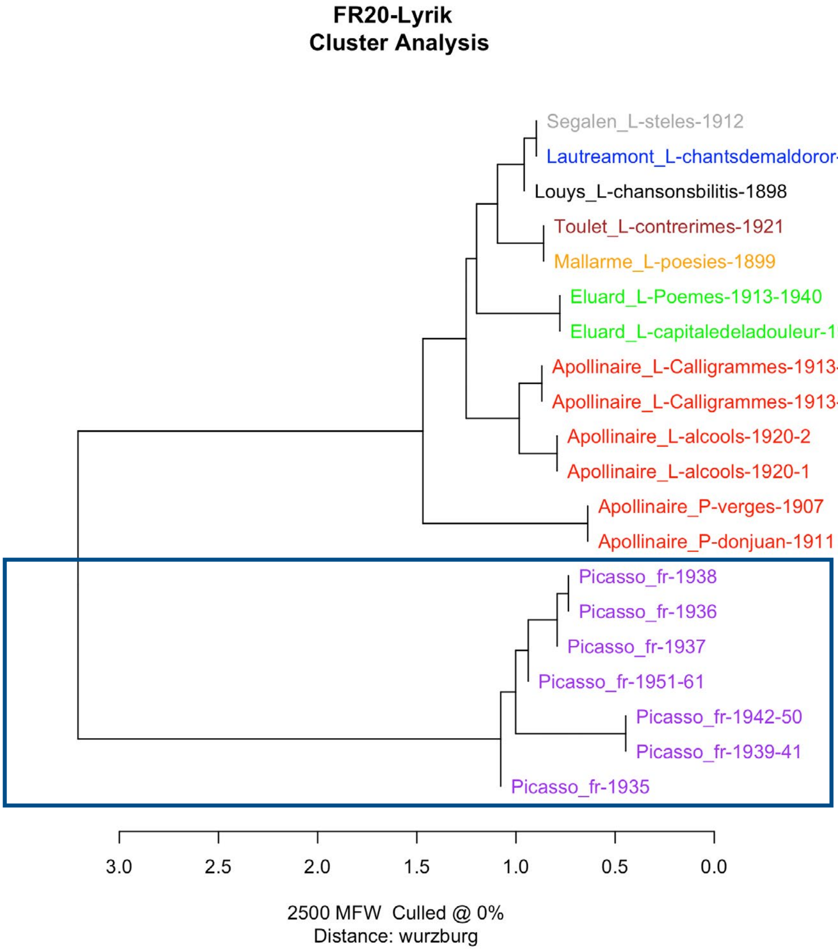


Fig. 3 Dendrogram for Cluster Analysis of the French texts (Rißler-Pipka, CC BY).

In Figure 3, Picasso's French writings take an own branch in the dendrogram and are not clustered with the surrealistic poetry or prose (or prose poetry) of the same period. The genre signal shows here as much effect as the author signal. Apollinaire is neatly divided in prose (*Les exploits d'un jeune Don Juan*; *Les onze mille verges*) and poetry (*Calligrammes; Alcools*) as well as the single authors who are more or less separated along the genre criteria. Picasso's writings are nearly all lined up according to the dates of creation: Picasso_Poesia1, 2 and 5 (1935–36); 4 and 3 (1935); 7 and Parchemin (1940) and the rest 6, 8 and the Entierro (1936–1957) are clustered

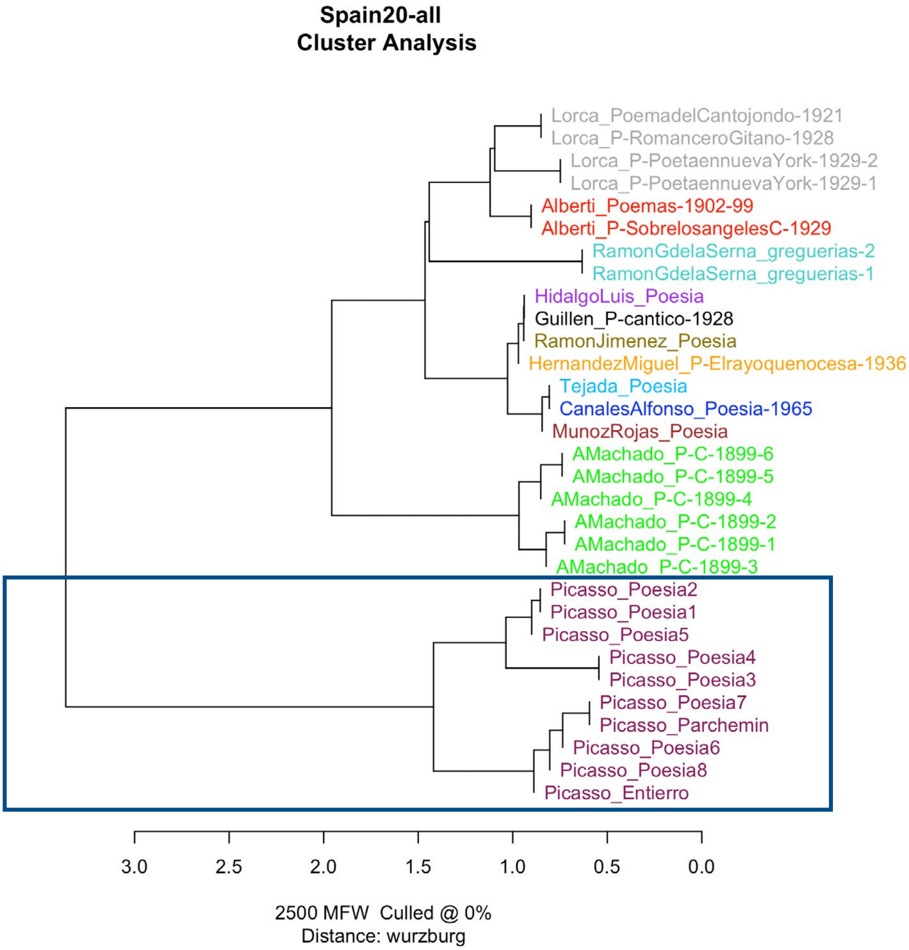


Fig. 4 Dendrogram for Cluster Analysis of the Spanish texts (Rißler-Pipka, CC BY).

together. In direct comparison to the Spanish corpus, we observe a similar effect as shown in Figure 4.

From previous analysis, I would expect that Picasso might cluster together with Ramón Gómez de la Serna but again (cf. Figure 4) his writings are completely separated from the rest of the corpus. Another outlier is—as expected—Antonio Machado because his poetry represents more the style of *generación 98* than surrealism and avant-garde. Regarding the motifs and subjects, we know from close reading that there is a similarity between Picasso and the upper group (Lorca, Gómez de la Serna, Alberti and particularly

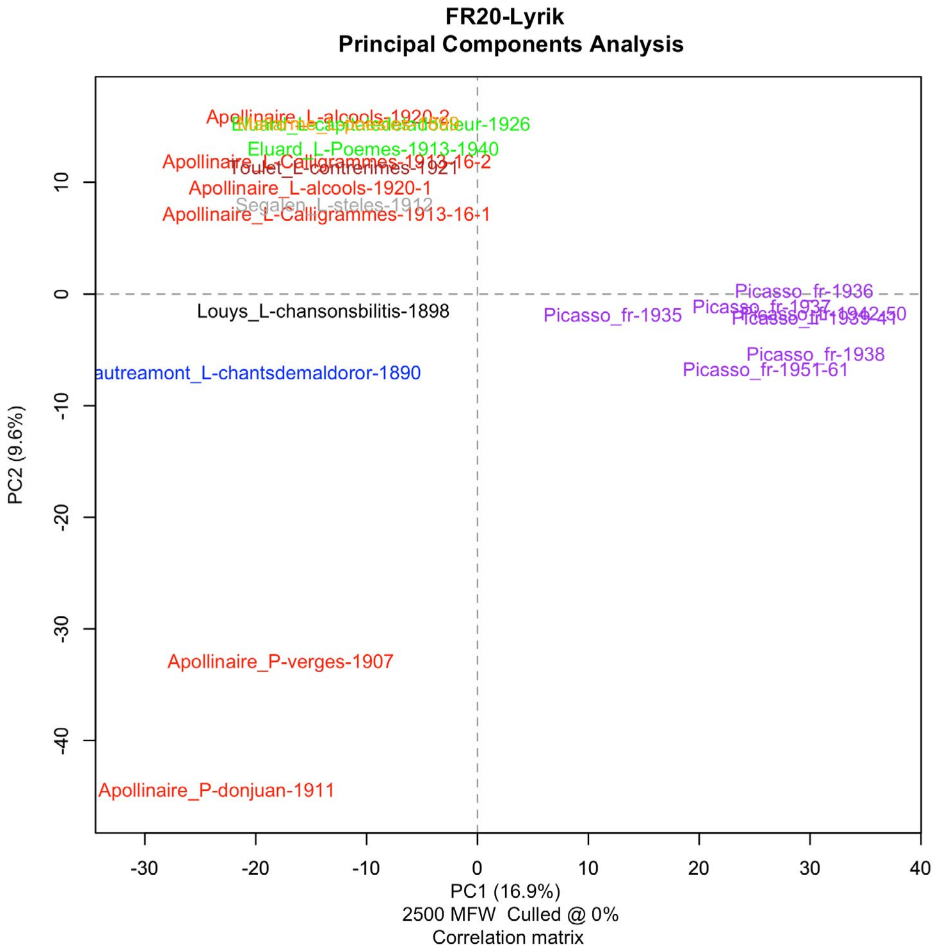


Fig. 5 Principal component analysis of the French texts (Rißler-Pipka, CC BY)

Miguel Hernández). They all use the technique of montage and the surreal description of the visible and palpable reality, which is at the same time abstract and concrete. This is probably more on the content level and that is the reason why the signal is not as strong as the stop-words which are responsible for the distribution of the MFW. The open question is accordingly: Why does Picasso's style also differ on the level of stop-words/MFW?

As a next step, I tried a principal component analysis (PCA) with the very same parameters in *stylo* and the same text collection. With this additional method I intended to get a better view of genre and author signal in comparison and will be able to visualize

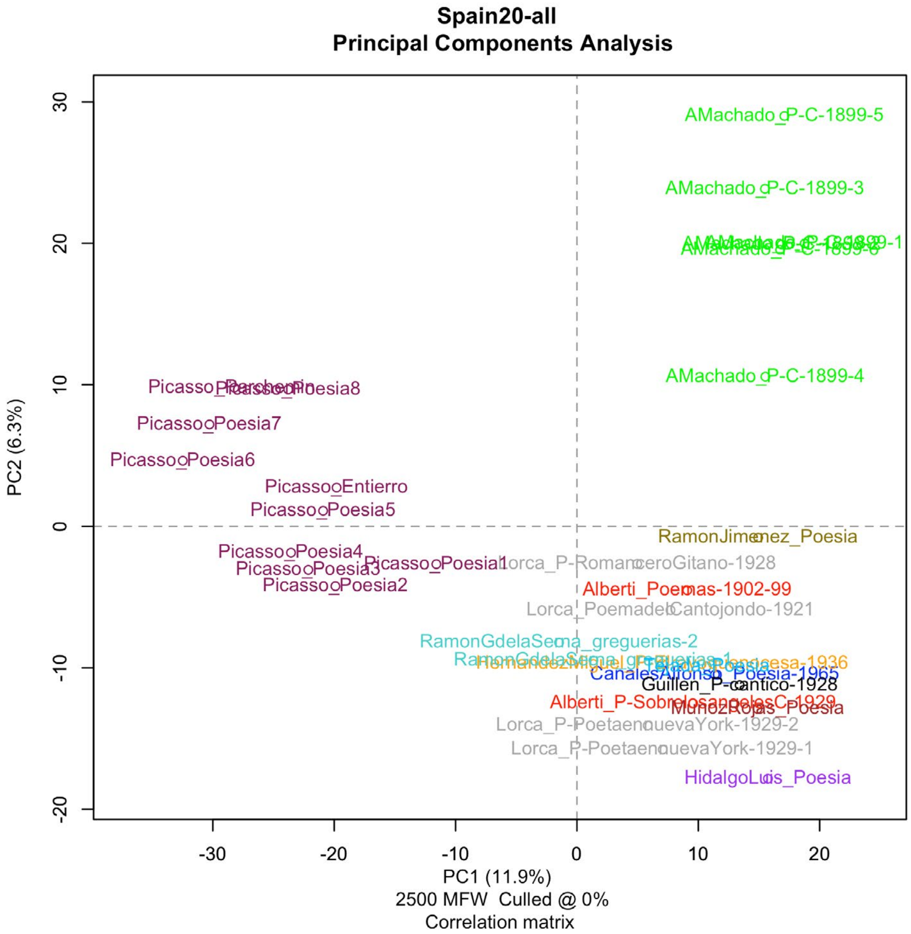


Fig. 6 Principal component analysis of the Spanish texts (Rißler-Pipka, CC BY).

the distance between the texts.[2] In Figure 5 we can also check the inner consistency of the two corpora. As we know from further research in stylometry (Eder 2010), the composition of the individual text collections may influence the results heavily.

In the upper part of the matrix (Figure 5) clusters the poetry—in the bottom the prose. It is very nice to see that Picasso as well as Louys and Lautréamont are marking

[2] The complexity of the initial distance table is reduced by the PCA, this may produce different results (for a critical explanation see Craig 2004).

the frontier between both genres. All three of them write prose in a surrealistic and poetic way. In Lourys and Lautréamont this aspect is expressed in the word *chant* ('song', 'lyrics') in the title (*Les Chants de Maldoror* and *Chansons de Bilitis*) which also hints to a subgenre of poetry. Picasso does not care to give his writings a formal literary category. However, they are clearly nearer to the prose genre than to the prose poems by Éluard or Apollinaire, which contributes to a debate highly relevant in literary studies of that time and later on (Follet 1987). The analysis clusters two texts by Apollinaire together and separates them from the rest of the texts by the same author. This apparently confirms the genre difference: the upper cluster represents poetry and the lower prose (two erotic novels by Apollinaire). The effect may also be due to other indicators like a specific vocabulary (as it is to be expected in the erotic novels by Apollinaire)—here again the composition of the corpus is relevant and could be improved in this case. Even more than in the cluster analysis, the PCA shows Picasso as an outlier on the very right side of the matrix (compare Figures 4 and 5). Again, we observe a quite similar effect in the Spanish corpus (see Figure 6).

The PCA shows three different clusters but this time not necessarily due to genre: above, on the right side clusters the more traditional classic poetry by Antonio Machado, at the bottom the avant-garde/ Surrealism poetry and on the left side Picasso, isolated, clustering half in the more experimental poetry and half in not yet categorized part of the PCA matrix. In the Spanish corpus, there is no initial mixture in genre composition (poetry versus prose) as in the French corpus. The only prose-poets in this corpus are Picasso and Ramón Gómez de la Serna. Nevertheless, in both analyses, the PCA and cluster analysis, Gómez de la Serna is clustered together with the other avant-garde poetry by Lorca and Alberti. Consequently, we need to think of another style marker beyond genre that is responsible for the Picasso outlier and that also works for Spanish and French in quite the same way.

For both languages and in different statistical settings (dendrogram, PCA), Picasso proved to be stylistically an outlier in comparison to his contemporaries. This finding supports the hypotheses (1) and particularly (2) regarding the difference of everyday language, poetic language and *Picasso-language* and consequently the uniqueness of his style.

3.3 Part of Speech Repetition

Close reading Picasso we get the impression that he loves to repeat not only of motifs, topics, objects but also grammatical entities like POS. A simple example, chosen randomly out of a longer prose poem written in summer 1940 (see Table 3) shows how two pieces of text belong together without really repeating themselves word by word and without speaking about the same thing.

Table 3 Example for repetition of grammatical entities in Picasso's poetry [red color = word repetition; green color = repetition of stop-words; blue color = strong resemblance in word melody and sound; italic = repetition of grammatical constructions]

19 July1940	01 August 1940
"buscando dentro del oro ***derretido que*** *corre* la cortina ***sus*** *fuegos* fatuos sobre el filo del estoque por el amor del dios amor *su limosnita* la línea que sube"[a]	"que vomita la boca sin dientes del calor ***derretido que*** *extiende* ***sus*** *piernas* por encima del almohadón de las faldas del veneno píldora del amor la boca abierta a la sed del estoque que la enlaza azucena que sube"[b]

[a] "looking inside the melted gold which runs the curtain its vain fires over the wire of the sword of the love of the gods of love his donation the rising line" (translation by N. R.-P.).

[b] "which vomits (or vomiting) the mouth without teeth (because) of melted heat which spreads its legs over the pillow of the skirts of the poison pill of love (with) open mouth to the thirst of the sword which ties her lily which is rising" (translation by N. R.-P.).

The semantic meaning of these pieces of poetry is clearly not relevant in terms of understanding and detecting the literary context. Picasso uses a highly poetic language and every content word like *dientes* ('teeth'), *calor* ('heat'), *estoque* ('sword') has a symbolic meaning and is at the same time very ambiguous. Here, we could easily detect words or symbols belonging to the bullfight motif and analyse them regarding their symbolic meaning and context in Picasso's other pieces of art. For example, 'teeth' and 'heat' are very important in the bullfight context in his writings but can rarely be detected explicitly in his visual art. Therefore, one could argue that by writing prose poems Picasso adds the invisible to his visual art.

However, what is he doing inside his poetry that distinguishes his writings from the French and Spanish avant-garde poets and writers? It might be the repetition of POS elements. Looking at the example in Table 3, we observe that the construction of the sentence—or part of speech—is similar with tiny changes: *dentro* 'inside' / *dientes del oro* 'teeth of melted gold' / *calor derretido que corre* 'heat which runs' / *extiende [la cortina] sus fuegos* 'spreads [the curtain] its fires' / *pierna* 'leg') etc. Using the deductive method of close reading, we could now extrapolate from here to the conclusion that the repetition of POS elements is the unique characteristic which separates Picasso's writings from his contemporaries. This also underlines the findings out of stylometry (see 3.2) because a high frequent use of prepositions, for example, is also detectable in the MFW. In Figures 3–6 the position of Picasso as an outlier may be explained in combination with the POS frequencies.

The quantitative analysis of POS elements and cross-language comparison using stylometry may prove this hypothesis not only by the observation of recurring and randomly selected paragraphs but consistently by counting the repetition of POS elements and comparing them relatively in a whole corpus. The differences in languages between Spanish and French still do not allow us to compare the POS elements or

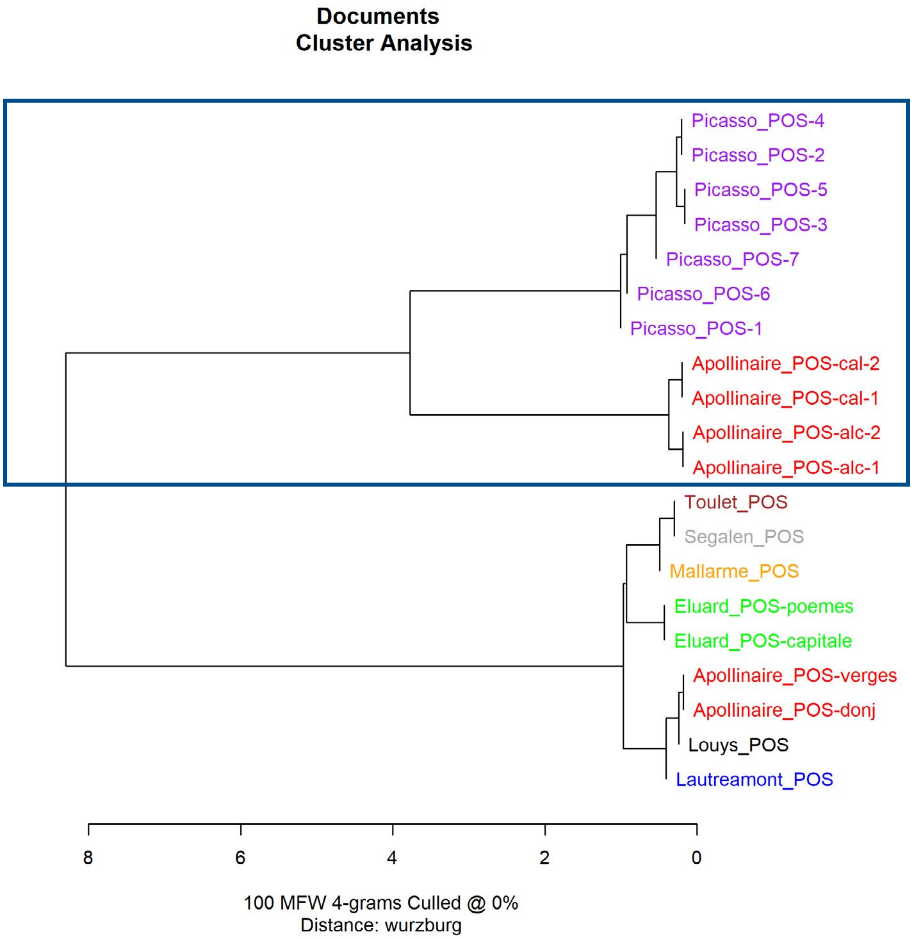


Fig. 7 Cluster analysis of POS 4-grams of the French texts (Rißler-Pipka, CC BY).

sequences directly even if we are freed of vocabulary differences. The reason is the same as for the MFW differences (see Table 1). Still, POS elements are less variable as MFW.

Jeremy Ochab tested POS sequences as style marker using the *stylo* package and comparing the most frequent POS ngrams in a corpus (Ochab 2017). Looking at the very same corpus of Spanish and French poetry including Picasso, the difference in language is visible via the number of POS tokens: in the Spanish corpus we count 157,863 word tokens and 181,905 POS tokens, while in the French corpus we count 290,131 word tokens and 316,073 POS tokens. I used the Standford POS tagger (Toutanova

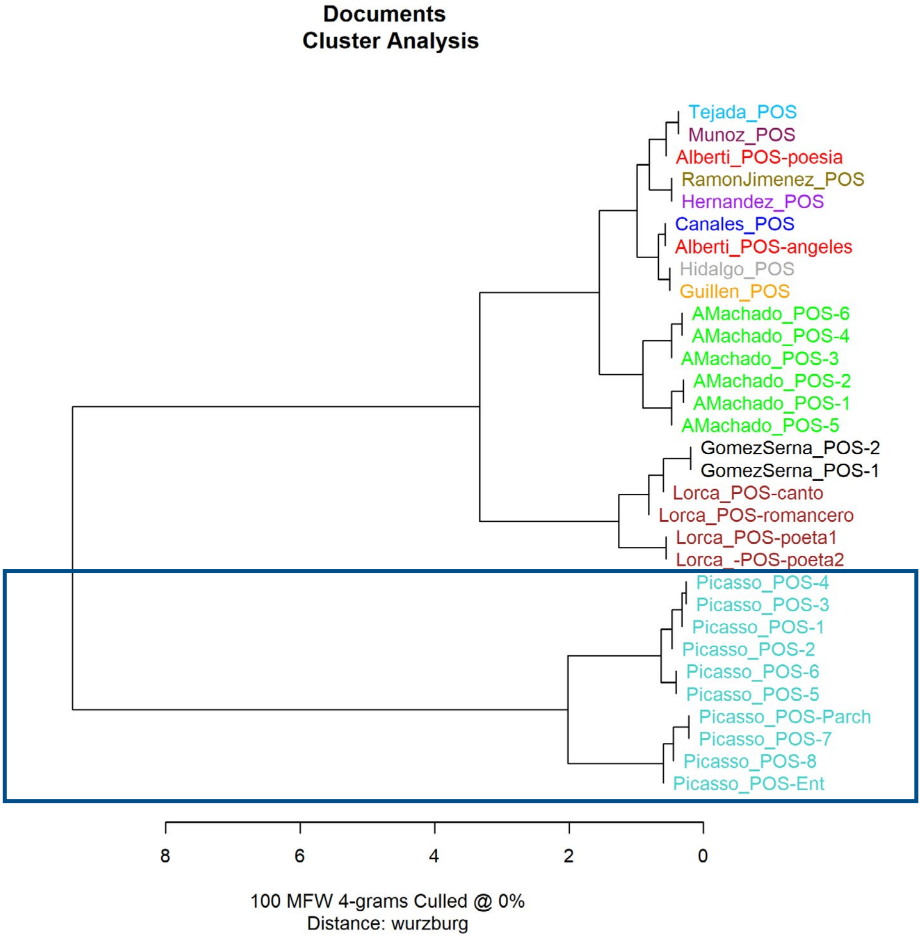


Fig. 8 Cluster analysis of POS 4-grams of the Spanish texts (Rißler-Pipka, CC BY).

et al. 2003) with different tagsets for Spanish and French. For the clustering experiment with *stylo*, I decided to test 4-grams of POS tags for each corpus with a maximum of one hundred MFW because there are not many more different 4-grams in the corpus.

At first sight, there is not much difference between the clustering with the MFW feature (Figures 3 and 4) or the POS 4-grams as a feature (Figures 7 and 8). Both cluster Picasso's writings for both languages separately. However, in the French corpus Picasso and the poetry by Apollinaire are clustered on the same branch and in the Spanish corpus Lorca and Gómez de la Serna are clustered together. All in all, the author signal is still visible in the preferences regarding frequencies for POS sequences. The clustering of Apollinaire and Picasso does not necessarily mean that both prefer similar POS sequences but may also hint to a common characteristic that distinguishes both from the rest of the corpus. As we know, Picasso and Apollinaire both nearly never use punctuation in their poetry—this might be a simple explanation.

For a deeper look, I checked the most frequent POS 4-grams for their characteristics in each language according to different tagsets (using the Stanford NLP POS tagger) (Table 4).

Table 4 The three most frequent POS 4-grams in the French and Spanish corpus: nc / ncs = noun, p / sp = preposition, det / da = determiner, punc = punctuation, cs = conjunction, ao = adjective

French	Spanish
det nc p nc	ncs sp da ncs
nc p det nc	sp da ncs sp
p det nc p	da ncs sp ncs

The problem that disturbs the comparability here is not only the different grammar and syntax but also a different POS tagset used by the Stanford NLP POS tagger: punctuation is named differently and separated into the different kinds of punctuation in the Spanish tagset, but in the French tagset it is just all assembled as *punc* ('punctuation').[3] That means that two authors not using punctuation as Picasso and Apollinaire will most likely be clustered together simply because of this very striking style marker. Tagged or not, punctuation might influence the POS tagger whether to decide between ambivalent POS elements. The most difficult question is whether a word is used as noun or verb. Without punctuation, POS taggers are more prone to errors because of missing indicators. I discussed this for Picasso in an earlier study using the FreeLing POS tagger (Rißler-Pipka 2019b). Recently Byszuk et al. discussed the accurateness of detecting direct speech in multilingual corpora, pointing out that "while heavily relying on punctuation. The last one seems particularly important for misclassifications" (Byszuk et al. 2020). That means we have to look at the results doubtfully. Also, most recently Eder and Górski highlighted the limitations of POS for different languages (Eder and Górski 2023).

[3] Mapping the tags to a universal tagset would be a solution here even if details may get lost. For me, this was not practicable at the time of the analysis. Plus, we have to keep in mind that all POS taggers coming out of the computational linguistics context are trained on contemporary everyday speech and we do not have any indication about their correctness when applying them on historic literary texts.

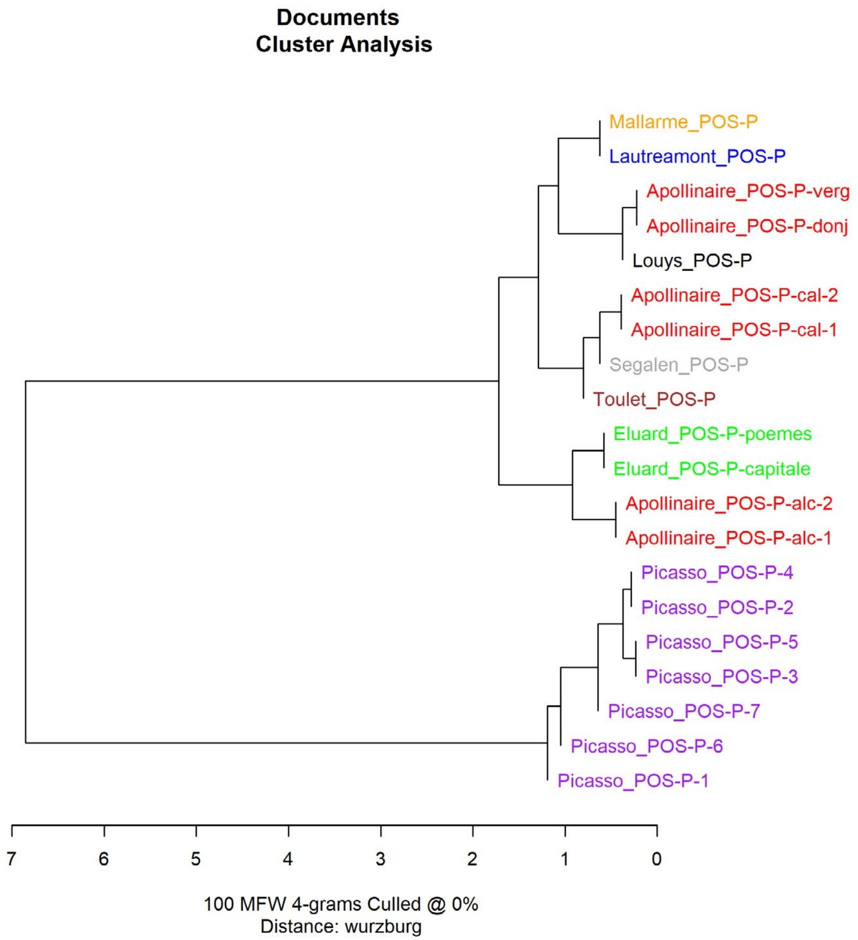


Fig. 9 Cluster analysis of POS 4-grams of the French texts without punctuation-tag (Rißler-Pipka, CC BY).

To check whether the punctuation is responsible for the clustering of Picasso and Apollinaire, I used the same parameters for the cluster analysis but deleted all *punc* ("punctuation") in the whole French corpus (Figure 9). That does not mean that the result is not influenced anymore by punctuation but for the most frequent POS 4-grams it could change the clustering.

Now again Picasso's prose poems are clustered on a single branch. Astonishingly Apollinaire is exactly divided by his anthologies *Alcools* (1913) / *Calligrammes* (1913–16) and his erotic novels *Les exploits d'un jeune Don Juan* (1911) / *Les onze mille verges* (1907).

Zooming into the results of the most frequent POS tags in the two corpora, the Spanish and the French, we might better understand the difference Picasso makes by using a very rare style of POS repetition. Comparing the tables of frequencies for both corpora represented as a heatmap, we can see which POS 4-grams Picasso uses far more often than his contemporaries and what is the combination most common for other writers, but not used by Picasso.

The visualization (Figure 10) of the frequency table is a simple csv-excel-sheet—an output of *stylo*—showing the ranked POS 4-grams on the left-hand column. For the heatmap the statistical values for each text in the corpus (here the Spanish collection) are transformed accordingly into colors (from dark blue = 0 to dark red = high frequency). As explained according to the corpus composition, Picasso's texts form a larger part and are easily detectable in the visualization.

In many texts Picasso shows a very distinctive preference for dominant POS 4-grams which are very frequently used, but only by himself. On the one hand, the most frequent POS sequences like *ncs sp da ncs* ('noun, preposition, determiner, noun') of the overall corpus are used by Picasso. Also, in the lower rows Picasso uses many POS 4-grams much more often than other authors in the corpus. On the other hand, some sequences are not at all used by Picasso (the green lines in the Picasso-block) which is not surprising because they are indicating punctuation: for example, rows 18 and 25 of the table: *sp da ncs fp* ('preposition, determiner, noun, full-stop') and *sp da ncs fc* ('preposition, determiner, noun, comma'). Even if Picasso's texts are overrepresented in the whole corpus, this cannot explain or excuse the exceptional preference for some PoS4-grams. Moreover, through the visualization we could now detect some texts by Picasso in which he focused on the repetition of the very same POS sequence and a development over time. The texts are in chronological order (from 1 to 8) apart from the first two, *El Entierro del Conde de Orgaz* (1957–59) and *Carnet Parchemin* (1940). From those two, the *Carnet Parchemin* significantly deviates and differs from others texts by Picasso. The *Carnet* consists of just one very long text dated 7 November 1940—a time when Picasso was immobile in the city of Paris occupied by Nazi Germany during World War II (1940–44). Apart from this long prose poem the rest of the *Carnet Parchemin* is lost and as we know from the dates in Table 2, Picasso nearly stopped writing in his native language Spanish from 1938 to 1957, only the year 1940 marks an exception with 61 texts (19,376 words). The manuscript of the *Carnet* (which is not freely accessible, but an example can be seen in López Sánchez 2015; Rißler-Pipka 2015, 39; Bernadac and Piot 1989, 243–51) shows the writing method of Picasso in an exceptional way: the number of insertions, arrows and deletions makes the manuscript nearly indecipherable. As we now recognize by looking at the most frequent POS sequences, the impression of repetition and preferences in grammatical order followed up to details of every line (not sentence) makes even more sense in light of this context. The nearly obsessive character of insertion and working

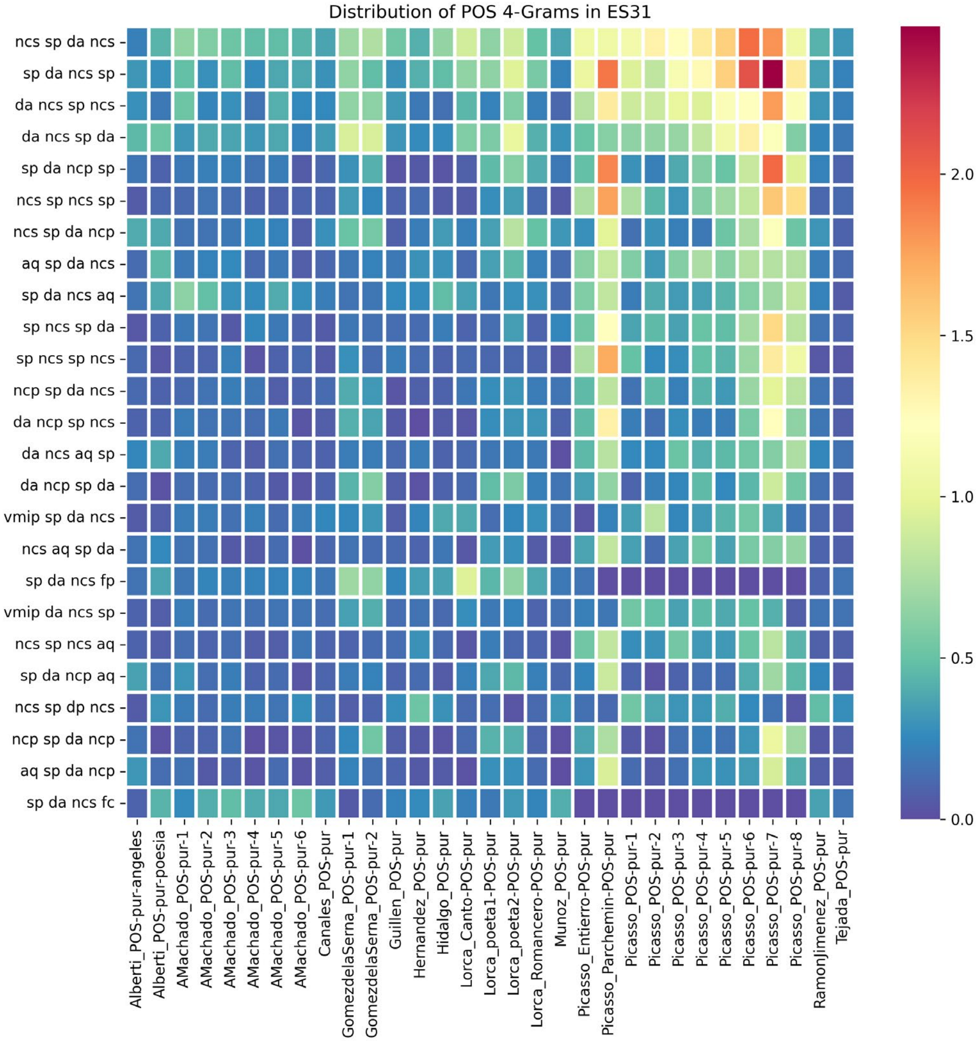


Fig. 10 Heatmap of the table of frequencies for the cluster analysis of the Spanish corpus of PoS4-grams as shown in Figure 8 (Rißler-Pipka, CC BY).

on the texts shows that he is not adding any content words but rather repeating POS sequences.

Apart from this rather detailed point of view on one particular prose poem by Picasso, the clustering of POS 4-grams of the Spanish corpus also supports the hypothesis

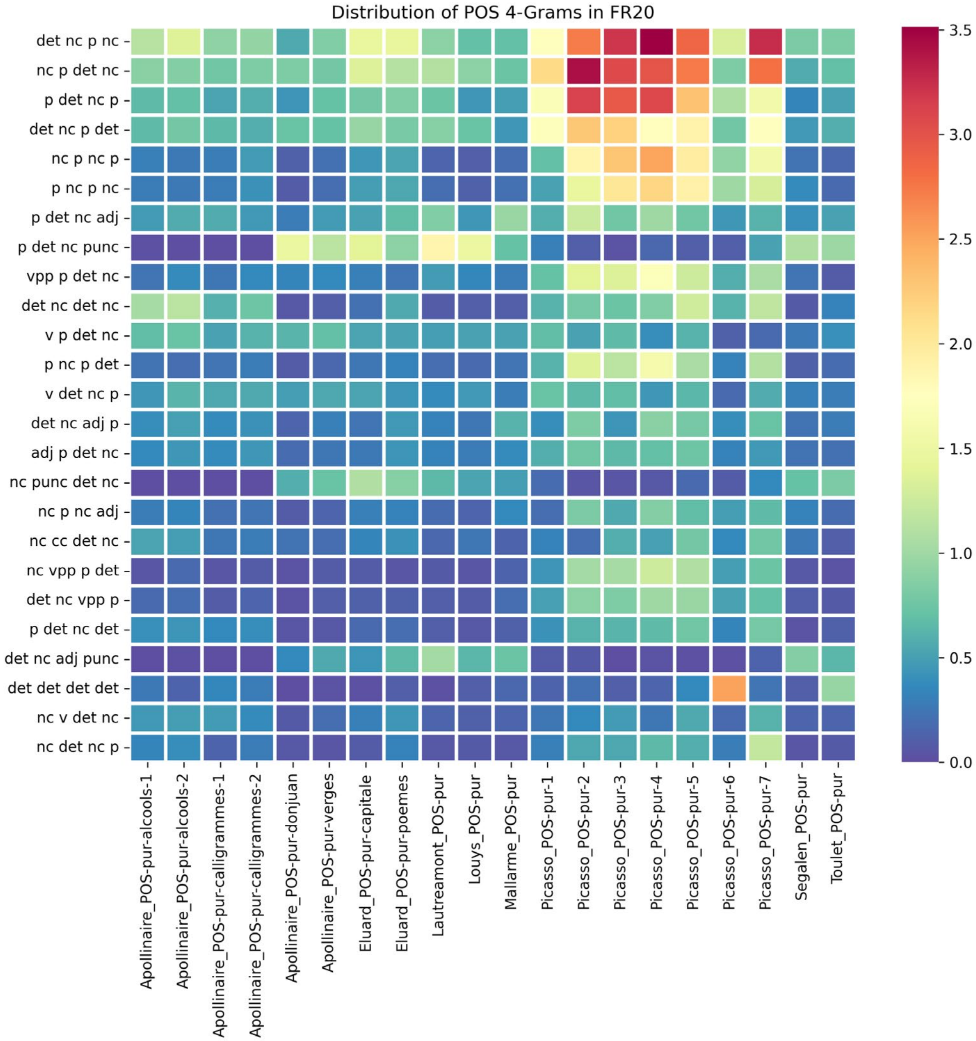


Fig. 11 Heatmap of the table of frequencies for the cluster analysis of the French corpus of POS 4-grams as shown in Figure 7 (Rißler-Pipka, CC BY).

that Picasso uses a striking system of grammatical repetition which is not comparable to other—even poetic—language (Rißler-Pipka 2015; but also Mallen 2012). In order to test if this is also true for the French corpus in which Apollinaire as a predecessor shows close resemblance in style with Picasso, I made the same visualization with the French texts (Figure 11).

As described for the Spanish corpus heatmap (Figure 10) we see the most frequent POS 4-grams ranked in the left-hand column and the heatmap shows their relative frequency for each text of the corpus. The texts by Apollinaire are easily detectable because of the dark blue brackets indicating POS 4-grams with punctuation (p det nc punc; nc punc det nc; det nc adj punc)—not used neither by Apollinaire in the four parts of his poetry anthologies (*Calligrammes* 1+2; *Alcools* 1+2) nor in most of Picasso's text. Picasso's texts present the most significant pattern. This is similar to the Spanish corpus (Figure 10). Picasso seems to have a clear preference for a set of POS sequences, which he repeats again and again. According to the heatmap for the Spanish corpus, Picasso's French texts are in chronological order beginning in 1935 up to 1961. The first year when Picasso began to write prose poems, he started writing in his native language, Spanish. The 52 French prose poems of 1935 were rather short (9,454 words all in all) while he was writing in the same year 82 prose poems in his native language Spanish (25,377 words). In the foreign language he had to find his personal style as a writer and started rather normally with very long 'sentences.' A style that can be compared to his surrealist friends, but not to the noun-preposition-noun-style he elaborated during his writer-period. As an example, we could have a look at the prose poem dated November 5, 1935:

> […] mais l'esprit de sel ne compte ses blessures sur le dos de l'amande qui se rit des regards courroucés de la foule et ne tient sa raison à la chaîne que pour mieux se moquer devant sa glace de la poche d'argent que dans sa main fait la monnaie rendue à l'ennemi […] (Picasso, November 5, 1935).
>
> "[…] but the spirit of the salt does not count its wounds on the back of the almond which laughs about the angry looks of the crowd and does not hold its reason at a chain which for better mockery in front of its mirror of the pocket of money which in its hand made the money given back to the enemy […]" (translation by N. R.-P.).

The verbs are highlighted (in green) in the citation because of the rather normal (= frequent) use of verbs which is not at all typical for Picasso's style from 1936 on or for his Spanish writings. As a counterexample we can read the prose poem dated February 6, 1938:

> […] entouré par les dents de la mâchoire du soleil plantées dans sa chair le carré de l'arène rempli d'eau soutenu par des cierges grelotte ses poils et secoue la cendre tombée sur la nappe chiffonnée déjà pleine de taches après le déjeuner collée au plafond de la grêle de flèches du clairon fixé au mât de l'épée qui […] (Picasso, February 6, 1938).

> "[…] surrounded by the teeth of the jawbone of the sun planted in its flesh the square of the arena filled with water supported by candles rings its hair and shakes the ash fallen out of the crumpled tablecloth already full of stains after the meal sticked to the ceiling of the hail of arrows of the horn attached to the pole of the sword which […]" (translation by N. R.-P.).

Here, active verbs are transformed to adverbs or adjectives describing the nouns, which are linked together, by prepositions and conjunctions (highlighted in red). Nobody or nothing does anything anymore. In this rather long paragraph, we only find two active verbs (highlighted in green). This striking preference for an endless description means a preference for certain POS sequences. The absence of verbs is rather unusual also in comparison to his contemporaries in Spanish and French avant-garde. By observing the frequencies and distances regarding every text we can also detect a development over time inside of Picasso's writings—a phenomenon we observed in a similar way in the Spanish corpus. After the first year of struggling and searching, 1935, Picasso found his typical style also in the foreign language French.

4. Conclusion

Summing up the discussion of this paper and others, it remains uncertain, if we can really compare Picasso's method of writing—or rather building texts out of a laboratory of words—to modern information capturing with the help of algorithms like the artist and art historian López Sánchez suggested already in 2015:

> […] tiene mucho que ver la visión crítica que Picasso hace de sus 'metadatos' con la labor que ejerce un analista de Big-data hoy día, pues ambos saben (quizás Picasso de manera inconsciente y por supuesto de una manera mucho más rudimentaria) que gestionando datos, analizándolos y extrayéndoles información cuando es necesario, es la mejor manera para generar el conocimiento (López Sánchez 2015).
>
> "[…] has much to do with the critical vision of Picasso regarding his 'metadata' and with the work a Big-data analyst would do today. We all know (might it be because Picasso does it unconsciously or on purpose in a much more rudimental way) that managing data, analyzing and extracting information out of them when necessary is the best way to gain and create knowledge" (translation by N. R.-P.).

Picasso was not a modern data scientist analyzing his metadata to get an illuminating knowledge graph in the end. The striking difference is the effect of destroying the grammatical logic of a language system to invent and to create something new which might lead to some knowledge about the construction of sense via syntax and semantics. Certainly, his aim was not to come to a deeper knowledge about the things and themes he was speaking about. Picasso writes about basic things (like food) and old things (like myths) and varies the perspective by new combinations in the very same system: how many combinations are possible sticking to the same vocabulary and some preferred POS 4-grams? What looks like a simple algorithmic or combinatoric question does not lead to a poetic machine (invented by Raymond Queneau at the same time, see Wolff 2016). Neither does it mean a constant variation without following any system or rule as Michaël suggests: "Il varie les écritures, passe de la prose au vers, change la disposition graphique sur la page, joue sur la linéartité et la polysémie" (2012, 169; "He varies different kinds of writing, switches from prose to verse, changes the graphical order of the page, plays with linearity and polysemy;" translation by N. R.-P.). The impression of constant variation is true, but the reason is not constant invention but a "mirror game and a game of repetition" ("jeu de miroir et de répétition") as Michaël (ibid., 179) also adds.

By stylometric analyses in a wider cross-language sense by comparing two text collections of two languages with one common author (Picasso), and by testing the hypothesis that Picasso uses a recognizable style in both languages that differs from his contemporaries, we were able to support the impression gained by close reading in various studies and to find out about the reasons behind this. This study is part of explaining the rules of the game Picasso is playing with his reader and spectator. The preferred POS 4-grams and the MFW might only represent a tiny part of the rules to be followed up. The three parts analyzed in a quantitative way, vocabulary, stylometry and POS 4-grams, all help to understand what makes the prose poems by Picasso in both languages, French and Spanish, recognizable and outstanding. It helps us to understand how he evokes the impression of a constant metamorphosis. Picasso keeps a rather fixed and small set of vocabulary (content words) in both languages but constantly changes the smaller grammatical units, while keeping some preferred and very often used stop-words and POS 4-grams.

ORCID®

Nanette Rißler-Pipka https://orcid.org/0000-0002-0719-9003

References

Béhar, Henri. 1993. "Picasso au miroir d'encre." In *L'artiste en représentation*, edited by Réné Démoris, 199–213. Paris: Desjonquères.

Bernadac, Marie-Laure, Christine Piot. "Critical Comments, Annex", in : Picasso, Pablo. Picasso: Collected Writings. Edited by Marie-Laure Bernadac. 1. publ. London: Aurum, 1989.

Byszuk, Joanna, and Maciej Eder. 2019. "Feature Selection in Authorship Attribution: Ordering the Wordlist." In *Digital Humanities 2019: Conference Abstracts*. https://doi.org/10.34894/RCOIXS.

Byszuk, Joanna, Michał Woźniak, Mike Kestemont, Albert Leśniak, Wojciech Łukasik, Artjoms Šeļa, and Maciej Eder. 2020. 'Detecting Direct Speech in Multilingual Collection of 19th-Century Novels'. In Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages, 100–104. Marseille, France: European Language Resources Association (ELRA). https://aclanthology.org/2020.lt4hala-1.15.

Calvo Tello, José. 2019. "Delta Inside Valle-Inclán: Stylometric Classification of Periods and Groups of His Novels." In *Theorien von Autorschaft und Stil in Bewegung: Stilistik und Stilometrie in der Romania*, edited by Nanette Rißler-Pipka, 151–63. München: AVM edition.

Clement, Tanya. 2013. "Text Analysis, Data Mining, and Visualizations in Literary Scholarship." In *Literary Studies in the Digital Age*, edited by Kenneth M. Price and Ray Siemens. New York: Modern Language Association of America. https://dlsanthology.mla.hcommons.org/text-analysis-data-mining-and-visualizations-in-literary-scholarship/.

Craig, Hugh. 2004. "Stylistic Analysis and Authorship Studies". In *Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell Publishing Professional. http://www.digitalhumanities.org/companion/.

Eder, Maciej. 2010. "Does Size Matter? Authorship Attribution, Short Samples, Big Problem." *Digital Humanities 2010: Conference Abstracts*, 132–35.

Eder, Maciej. 2011. "Style-Markers in Authorship Attribution. A Cross-Language Study of the Authorial Fingerprint." *Studies in Polish Linguistics* 6: 99–114.

Eder, Maciej, Mike Kestemont, and Jan Rybicki. 2013. "Stylometry with R: A Suite of Tools." *Digital Humanities 2013. Conference Abstracts*, 487–89.

Eder, Maciej, and Rafał L. Górski. 2023. 'Stylistic Fingerprints, POS-Tags and Inflected Languages: A Case Study in Polish'. Journal of Quantitative Linguistics 30, no. 1: 86–103. https://doi.org/10.1080/09296174.2022.2122751.

Éluard, Paul. 1935. "Je parle de ce qui est bien," in: "Picasso 1930–1935," special issue, *Cahiers d'Art* 10, no. 7–10: 29–32.

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. "Understanding and Explaining Delta Measures for Authorship Attribution." *Digital Scholarship in the Humanities* 32: ii4–ii16. https://doi.org/10.1093/llc/fqx023.

Fernández Molina, Antonio. 1988. *Picasso Escritor*. Madrid: Prensa y Ed. Iberoamericanas.

Follet, Lionel. 1987. "Apollinaire entre vers et prose – de « L'Obituaire » à « La Maison des morts »." *Semen – Revue de sémio-linguistique des textes et discours* 3. https://doi.org/10.4000/semen.5523.

Goddard, Linda. 2006. "Mallarmé, Picasso and the Aethetics of the Newspaper." *Word & Image* 22 (4): 293–303.

Heuser, Ryan. 2020. "Abstraction: A Literary History." Talk given at King's College, Cambridge on February 18. https://ryanheuser.org/talks/kingscollege2020/.

Heydel, Magda and Jan Rybicki. 2012. *The stylometry of collaborative translation*. Digital Humanities 2012: Conference Abstracts. Hamburg: Hamburg University Press, pp. 212–14. https://www-archiv.fdm.uni-hamburg.de/dh2012/conference/programme/abstracts/the-stylometry-of-collaborative-translation/.

Heydenreich, Titus. 1979. "'Kilómetros y Leguas de Palabras…'. Pablo Picasso als Schriftsteller." *RZLG* 3: 154–68.

Jannidis, Fotis, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2015. "Improving Burrows' Delta – An Empirical Evaluation of Text Distance Measures." In *Digital Humanities 2015: Conference Abstracts*. Sydney: University of Western Sydney. http://dh2015.org/abstracts/index.php.

Juola, Patrick, and George K Mikros. 2016. "Cross-Linguistic Stylometric Features: A Preliminary Investigation." *JADT 2016: 13ème Journées internationales d'Analyse statistique des Données Textuelles*. https://jadt2016.sciencesconf.org/80692/Draft_Paper_style_jadt_2016_2.pdf.

Leiris, Michel. 1966. *Brisées*. Paris: Mercure.

López Sánchez, Santiago. 2015. "Picasso al margen: Un viaje por los metadato." In *II Taller Experimental de Investigación sobre Picasso y el Arte del siglo XX*. Málaga: Fundación Picasso Málaga. https://fundacionpicasso.malaga.eu/export/sites/default/cultura/fpicasso/portal/menu/portada/documentos/SANTIAGO_LxPEZ_Picasso_al_Margen.pdf.

Mallen, Enrique. 2009. "The Multilingual Poetry of Pablo Picasso." *Interdisciplinary Journal for Germanic Linguistics & Semiotic Analysis* 14 (2): 163–202.

Mallen, Enrique. 2012. "La poesía simpatética de Pablo Picasso." In *Picasso – Poesie – Poetik: Picassos Schaffen aus literatur-, sprach- und medienwissenschaftlicher Sicht = Picasso, his poetry and poetics*, edited by Nanette Rißler-Pipka, 105–40. Aachen: Shaker.

Mallen, Enrique, and Luis Meneses. 2019. "Adjoined Conceptual Domains in the Bilingual Poetry of Pablo Picasso." *Digital Studies/Le Champ Numérique* 9 (1): 20. https://doi.org/10.16995/dscn.320.

Meneses, Luis, and Enrique Mallen. 2019. "Semantic Domains in Picasso's Poetry." *Digital Scholarship in the Humanities* 34 (1): i123–i28. https://doi.org/10.1093/llc/fqy078.

Michaël, Androula. 2008. *Picasso Poète*. Paris: ENSBA.

Michaël, Androula. 2012. "Picasso écrivain, la réactualisation des préoccupations." In *Picasso – Poesie – Poetik: Picassos Schaffen aus literatur-, sprach- und medienwissenschaftlicher Sicht = Picasso, his poetry and poetics*, edited by Nanette Rißler-Pipka, 165–82. Aachen: Shaker.

Rißler-Pipka, Nanette. 2019. "In Search of a New Language: Measuring Style of Góngora and Picasso." In *Theorien von Autorschaft und Stil in Bewegung: Stilistik und Stilometrie in der Romania*, edited by Nanette Rißler-Pipka, 117–50. München: AVM edition. https://www.romanischestudien.de/index.php/rst/article/view/639.

Rißler-Pipka, Nanette. 2019a. "L'esthétique numérique de Picasso." *Philologie im Netz,* Beiheft 16: 39–58. http://web.fu-berlin.de/phin/beiheft16/b16t04.pdf.

Rißler-Pipka, Nanette. [2015] 2019b. *Picassos schriftstellerisches Werk: Passagen zwischen Bild und Text*. Bielefeld: Transcript. / OA version without images published in 2019: https://nbn-resolving.org/urn:nbn:de:hbz:467-14397.

Rybicki, Jan, and Maciej Eder. 2011. "Deeper Delta across Genres and Languages: Do We Really Need the Most Frequent Words?" *Literary and Linguistic Computing* 26 (3): 315–21. https://doi.org/10.1093/llc/fqr031.

Sabartés, Jaime. 2017. *Picasso: retratos y recuerdos*. Aguadulce: Confluencias editorial.

Sinclair, Stéfan, and Geoffrey Rockwell. 2016. *Voyant Tools*. Web. http://voyant-tools.org/.

Ochab, Jeremi K. "Stylometric Networks and Fake Authorships". *Leonardo* 50, Nr. 5 (Oktober 2017): 502–502. https://doi.org/10.1162/LEON_a_01279.

O'Sullivan, James, Katarzyna Bazarnik, Maciej Eder, and Jan Rybicki. 2018. "Measuring Joycean Influences on Flann O'Brien." *Digital Studies/Le Champ numérique* 8 (1): 6. https://doi.org/10.16995/dscn.288.

O'Sullivan, James. 2014. "Finn's Hotel and the Joycean Canon." *Genetic Joyce Studies* 14: 8.

Toutanova, Kristina, et al. 2003. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network." In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology,* vol 1. 173–80. Edmonton, Canada: Association for Computational Linguistics. https://doi.org/10.3115/1073445.1073478.

Wolff, Mark, 2016. 'Oulipian Code', http://markwolff.name/wp/digital-humanities-2/oulipian-code/.

For the software: https://nlp.stanford.edu/software/tagger.html.

For the corpora of digitized texts: Biblioteca Virtual Miguel de Cervantes, Wikisource and Mallen, Enrique. n.d. Picasso: Online Picasso Project. Enrique Mallen. https://picasso.shsu.edu/.

Digital Approaches to Poetic Style
A Quantitative Stylistic Analysis of Italian Petrarchism

Jan Rohden

Abstract Petrarch can be considered one of the most influential poets of European literature. One of the main reasons for this is his collection of Italian love poems known as *Canzoniere*, which for centuries became a role model for many European poets trying to imitate Petrarch's poetic style. Research has acknowledged Petrarch's influence on later poetry and even created a term to describe this phenomenon: Petrarchism. Yet, despite the many studies describing Petrarch's impact on various European authors and texts, the notion of Petrarchism itself continues to be under discussion. This article raises the question to what extent digital methods can provide new impulses for research on Petrarchism. More specifically, a quantitative stylometric analysis of a corpus of Italian love poetry is conducted to find stylistically distinctive elements for Petrarchism.

Keywords Petrarch, Petrarchism, *stylo*, quantitative analysis, contrastive analysis

1. Petrarch: Italian Author, European Role Model

1.1 Petrarch's *Canzoniere*

The Italian Francesco Petrarca (1304–74), commonly anglicized as Petrarch, can be considered one of the most influential authors of European literature. His works left their mark on the literary landscape not only of his time, but also on the texts of many later writers. Based on the language, Petrarch's writings can be divided into two categories. On the one hand, he published different texts in Latin, including an epic poem, a collection of biographies of famous historic persons and a collection of letters. On the

Rohden, Jan: Digital Approaches to Poetic Style. A Quantitative Stylistic Analysis of Italian Petrarchism, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 81–100. DOI: https://doi.org/10.17885/heiup.1157.c19367

other hand, one of the main reasons for Petrarch's fame is however his well-received collection of love poems written in Italian and known as *Canzoniere*.[1]

The central theme of *Canzoniere* is the lyrical I's unrequited love for a married woman called Laura, a love that continues even after the woman's death. The 366 poems in the collection draw on central motifs of the Latin, Provençal and Italian literature and are arranged according to an elaborate structure. This structure not only provides a temporal order that is supported by references to specific dates and periods of the ecclesiastical year (Barolini 1989; Fornasiero 2001, 59–89), but also establishes a narrative dimension, creating an autobiographic tale that begins with the lyrical I's love at first sight for Laura and ends with the supposed renunciation of its love for her (Geyer 2009; Wehle 2009).[2] Thus, *Canzoniere* gives a detailed account of the lyrical I's feelings for Laura, which are often contradictory: the lyrical I is frequently torn between a sentiment of pleasure provided by its love and the pain resulting from its unfulfilled desire for Laura. This conflicting emotional state of the lyrical I, often referred to as *dolendi voluptas* by scholars, is a fine example of the oppositeness that permeates the form and content of Petrarch's *Canzoniere* (Friedrich 1964, 217–19; Forster 1969).

1.2 Approaches to a Definition of Petrarchism

Canzoniere had an enormous impact on European love poetry. In fact, the text became a role model for a great number of authors trying to imitate Petrarch's poetic style in Italy and beyond. Petrarch's influence appears most clearly in Italian collections of poems also entitled *Canzoniere*, alluding to Petrarch's work. Apart from such obvious references, literary scholars have pointed out elements they consider characteristic for Petrarch's poetic style in the texts of many other European authors. Research even invented the term *Petrarchism* to describe Petrarch's impact on his literary successors.

There is a vivid scholarly research discourse on Petrarchism (Hempfer, Regn, and Scheffel 2005). While most studies focus on the impact of Petrarch's poetic style on single authors or texts (e.g. Pyritz 1963; Regn 1987; Warning 1987; Morales Saravia 1998; Schiffer 2000; Schneider 2007; Marnoto 2015), some contributions aim to describe Petrarch's influence on the later literary landscape systematically (e.g. Baldacci 1957; Forster 1969, 61–83; Hoffmeister 1973; Nardone 1998; Bernsen 2011; Regn 2013). Among the latter, three approaches stand out: The first one conceives Pertrarchism as a literary system of elements (Hempfer 1987; 1991; Regn 1987; 1993). In other

[1] Over the many years that Petrarch worked on the collection, its title and structure changed several times (Wilkins 1948; Santagata 1993).

[2] For an overview of the sources and intertextual references in Petrarch's Canzoniere see Petrarca (2015).

words, a text has to include at least a minimum number of certain elements in order to be considered Petrarchistic. Referring to Mikhail Bakhtin's notion of *Dialogism*, the second concept considers Petrarchism a form of literary appropriation of competing lyrical dictions (Warning 1987). Warning argues that a text can be called Petrarchistic as long as the Petrarchistic discourse is the dominant one. If it is supplanted by other types of discourse instead, the text in question is not Petrarchistic. The third concept tries to combine the two opposing conceptions (e.g. Huß 2001).

The different approaches proved valuable for shedding light not only on Petrarchism in general, but also on the various ways single authors adapt Petrarch's poetic style. Research even managed to identify some recurring elements in European literature that seem characteristic for Petrarch's way of composing poetry, including:

- the above-mentioned concept of *dolendi voluptas* (see 1.1);
- stylistic devices that express contrast, e.g. the oxymoron (Friedrich 1964, 217; Regn 2013);
- the idealization of the beloved woman both ethically and aesthetically (Regn 2013).

Nonetheless, there is still no conclusive list of elements that would allow us to distinguish Petrarchistic texts from non Petrarchistic literature. In order to create a list of stylistically distinctive elements of Petrarchism, it would be useful to analyze a large corpus of texts considered to be Petrarchistic and written by different authors, instead of focusing on single authors or texts, like most of the research literature has done so far. In fact, even publications with the goal to study Petrarchism as a European phenomenon usually analyze small numbers or even single texts or authors.[3] Quantitative approaches however, which in recent years enjoyed increasing popularity thanks to the ascent of the digital humanities, give the chance to analyze Petrarchism on a large scale.

2. Digital Text Analysis and Poetry

Digital literary studies have experienced a veritable boom in the last years, as some widely received studies provided illustrative examples for the possibilities digital tools offer, especially for quantitative approaches to large text collections (e.g. Jockers 2013; Moretti 2013).[4] This led to a variety of different studies on major literary genres, in particular prose and drama. Although the number of digital analyses that deal with

[3] There are, however, some exceptions, e.g. Baldacci (1957); Hoffmeister (1973); Nardone (1998).

[4] For an overview of quantitative analyses in the digital humanities see Schöch (2017).

poetry is lower compared to the latter two genres, there are some publications that can be divided into three groups based on their approach: studies that analyze the metric structure of poetic texts (e.g. Beaudouin and Yvon 1996; Navarro Colorado 2018a), publications which focus on poetic language (e.g. Rhody 2012; Navarro Colorado 2018b) and stylometric approaches (e.g. Hoover 2008; Rojas Castro 2018).

Nonetheless, there have been no digital approaches to define or analyze Petrarchism on an international or even national level so far. However, a digital quantitative analysis could help identifying stylistically distinctive features of Petrarchistic poetry on a scale that goes beyond the level of a single text or author.

3. Digital Approaches to a Quantitative Stylistic Analysis of Petrarchism

3.1 Approach

If Petrarchism, as the name suggests, refers to a specific type of influence of Petrarch's *Canzoniere* on the texts of other poets, then in Petrarch's poetry and in the works of his poetic successors there must be aspects that distinguish these texts from non Petrarchistic literature. In order to find these distinctive elements, it is necessary to compare definitely Petrarchistic texts with poetry that is certainly not Petrarchistic on a broad basis. Contrastive analyses are suitable for this purpose, as demonstrated by various contrastive studies in recent years (e.g. Schöch 2018; Ilsemann 2019; Rebora et al. 2019) in which the distance measure Zeta has proven useful. Moreover, Zeta is quite user-friendly thanks to its implementation in the R package *stylo* (Eder, Rybicki, and Kestemont, 2016) and the Python library *pyzeta* (Schöch 2019) and will therefore also be used here.

In the following, such a contrastive analysis is conducted by using *stylo* to study a corpus of twelve collections of Italian poetry.[5] The analysis consists of four steps: the first step examines the whole corpus with regard to the stylistic similarity between the different collections of poems contained in it. In the second step the collections of Italian love poetry written before and Petrarch's *Canzoniere* are analyzed contrastively. The third step consists of a contrastive analysis between the Italian love poetry written before and poetic collections published after the *Canzoniere*. The fourth and last step compares the results of steps two and three.

[5] See 3.3 for the structure of the corpus in detail.

3.2 Genre and Language

When dealing with premodern Italian poetry, it is recommended to keep two things in mind. The first aspect concerns the structure of poetic texts in general. One of the most characteristic features of poetry is that it is often composed of verses. At first glance verses are usually associated with deliberate line breaks. Although this is often the case, the fact that a poem consists of verses may have more consequences for the text. Verses can also provide a metric structure and thereby a rhythm. This is an important point because it means that in order to analyze poetic texts it is frequently necessary to take into account not only the graphic dimension of the text, but also acoustic aspects.[6] Moreover, the fact that poetry is often composed of verses results in poetic texts being shorter on average than prose texts. The second aspect that needs to be considered when working with premodern Italian poetry is that the texts in question are centuries old. On the one hand, this means that there are authors of whom only few texts have survived. On the other, it implies that the texts were written at a time when the standardization of the Italian written language was in its early stages.[7] Therefore, it is not unusual to find various variants for the same linguistic phenomenon in different texts of the same time.

These two aspects have two consequences for the analysis. The first concerns the evaluation of possible results. The results of a quantitative stylometric analysis primarily refer to the graphic level of the text, whereas aspects relating to the acoustic dimension are not revealed. Therefore, whatever stylometric approaches may teach us about Petrarchistic poetry, it may only be one part of the solution and a first step toward a definition of Petrarchism. The second consequence regards the choice of the texts for the corpus. For a quantitatively sound analysis, every text in the corpus must have a certain length. This is the reason why only poetic collections with a length of at least 1,000 verses were taken into account. Especially in the twelfth and thirteenth century there are however quite a few authors, of whom only a few Italian poems have been preserved. Leaving out such texts would have meant ignoring a substantial part of early Italian poetry. In order to solve this problem, anthologies of early Italian poetry were included in the corpus.

3.3 Corpus

Table 1 provides an overview of the corpus analyzed in this essay, which includes three parts. The first part consists of four collections of Italian love poetry written before Petrarch's *Canzoniere*: poems of the *Scuola Siciliana*, Tuscan poetry, *Dolce Stil Novo* and

[6] Apart from the verse, e.g. rhymes and stylistic devices may influence the acoustic dimension of a poetic text.

[7] For an overview of the history of the Italian language see Blasi (2008, especially 3–70).

Table 1 Corpus

Part	Author(s)	Type of text	Title/school	Centuries	Number of characters
A	various	Anthology	*Scuola siciliana*	12th–13th	350,359
	various	Anthology	Tuscan Poetry	13th	130,950
	various	Anthology	*Dolce Stil Novo*	13th–14th	162,349
	Dante	Love poems	*Rime*	13th	59,842
B	Petrarch	Collection	*Canzoniere*	14th	287,288
C	Bandello, Matteo	Collection	*Canzoniere*	16th	182,676
	Conti, Giusto de'	Collection	*Canzoniere*	15th	164,530
	de Medici, Lorenzo de'	Collection	*Canzoniere*	15th	119,608
	Galli, Angelo	Collection	*Canzoniere*	15th	353,230
	Rossi, Niccolò de'	Collection	*Canzoniere*	14th	221,159
	Sforza, Alessandro	Collection	*Canzoniere*	15th	227,557
	Tansillo, Luigi	Collection	*Canzoniere*	16th	327,730

Dante's love poetry.[8] Petrarch's *Canzoniere* is the second part. The third part consists of collections of Italian poems published after Petrarch's *Canzoniere*, but bearing the same title. These collections were deliberately chosen, because their title suggests that the respective author was familiar with Petrarch's work. All the texts are based on curated editions, whose electronic versions were obtained via the digital library *Biblioteca Italiana* (Quondam, Alfonzetti, and Asperti 2019). For the analyses, each collection was stripped of page and line numbers, titles, notes and editorial information and saved in a single plain text file (UTF-8).

3.4 Analysis of the Stylistic Similarity of the Texts

Figure 1 shows the stylometric similarity of all the texts in the corpus based on the 3,000 most frequent words according to the Classic Delta distance.[9]

[8] In the case of Scuola siciliana, Dolce Stil Novo, the Tuscan poetry and Dante's poems, the texts were extracted from larger collections, also available via Biblioteca Italiana (Scuola siciliana: Panvini 1962–64; Dolce Stil Novo: Contini 1960; the Tuscan poetry: Zaccagnini and Parducci 1915; Dante's love poetry: Contini 1973).

[9] Classic Delta is based on the distance measure Burrows' Delta introduced by John Burrows (2002). In order to calculate Classic Delta, the word frequencies in a corpus are first converted into relative word frequencies and then subjected to a Z-transformation. Based on the resulting values, the similarity between the texts in the corpus is then determined using the Manhattan distance.

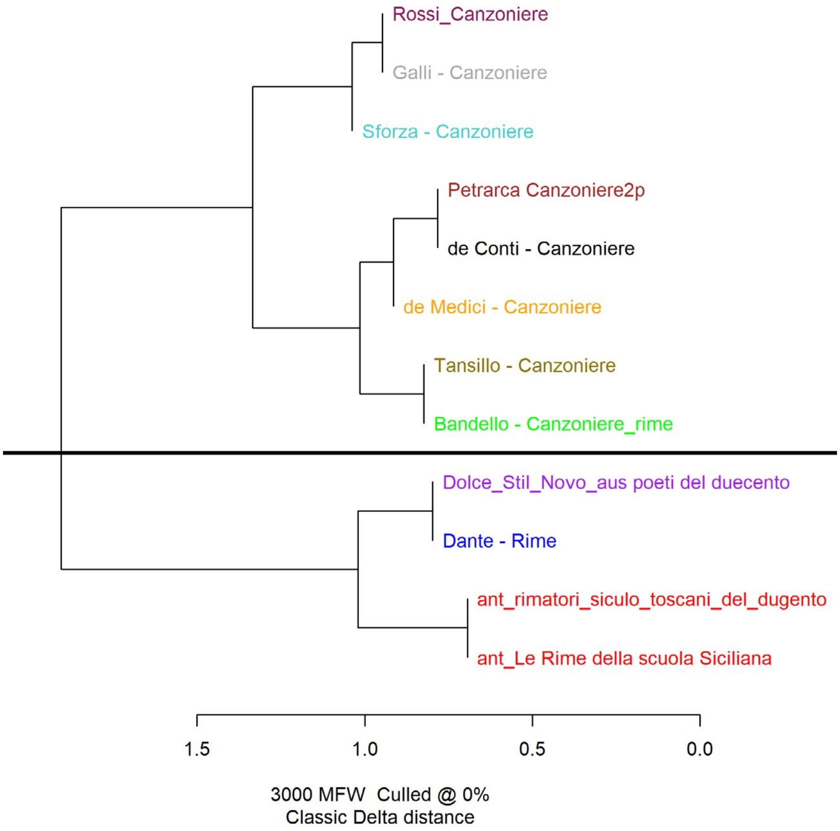


Fig. 1 Stylometric similarity, cluster analysis, 3,000 most frequent words, Classic Delta (corpus parts A, B, C), (Rohden, CC BY).

The dendrogram illustrates a clear distinction between all the collections of poems entitled *Canzoniere*, including Petrarch's text on the one hand (above the line), and the love poetry composed before on the other (below the line).[10] The dendrogram in Figure 2, based on the same number of most frequent words, but according to the Würzburg distance,[11] confirms the contrast between the *Canzonieri* since Petrarch and the love poetry composed before.

[10] The black line in the lower part of the dendrogram was added manually.

[11] In contrast to Burrows' Delta, the Würzburg distance uses the Cosine distance instead of the Manhattan distance to calculate the similarity (Jannidis et al. 2015).

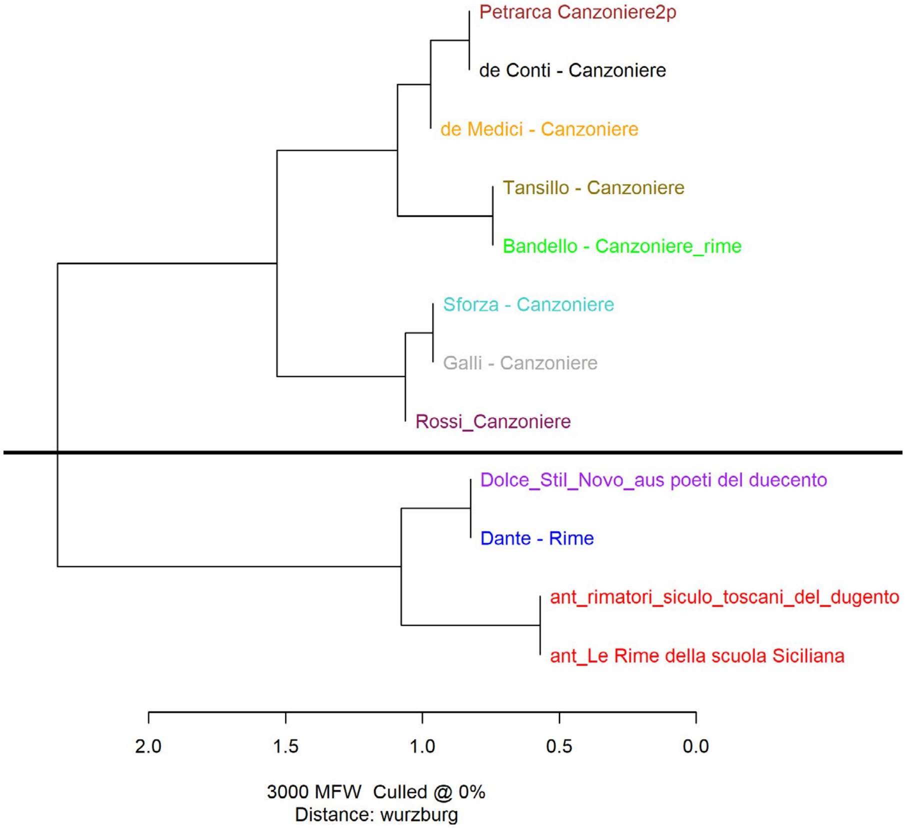


Fig. 2 Stylometric similarity, cluster analysis, 3,000 most frequent words, Würzburg (corpus parts A, B, C), (Rohden, CC BY).

From a stylometric point of view both dendrograms therefore suggest a clear difference between the *Canzonieri* since Petrarch and the love poetry written before. However, the question arises as to what causes these stylometric differences. Can they possibly be traced back to writing variants or instead to other stylistic elements? Contrastive text analyses can shed light on this.

3.5 Contrastive Analysis of Petrarch's *Canzoniere* and the Italian Love Poetry Written Before

A contrastive analysis reveals stylometric differences between Petrarch's *Canzoniere* and the Italian love poetry written before, as well, as Figure 3 shows.

First of all, what is conspicuous about Figure 3 is the fact that there are various variants of one and the same lemma. One example is the word *dolze* ('sweet'), written

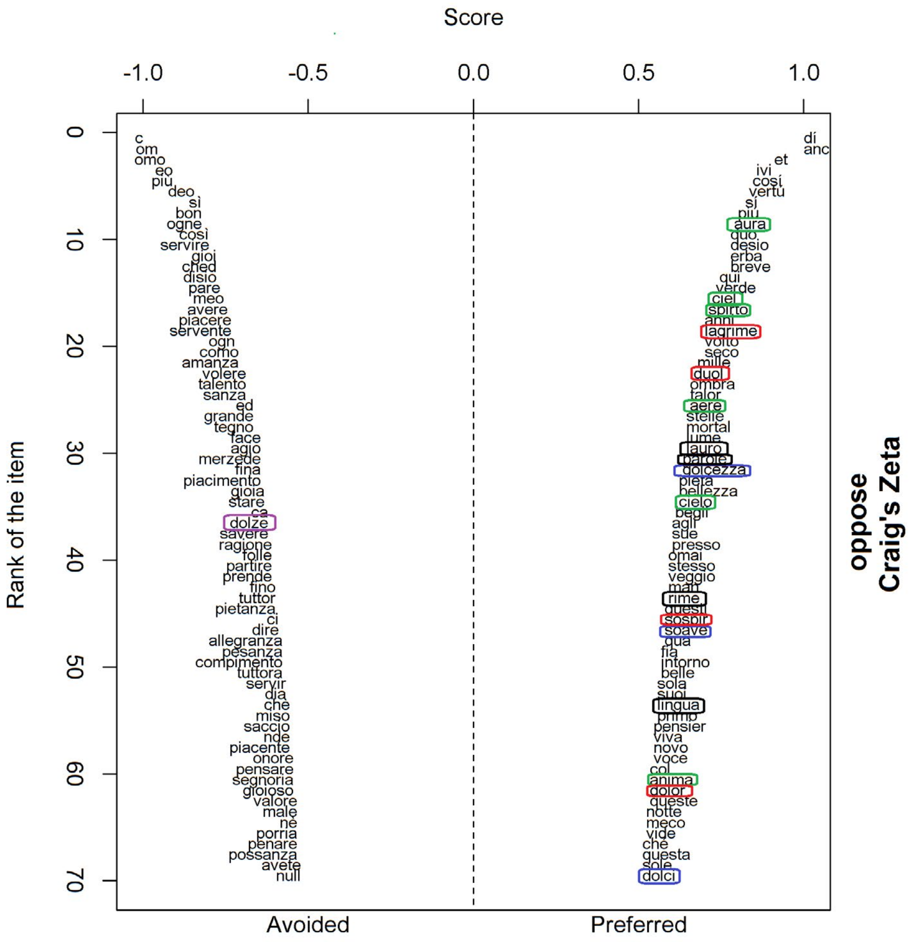


Fig. 3 Contrastive analysis of Petrarch's *Canzoniere* and the Italian love poetry written before, Oppose, slice length 3,000, Occurrence Threshold 30, Craig's Zeta (corpus parts A vs B), (Rohden, CC BY).

with a <z>, which is avoided in Petrarch's *Canzoniere*, whereas the plural *dolci* with a <c> as well as the noun *dolcezza* ('sweetness') are preferred by him.[12] This is a reminder of the fact that early Italian poetry may include different variants of the same lemma. Apart from *dolce*, there is also another expression designating sweetness: *soave*. The presence of various words for sweetness in Petrarch's *Canzoniere* can be regarded as a reference to the literary movement *Dolce Stil Novo*.[13] Another group of words preferred by Petrarch deals with different forms of pain, in particular *duol* and *dolor* (both referring to 'pain'), *lagrime* ('tears') and *sospir* ('sigh'). These words emphasize the importance of the motif of pain, not least for the concept of *dolendi voluptas*. Another interesting aspect of the words listed in Figure 3 are nouns that imply a spiritual dimension of Petrarch's poetry: *aura* ('aura, air'), *ciel* and *cielo* ('sky, heaven'), *spirto* ('spirit, soul'), *aere* ('sky, heaven') and *anima* ('soul, spirit'). These words suggest the idealization of the beloved woman, a common motif in the poetry of *Dolce Stil Novo* and in Dante's love poems (e.g. Seitschek 2014). Four nouns are noteworthy, as well: *lauro* ('laurel'), *parole* ('words'), *rime* ('rhymes') and *lingua* ('language, tongue'). These words can be seen as a reference to poetry itself, since words and rhymes are among its most basic components. The laurel is particularly interesting, as it has two meanings in Petrarch's *Canzoniere*. On the one hand it symbolizes the crown of the poet, the highest poetic award as well as a sign of poetic fame.[14] On the other *lauro* can also allude to the beloved Laura, due to the similar spelling of both words. The motif of the laurel therefore elucidates the duality of Petrarch's poetry, which deals with love, yet at the same time serves Petrarch's goal to become a famous poet (e.g. Wehle 2009).

3.6 Contrastive Analysis of the *Canzonieri* after Petrarch and the Italian Love Poetry Written Before

Although some expressions may differ, a contrastive analysis of the *Canzonieri* after Petrarch and the Italian love poetry composed before (Figure 4) confirms many of the observations described in 3.5.

A reference to poetry is present in form of the three words *stile*, *stil* ('style') and *carte* ('papers'). The nouns *gloria* ('glory') and *fama* ('fame') clearly refer to the motif of poetic fame, a motif symbolized in Petrarch's *Canzoniere* by the laurel. Furthermore, in

[12] While dolze is marked purple, all the other words referring to the semantic field of sweetness are marked blue. Color-marked in this and all following figures are words referring to: sweetness (blue), pain (red), spirit/soul/heaven (green), poetry (black), happiness/pleasure (orange).

[13] For an overview of Dolce Stil Novo see Pirovano (2014).

[14] In fact, in 1341 the Roman Senator Ursus d'Anguillara officially granted Petrarch this award by giving him a crown made of laurel, thus crowning him poeta laureatus (Suerbaum 1972).

the *Canzonieri* after Petrarch there is even stronger evidence for the idealization of the beloved woman, as the words *eterno* and *eterna* ('eternal'), *alma* ('spirit, soul'), *celeste* ('celestial, heavenly'), *sacro* ('holy') and *ciel* ('sky, heaven') demonstrate. The concept of sweetness can be found in Figure 4, as well, although in this case it is only represented by a single adjective (*suave*). Moreover, the element of pain is included in the wordlist, although also only with a single word (*duol*). Instead, there are three adjectives that imply happiness or pleasure: *felice* ('happy') and *lieto* as well as *lieta* ('happy'). These three words not only show that the aspect of duality which can be found in Petrarch's

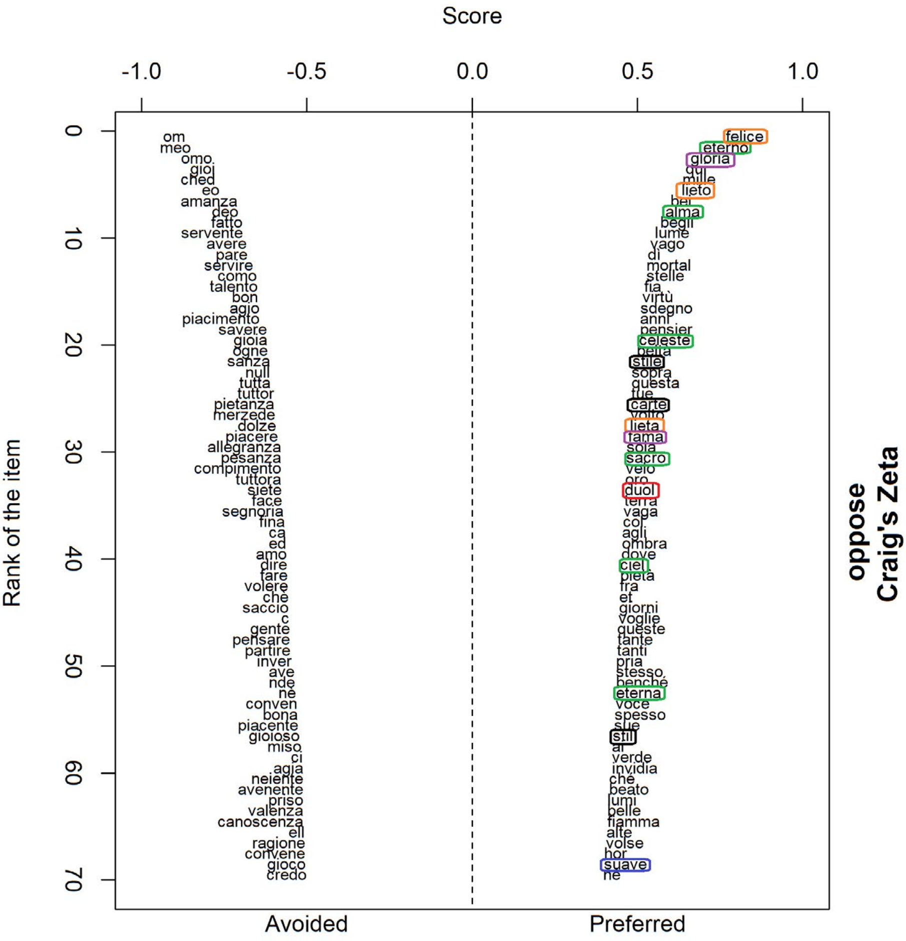


Fig. 4 Contrastive analysis of the *Canzonieri* after Petrarch and the Italian love poetry written before, Oppose, slice length 3,000, Occurrence Threshold 30, Craig's Zeta (corpus parts A vs C), (Rohden, CC BY).

collection is part of later *Canzonieri*, as well, but they also refer to the second component of the concept of *dolendi voluptas* besides pain: pleasure.

3.7 Comparison of the Results

By pointing out notable differences between the Italian love poetry written before Petrarch in contrast to the *Canzonieri* since Petrarch, Figures 3 and 4 confirm relevant observations of literary research on Petrarchism from a quantitative perspective (e.g. Forster 1969; Regn 2013).

A closer look at the wordlists generated by the two contrastive analyses however reveals two more differences. The first one does not concern the content of the words, but rather the numerical ratio between two parts of speech, nouns and verbs. According to the results of the contrastive analysis based on Craig's Zeta,[15] while there is only a difference of approximately 22 percent between the number of nouns avoided (73) and those preferred (89) in the love poetry written before Petrarch, in Petrarch's *Canzoniere*, the number of verbs avoided (88) is more than three times higher than the number of verbs preferred (23). A contrastive analysis of the same texts based on Eder's Zeta leads to similar results,[16] although in this case the number of nouns preferred (83) is nearly twice as high as the number of nouns avoided (48), while the number of verbs avoided (62) is more than three times higher than the number of verbs preferred (20). Thus, when compared to the love poetry written before, Petrarch's *Canzoniere* seems to prefer nouns at the expense of verbs. A similar preference for nouns can be observed for the *Canzonieri* written after Petrarch in contrast to the love poetry composed before the *Canzoniere*.

In the case of Craig's Zeta, the number of nouns preferred (205) is nearly two times higher than the number of those avoided (104), while the number of verbs avoided (114) is about 34 percent higher than the number of verbs preferred (85). An analysis based on Eder's Zeta obtains similar results, with the number of nouns preferred (205) being more than twice as high as the number of nouns avoided (96) and the number of verbs avoided (138) being approximately 47 percent higher than those preferred (94). Table 2 summarizes the preference for nouns at the expense of verbs in

[15] Craig's Zeta (Craig and Kinney 2009, 18–22) is a variant of the distance measure Burrows' Zeta (Burrows 2007) originally developed by John Burrows. For the calculation, the corpus to be analyzed is first divided into a target and a comparison partition. Then the document proportions are calculated for each feature in both partitions. To determine the Zeta values, the document proportions of the comparison partition are subtracted from those of the target partition. For the mathematical background of Zeta, see Schöch (2018).

[16] Unlike Craig's Zeta, the Zeta values for Eder's Zeta are not calculated by subtraction, but based on the Canberra distance.

the *Canzonieri* since Petrarch compared to the love poetry written before, as illustrated by the contrastive analyses.

Table 2 Number of nouns and verbs preferred and avoided in the *Canzonieri* since Petrarch compared to the love poetry written before according to the contrastive analyses (A vs B vs C)

	A vs B		A vs C	
	Craigs's Zeta	Eder's Zeta	Craigs's Zeta	Eder's Zeta
Verbs preferred	23	20	85	94
Verbs avoided	88	62	114	138
Nouns preferred	89	83	205	205
Nouns avoided	73	48	104	96

The tendency in the *Canzonieri* since Petrarch to prefer nouns at the expense of verbs, which can be seen in the example of the contrastive analyses, is confirmed by the absolute word frequencies. This shows a comparison of the 1,000 most frequent words in the love poetry before Petrarch, his *Canzoniere*, and the *Canzonieri* written after Petrarch. Whereas the 1,000 most frequent words of the love poetry written before Petrarch include 256 verbs and 195 nouns, Petrarch's *Canzoniere* comprises 198 verbs and 261 nouns, and the *Canzonieri* since Petrarch include 176 verbs and 235 nouns. Table 3 summarizes the preference for nouns at the expense of verbs in the *Canzonieri* since Petrarch compared to the love poetry written before, as illustrated by the 1,000 most frequent words.

Table 3 Number of nouns and verbs in the *Canzonieri* since Petrarch compared to the love poetry written before according to the 1,000 most frequent words (A, B, C)

	Number of verbs	Number of nouns
A	256	195
B	198	261
C	176	235

The results of the contrastive analyses and the 1,000 most frequent words of each of the three parts of the corpus thus suggest that the *Canzonieri* since Petrarch prefer nouns at the expense of verbs. An examination of the preferred words revealed by the contrastive analyses shows that certain kinds of nouns are preferred, namely words which belong to one of the categories listed in Table 4.

While the first three groups confirm what research literature has already pointed out, groups four and five are astonishing. The nouns in group four allude to a corporal dimension in the *Canzonieri* since Petrarch. This confirms Paolo Rigo's thesis

Table 4 Groups of nouns preferred in the *Canzonieri* since Petrarch in contrast to the Italian love poetry written before (A vs B & C)

Number	Category	Examples
1	poetry/language	*rime, stile, versi*
2	fame	*lauro, fama, gloria*
3	soul/spirit	*anima, alma, aura*
4	body	*volto, piede, seno*
5	landscape	*erba, valli, colli*

that the motif of the body can be found various works by Petrarch (Rigo 2017). The predominance of this topic in the *Canzoniere* is however remarkable, not only because Laura dies in the course of the collection. It is also noteworthy since the simultaneous preference for nouns belonging to the semantic field of the soul/spirit on the one hand and the category of the body on the other adds to the dualistic structure that has been described as characteristic for Petrarch's conception of love (e.g. Regn 2013). This contrast of body and soul/spirit appears to be an important aspect in the later *Canzonieri*, as well. In this respect the present study confirms Stephan Leopold's thesis, whose psychoanalytic reading of Petrarchism indicates the relevance of the body as a motif in the works of seven European authors apart from Petrarch (Leopold 2009).

The fifth group of words, which emphasizes the role of the landscape in all *Canzonieri*, is equally notable. Studies have described the importance of the landscape in Petrarch's texts, although this is in a famous letter describing Petrarch's ascent of *Mont Ventoux*, rather than in his poems.[17] Yet in Petrarch's *Canzoniere* the landscape represents more than an earthly opposite of the idealized, heavenly Laura: as Elissa Tognozzi argues, the landscape in Petrarch's *Canzoniere* corresponds with the lyrical I's psychological and sentimental condition (Tognozzi 1998). The fact that nouns belonging to the semantic category of the landscape are among the preferred nouns in the *Canzonieri* since Petrarch, however, suggests that the relation between lyrical I, beloved and landscape may be a relevant motif of Italian Petrarchism in general.

Perhaps even more illuminating than the preferred words is the second aspect, the avoided ones. Among the latter there are surprisingly many variants of a notion that forms the basis of all love poetry: love, as the ratio of the preferred and avoided writing variants of *amore*/*amare* ('love/to love') in Table 5 shows.

[17] For a study on the landscape in Petrarch's Canzoniere, see Stierle (1979); the critical review of that study in König (1980); Güntert (2012). For literature on Petrarch's ascent of Mont Ventoux, see e.g. Kablitz (1994); Pfeiffer (1997); Ulmer (2010, 34–47); Campana Comparini (2010); Behrens (2016). For contributions on Petrarch's conception of the landscape in general, see Luciani and Mosser (2009); Tosco (2011, 103–30).

Table 5 Avoided and preferred variants of *amore/amare* ('love/to love') according to the contrastive analyses

Analysis	Writing variants of *amore/amare* preferred	Writing variants of *amore/amare* avoided
A vs B, Craig's Zeta	0	8
A vs B, Eder's Zeta	1	5
A vs C, Craig's Zeta	0	8
A vs C, Eder's Zeta	0	7

According to the word list based on Craig's Zeta, compared to the love poetry written before, Petrarch's *Canzoniere* avoids eight variants and prefers none.[18] The situation is similar for Eder's Zeta (five variants avoided, one preferred).[19] A similar picture emerges for the *Canzonieri* written after Petrarch. In comparison to the love poetry written before Petrarch, according to Craig's Zeta eight variants are avoided and none is preferred,[20] and for Eder's Zeta seven are avoided and none is preferred. [21]

The results of the contrastive analyses thus indicate that the *Canzonieri* since Petrarch, one of the most important role models for love poetry in history, mostly avoid addressing their central theme literally: love. A comparison of the cumulative frequencies of the occurring variants of *amore*/*amare* ('love/to love') supports this finding (see Table 6).[22]

Table 6 Cumulative frequencies of the occurring variants of *amore/amare* x 100

Part	Total number of words	Absolute frequency	Relative frequency
A	161,270	1,398	0.866869226
B	69,773	187	0.268011982
C	435,367	1,560	0.358318384

Table 6 illustrates that the cumulative relative frequencies of the occurring variants of *amore*/*amare* ('love/to love') are considerably higher in the love poetry before Petrarch than in Petrarch's *Canzoniere* and in the *Canzonieri* after Petrarch. One possible explanation could be the highly personal dimension that characterizes Petrarch's *Canzoniere*. Through his preoccupation with his beloved Laura, the lyrical I also comes to terms

[18] The variants avoided are: ama, amante, amanza, amare, amato, amo, amore, amoroso.
[19] The variants avoided are: ama, amante, amare, amato, amo. The variant preferred is amorosi.
[20] The variants avoided are: amante, amanti, amanza, amare, amato, amo, amorosa, amoroso.
[21] The variants avoided are: amante, amanti, amare, amato, amo, amorosa, amoroso.
[22] To determine the cumulative relative frequencies, the tool TXM was used.

with his own emotional state, so that the poetic description of his love for Laura leads to a profound exploration of his personal feelings (Geyer 2009; Wehle 2009). The lyrical I's feelings are verbalized by expressions that go beyond the semantic field of love in the literal sense, which could also explain the occurrence of words from the semantic fields of body and landscape (Tognozzi 1998).

4. Conclusion

The aim of this study was to find elements that make it possible to distinguish Petrarchistic texts from non Petrarchistic literature through quantitative stylistic analyses of a corpus of Italian poetry. These analyses lead to two main results. On the one hand, some aspects that were already described in research on Petrarchism could be confirmed: first, a preference for words which express pain and pleasure in the *Canzonieri* since Petrarch, contributing to a dichotomy fundamental for the concept of *dolendi voluptas*. Second, a predilection for expressions referring to sweetness, which can be regarded an allusion to the poetry of *Dolce Stil Novo*. Third, the existence of a group of preferred words with a spiritual connotation in all *Canzonieri* since Petrarch, implying the idealization of the beloved woman. Fourth, a preference for nouns that allude to poetry. On the other hand, some elements, which have received only little attention in the scholarly discourse about Italian Petrarchism so far, could be revealed. Regarding the parts of speech, the results show a predilection for nouns at the expense of verbs in all *Canzonieri* since Petrarch. The nouns preferred can be divided into five categories, two of which are notable, because they were only occasionally taken into account in the research literature on Italian Petrarchism: a group of expressions belonging to the semantic field of the human body and a number of words from the semantic field of the landscape. Moreover, and maybe most notably, the results of the contrastive analyses and the cumulative frequencies suggest that the *Canzonieri* since Petrarch mostly avoid using variants of the word 'love', which suggests that these texts tend to avoid addressing their fundamental theme directly.

In the future, it would be interesting to analyze later collections of poems considered Petrarchistic in contrast to poetry of the same time which is not, e.g. collections belonging to other poetic schools. Moreover, it would be enlightening to study potentially Petrarchistic literature in other languages. This would help to answer the question whether it is possible to distinguish diverse kinds of Petrarchism in different areas or time periods. Only a broad study of European poetry from different periods and in diverse languages can provide a clearer picture of Petrarch's impact on world literature and therefore a more comprehensive understanding of Petrarchism.

ORCID®

Jan Rohden https://orcid.org/0000-0001-7998-8629

References

Alighieri, Dante. 1973. *Rime*, edited by Gianfranco Contini. 3rd ed. Torino: Einaudi.

Baldacci, Luigi. 1957. *Il Petrarchismo Italiano Nel Cinquecento.* Milano: Ricciardi.

Barolini, Teodolinda. 1989. "The Making of a Lyric Sequence: Time and Narrative in Petrarch's Rerum Vulgarium Fragmenta." *Modern Language Notes* 104: 1–38.

Beaudouin, Valérie, and Francois Yvon. 1996. "The Metrometer: a Tool for Analysing French Verse." *Literary and Linguistic Computing* 11 (1): 23–31.

Behrens, Christoph. 2016. "Gehört die Landschaft der Moderne? Zur historischen Dynamik des Mont Ventoux." In *Räume der Romania: Beiträge zum 30. Forum Junge Romanistik*, edited by Nadine Chariatte, Corinne Fournier Kiss, and Etna R. Krakenberger, 137–51. Frankfurt a. M.: Lang.

Bernsen, Michael. 2011. "Der Petrarkismus, eine lingua franca der europäischen Zivilisation." In *Der Petrarkismus – ein europäischer Gründungsmythos*, edited by Michael Bernsen and Bernhard Huss, 15–30. Göttingen: V&R unipress.

Blasi, Nicola de. 2008. *Piccola storia della lingua italiana.* Domini. Napoli: Liguori.

Burrows, John. 2002. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17: (3) 267–87. https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/17.3.267.

Burrows, John. 2007. "All the Way Through: Testing for Authorship in Different Frequency Strata." *Literary and Linguistic Computing* 22 (1): 27–47. https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqi067.

Campana Comparini, Francesca. 2010. "Petrarca tra Antichità e Modernità: La nascita del Paesaggio." *Città di Vita: Bimestrale di Religione Arte e Scienza* 65: 389–96.

Contini, Gianfranco, ed. 1960. *Poeti del Duecento. La letteratura italiana: storia e testi 1–2*. Milano: Ricciardi.

Craig, Hugh, and Arthur F. Kinney, eds. 2009. *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. "Stylometry with R: A Package for Computational Text Analysis." *R Journal* 8 (1): 107–21. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

Fornasiero, Serena. 2001. *Petrarca: Guida al Canzoniere.* Roma: Carocci.

Forster, Leonard. 1969. *The Icy Fire: Five Studies in European Petrarchism.* Cambridge: Cambridge University Press.

Friedrich, Hugo. 1964. *Epochen der Italienischen Lyrik.* Frankfurt a. M.: Klostermann.

Geyer, Paul. 2009. "Petrarcas Canzoniere als Bewusstseinsroman." In *Petrarca und die Herausbildung des modernen Subjekts*, edited by Paul Geyer and Kerstin Thorwarth, 109–51. Göttingen: V&R Unipress.

Güntert, Georges. 2012. "Natur ist nie nur Natur: topographische Elemente im Canzionere Petrarcas und im spanischen Petrarkismus des 16. Jahrhunderts." In *Begriff und Darstellung der Natur in der spanischen Literatur der Frühen Neuzeit*, edited by Wolfgang Matzat, 207–23. München: Fink.

Hempfer, Klaus W. 1987. "Probleme der Bestimmung des Petrarkismus. Überlegungen zum Forschungsstand." In *Die Pluralität der Welten: Aspekte der Renaissance in der Romania*, edited by Wolf-Dieter Stempel and Karlheinz Stierle, 253–77. München: Fink.

Hempfer, Klaus W. 1991. "Intertextualität, Systemreferenz und Strukturwandel: Die Pluralisierung des erotischen Diskurses in der italienischen und französischen Renaissance-Lyrik (Ariost, Bembo, Du Bellay, Ronsard)." In *Modelle des literarischen Strukturwandels*, edited by Michael Titzmann, 7–44. Tübingen: Niemeyer.

Hempfer, Klaus W., Gerhard Regn, and Sunita Scheffel, eds. 2005. *Petrarkismus-Bibliographie: 1972–2000.* Stuttgart: Steiner.

Hoffmeister, Gerhart. 1973. *Petrarkistische Lyrik.* Stuttgart: Metzler.

Hoover, David L. 2008. "Searching for Style in Modern American Poetry." In *Directions in Empirical Literary Studies*, edited by Sonia Zyngier, Marisa Bortolussi, Anna Chesnokova, and Jan Auracher, 211–27. Amsterdam: Benjamins.

Huß, Bernhard. 2001. "'Cantai colmo di gioia, e senza inganni'. Benedetto Varchis Sonetti (Parte Prima) im Kontext des italienischen Cinquecento-Petrarkismus." *Romanistisches Jahrbuch* 52 (1): 133–57.

Ilsemann, Hartmut. 2019. "Forensic Stylometry." *Digital Scholarship in the Humanities* 34 (2): 335–49. https://doi.org/10.1093/llc/fqy023.

Jannidis, Fotis, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2015. "Improving Burrows' Delta. An Empirical Evaluation of Text Distance Measures." In *Digital Humanities 2015: Conference Abstracts*. Sydney: Universirsity of Western Sydney. http://dh2015.org/abstracts.

Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History.* Urbana: University of Illinois Press.

Kablitz, Andreas. 1994. "Petrarcas Augustinismus und die écriture der Ventoux-Epistel." *Poetica* 26: 31–69.

König, Bernhard. 1980. "Petrarcas Landschaften. Philologische Bemerkungen zu einer neuen Deutung." *Romanische Forschungen* 92: 251–82.

Leopold, Stephan. 2009. *Die Erotik der Petrarkisten: Poetik, Körperlichkeit und Subjektivität in romanischer Lyrik Früher Neuzeit.* München: Fink.

Luciani, Domenico, and Monique Mosser, eds. 2009. *Petrarca e i suoi luoghi: Spazi reali e paesaggi poetici alle origini del moderno senso della natura.* Treviso: Fondazione Benetton Studi Ricerche.

Marnoto, Rita. 2015. *O Petrarquismo Portugues do Cancioneiro Geral a Camões.* Lisboa: Imprensa Nacional – Casa da Moeda.

Morales Saravia, José. 1998. "Vanitas y petrarquismo en el soneto XXIII de Garcilaso de la Vega." *Iberoromania* 47: 47–71.

Moretti, Franco. 2013. *Distant Reading.* London, New York: Verso.

Nardone, Jean-Luc. 1998. *Pétrarque et le pétrarquisme.* Paris: PUF.

Navarro Colorado, Borja. 2018a. "A Metrical Scansion System for Fixed-Metre Spanish Poetry." *Digital Scholarship in the Humanities* 33 (1): 112–27.

Navarro Colorado, Borja. 2018b. "On Poetic Topic Modeling: Extracting Themes and Motifs from a Corpus of Spanish Poetry." *Frontiers in Digital Humanities* 5. https://doi.org/10.3389/fdigh.2018.00015.

Panvini, Bruno, ed. 1962–64. *Le Rime Della Scuola Siciliana.* Firenze: Olschki.

Petrarca, Francesco. 2015. *Canzoniere*, edited by Marco Santagata. 2nd ed. Milano: Mondadori.

Pfeiffer, Jens. 1997. "Petrarca und der Mont Ventoux (zu Familiares IV,1)." *Germanisch-Romanische Monatsschrift. N. F* 47: 1–24.

Pirovano, Donato. 2014. *Il dolce stil novo.* Roma: Salerno.

Pyritz, Hans. 1963. *Paul Flemings Liebeslyrik: zur Geschichte des Petrarkismus.* Göttingen: V&R.

Quondam, Amedeo, Beatrice Alfonzetti, and Stefano Asperti. 2019. "Biblioteca Italiana." http://www.bibliotecaitaliana.it/ (Accessed September 27, 2019).

Rebora, Simone, J. Berenike Herrmann, Gerhard Lauer, and Massimo Salgaro. 2019. "Robert Musil, a War Journal, and Stylometry: Tackling the Issue of Short Texts in Authorship Attribution." *Digital Scholarship in the Humanities* 34 (3): 582–605. https://doi.org/10.1093/llc/fqy055.

Regn, Gerhard. 1987. *Torquato Tassos zyklische Liebeslyrik und die petrarkistische Tradition: Studien zur Parte Prima d. Rime (1591/1592).* Tübingen: Narr.

Regn, Gerhard. 1993. "Systemunterminierung und Systemtransgression. Zur Petrarkismus-Problematik in Marinos Rime amorose (1602)." In *Der Petrarkistische Diskurs: Spielräume und Grenzen; Akten des Kolloquiums an der Freien Universität Berlin, 23.10.–27.10.1991*, edited by Klaus W. Hempfer, 255–81. Stuttgart: Steiner.

Regn, Gerhard. 2013. "Petrarkismus." In *Historisches Wörterbuch der Rhetorik Online*, edited by Gert Ueding. Berlin/Boston: De Gruyter. https://www.degruyter.com/view/HWRO/petrarkismus?pi=0&moduleId=common-word-wheel&dbJumpTo=Petra. (Accessed September 27, 2019).

Rhody, Lisa M. 2012. "Topic Modeling and Figurative Language." *Journal of Digital Humanities* 2 (1). http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody.

Rigo, Paolo. 2017. "Petrarca e il corpo: una ricognizione del tema." *Arzanà* 19: 55–77. https://doi.org/10.4000/arzana.1050.

Rojas Castro, Antonio. 2018. "¿Cuántos 'Góngoras' podemos leer? Un análisis contrastivo de la poesía de Luis de Góngora." *e-Spania* 29. https://doi.org/10.4000/e-spania.27448.

Santagata, Marco. 1993. *I frammenti dell'anima: storia e racconto nel Canzoniere di Petrarca.* Bologna: il Mulino.

Schiffer, James. 2000. "Shakespeare's Petrarchism." In *Shakespeare's Sonnets: Critical Essays*, edited by James Schiffer, 163–83. New York, NY: Garland.

Schneider, Ulrike. 2007. *Der weibliche Petrarkismus im Cinquecento: Transformationen des lyrischen Diskurses bei Vittoria Colonna und Gaspara Stampa.* Stuttgart: Steiner.

Schöch, Christof. 2017. "Quantitative Analyse." In *Digital Humanities: Eine Einführung*, edited by Fotis Jannidis, Hubertus Kohle, and Malte Rehbein, 279–98. Stuttgart: Metzler.

Schöch, Christof. 2018. "Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie." In *Quantitative Ansätze in den Literatur- und Geisteswissenschaften*, edited by Toni Bernhart, Marcus Willand, Sandra Richter, and Andrea Albrecht, 77–94. Berlin, Boston: De Gruyter.

Schöch, Christof. 2019. "pyzeta." Accessed September 27, 2019. https://github.com/cligs/pyzeta.

Seitschek, Gisela. 2014. "Von der Donna angelicata zur gloriosa Beatrice. Stilo della loda oder Lobpreis der Herrin beim frühen Dante und den Stilnovisten." In *Das literarische Lob: Formen und Funktionen, Typen und Traditionen panegyrischer Texte*, edited by Norbert P. Franz, 55–84. Berlin: Duncker & Humblot.

Stierle, Karlheinz. 1979. *Petrarcas Landschaften: Zur Geschichte ästhetischer Landschaftserfahrung.* Krefeld: Scherpe.

Suerbaum, Werner. 1972. "Poeta laureatus et triumphans." *Zeitschrift für Sprach- und Literaturwissenschaft* 5: 293–328.

Tognozzi, Elissa. 1998. "La natura come corrispondente esterno ad una condizione psicologica-sentimentale nel Canzoniere (con particolare enfasi su *Zephiro Torna*)." *Forum Italicum* 32: 51–62.

Tosco, Carlo. 2011. *Petrarca: Paesaggi, città, architetture.* Macerata: Quodlibet.

Ulmer, Birgit. 2010. *Die Entdeckung der Landschaft in der italienischen Literatur an der Schwelle zur Moderne.* Frankfurt a. M.: Lang.

Warning, Rainer. 1987. "Petrarkistische Dialogizität am Beispiel Ronsards." In *Die Pluralität der Welten: Aspekte der Renaissance in der Romania*, edited by Wolf-Dieter Stempel and Karlheinz Stierle, 327–58. München: Fink.

Wehle, Winfried. 2009. "Im Labyrinth der Leidenschaften: zur Struktureinheit in Petrarcas Canzoniere." In *Petrarca und die Herausbildung des modernen Subjekts*, edited by Paul Geyer and Kerstin Thorwarth, 71–106. Göttingen: V&R Unipress.

Wilkins, Ernest H. 1948. "The Evolution of the Canzoniere of Petrarch." *Publications of the Modern Language Association of America* 63 (2): 412–55.

Zaccagnini, Guido, and Amos Parducci, eds. 1915. *Rimatori Siculo-Toscani del Dugento.* Scrittori d'Italia. Bari: Laterza.

Stylometry and Spanish Golden Age Theatre
An Evaluation of Authorship Attribution in a Control Group of One Hundred Undisputed Plays

Álvaro Cuéllar

Abstract The aim of this study is to perform an evaluation of one hundred Spanish Golden Age theatre plays of undisputed authorship using the R package *stylo*, the stylometric analysis tool developed by Eder, Rybicki, and Kestemont (2016). In this paper we will determine which algorithms obtain best results on authorial classification (method, MFW, culling, and word n-grams). We will also evaluate the text length at which stylometry begins to be an effective diagnostic tool for authorship attribution in our corpus. This cross-validation evaluation can serve future analysis of similar corpora and will show the possibilities of applying stylometry to Spanish Golden Age theatre, which presents many cases of dubious authorship.

Keywords stylometry, Spanish Golden Age theatre, *stylo*, author identification, text length, most frequent words (MFW), culling, n-grams

1. Spanish Golden Age Theatre Authorship Issues

Menéndez Pidal (1949, xxviii–xxxii) explained that the frequency of authorship issues in Spanish literature can be attributed to three factors: its tendency to anonymity, its collectivism, and its collaborative writing. These three elements, though present throughout the history of Spanish literature, converge particularly in the Spanish Early Modern period, with canonical prosaic texts such as *La vida de Lazarillo de Tormes* or the apocryphal *Quijote* by Avellaneda serving as famous examples of this authorial

Cuéllar, Álvaro: Stylometry and Spanish Golden Age Theatre. An Evaluation of Authorship Attribution in a Control Group of One Hundred Undisputed Plays, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 101–117.
DOI: https://doi.org/10.17885/heiup.1157.c19368

uncertainty.[1] Nevertheless, it is in the vast theatrical production of the Spanish Golden Age—over 10,000 texts are estimated to have been written for the stage during this period, although it is difficult to establish an exact number—where we find the highest quantity of authorship disputes that need to be resolved.

Issues of authorship in the field of Spanish Golden Age theatre are immeasurable and the academic literature surrounding them is substantial. For example, Tirso de Molina's famous prologue to the *Segunda parte de sus comedias* (1635) includes the following statement:

> Destas doze comedias quatro, que son mias, en mi nombre, y en el de los dueños de las otras, ocho (que no se porque infortunio suyo, siendo hijas de tan ilustres padres, las echaron a mis puertas) las que restan.
>
> "Of these twelve plays, only four are mine, in my name, and the remaining eight in the name of their owners (for I do not know why, being the daughters of such illustrious fathers, their misfortune put them at my doorstep)."[2]

This volume includes the play *El condenado por desconfiado*, crucial for its theological ideas, whose authorship has been in doubt precisely because of this statement by the playwright.

Calderón de la Barca is the greatest exponent of authorship confusion in Golden Age theatre, to the extreme that he is known to have even rejected comedies that he had actually written.[3] We know that he wrote plays that he did not include in several lists of his authentic texts that he prepared toward the end of his life. Moreover, some of his plays appear in a list of apocryphal works that circulated in his name that Calderón included at the beginning of his *Parte cuarta* (1672), and which is inexact in several cases. This situation allows us to appreciate the complexities presented by this extreme case: we cannot even take for granted the information provided by authors regarding their own production. To this point, Germán Vega García-Luengos sums up the extent to which authorial uncertainty surrounds the attributed production of the two major playwrights of Spain's Golden Age theatre:

[1] Some articles have recently been published addressing these issues with digital methods, especially using stylometry. See De la Rosa and Suárez (2016) for *La vida de Lazarillo de Tormes*, and Fradejas Rueda (2016), Rißler-Pipka (2016), and Rodríguez López-Vázquez (2018) for the apocryphal *Quijote*.

[2] All translations are my own.

[3] For an explanation about Calderón's textual issues and the problematics of his lists, see Coenen (2009a; 2009b; and 2019) or Vega García-Luengos (2008)

> De las 506 comedias atribuidas en algún momento a Lope de Vega que estudian S. G. Morley y C. Bruerton en su *Cronología* (Morley y Bruerton 1968), 316 serían auténticas, 27 probablemente, 73 se califican como dudosas y 90 como ajenas. Calderón, el otro grande del teatro español, aún superaría estos porcentajes, tal como se apunta ya en los testimonios más tempranos: las 108 piezas de sus nueve partes se ven superadas por las 115 de las listas de 'comedias supuestas' que Vera Tassis dio en la *Verdadera quinta parte* (1682) y en la *Séptima* (1683).

> "Of the 506 plays attributed at some point in time to Lope de Vega that S. G. Morley and C. Bruerton study in their *Cronología*, 316 would be authentic, 27 probably so, 73 dubious, and 90 spurious. Calderón, the other great writer of Spanish theatre, even exceeds these percentages, as already indicated in the earliest testimonies: the 108 texts of his nine volumes are exceeded by the 115 plays in the lists of 'supposed comedies' that Vera Tassis included in the *Verdadera quinta parte* (1682) and the *Séptima* (1683)." (2002, 16)

We may think that authorship problems affect only secondary works of Golden Age theatrical production, those that have exerted less sway of interest and influence in the literary tradition, effectively passing unnoticed until a scholar decides to recover them. However, some of the most celebrated plays fall into this category of dubious authorship, such as *Tan largo me lo fiáis* and *El burlador de Sevilla*, the two plays that started the Don Juan myth and whose authorship has been profusely debated.[4]

Among the possible explanations for the authorship problems that Spanish Golden Age theatre presents, the economic factor seems to be key. As a form of property, the text legally belonged to the theatrical director or the bookseller, not to the author of the play. The owners could change the real author's name to attract audiences, or add or delete verses from the text. Although it may seem surprising today, Vega García-Luengos asserts that writers, "in general, did not make any great effort to resolve this state of affairs" (2009, 97; "en líneas generales, no se aprecia que dedicaran demasiados esfuerzos a solucionar tal estado de cosas.") for two main reasons. First, it is worth noting that poets wrote their plays for the stage, and that publication was considered secondary. Second, the economic and ownership status of the theatrical text meant that textual stability was not legally protected or assured: playwrights did not keep the publishing rights when selling their original manuscripts with the final version of their plays to a theatrical company and, when they did recover those rights, there was no guarantee that the text would match the original.

[4] Rodríguez López-Vázquez has been studying this issue for decades in numerous articles and books. For instance, see Ródríguez López-Vázquez (1983).

2. Stylometry

Stylometry is a technique that tries to automatically compare texts by their *style* and, among its many applications, it can be useful for authorship attributions of literary works. Within the category of *style*, which can be quite abstract, we are going to be working specifically with the concrete criterion of word usage frequency.[5] Stylometry is based on this simple hypothesis: each writer uses some words more frequently than others, and these data can be used to establish relationships between different texts. For instance, in *La dama boba* by Lope de Vega, for each 100 words, 4.90 are *que*, 3.64 are *de*, and 2.90 are *y*, while in *Don Gil de las calzas verdes* by Tirso de Molina, these proportions are 4.86, 3.91 and 3.03. These tiny differences are not usually relevant to researchers; nevertheless, they are decisive for stylometry. Stylometry uses these proportions to establish relations of proximity between the different texts or classify them in a specific group through statistical models. Usually, stylometric analysis is performed on the basis of hundreds of these words and their proportions. Sometimes other parameters are more relevant, such as analyzing groups of words (n-grams) or accepting only words found in a certain percentage of texts, which is known as culling.

In the last years, stylometry has been extensively applied to authorship issues of Spanish Golden Age theatre. Some good examples are the attribution of *Siempre ayuda la verdad* to Lope de Vega (García-Reidy 2019); the study of *La conquista de Jerusalén* (Cerezo Soler and Calvo Tello 2019), where stylometric analysis seems to corroborate the consideration of Cervantes as the author of the text; the study of *La segunda Celestina*, a text attributed to Sor Juana Inés de la Cruz by Schmidhuber de la Mora some decades ago (Hernández Lorenzo and Byszuk 2019); the study of the *entremés* titled *Los mirones*, a short type of play that was performed in between acts of a play (Blasco 2019); and the stylometric study of Agustín Moreto y Cavana's production, which is one of the most complex repertoires of the period because of the author's tendency toward collaboration and rewriting (Ulla Lorenzo, Martínez Carro, and Calvo Tello 2020).

Professor Vega García-Luengos and I are currently developing the project *ETSO: Estilometría aplicada al Teatro del Siglo de Oro* (Stylometry applied to the Spanish Golden Age Theatre), https://etso.es/ (Cuéllar and Vega 2017–2022), a collaborative project that currently includes dozens of scholars and seeks to build the biggest possible digital repository of Spanish Golden Age plays for the application of stylometric analysis. Our working philosophy is not to analyze a specific authorship issue within a controlled corpus. Instead, we use all the texts we have and let the corpus show unexpected connections. This is exactly what happened with *La monja alférez*, a play traditionally attributed to Juan Pérez de Montalbán, Lope de Vegas's disciple. We were not specifically

[5] For an introduction to stylometry and the stylo package, see Fradejas Rueda (2019).

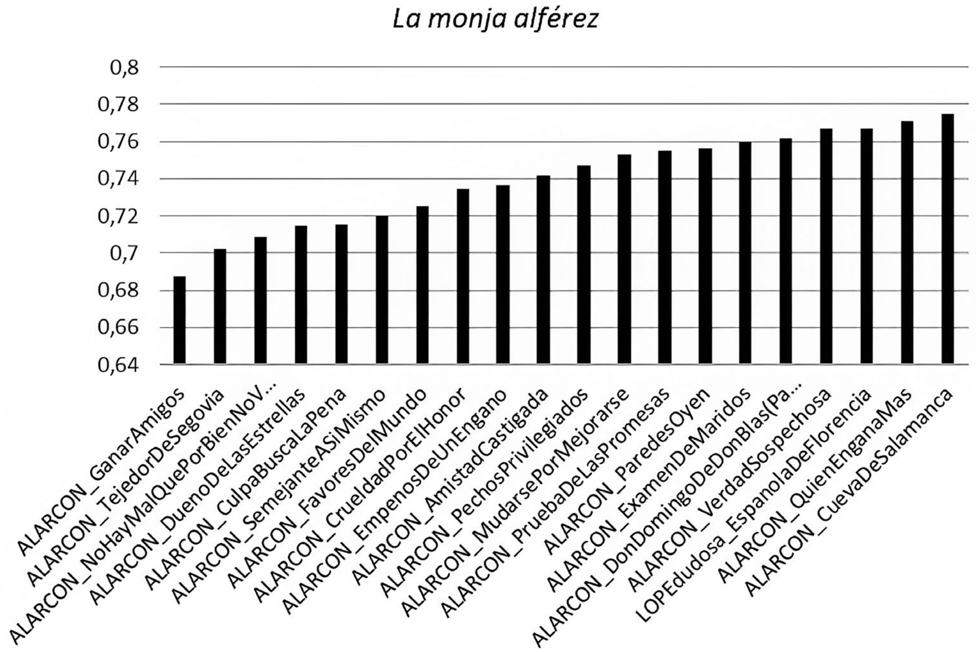


Fig. 1 20 most proximate plays to *La monja alférez* in a corpus of 1,028 texts (ETSO, September 1, 2020), Stylo (500 MFWs, Classic Delta, zero percent culled). (Vega García-Luengos 2021)

researching authorship issues surrounding this play, but when we ran an analysis of all the new texts we had collected, the following (Figure 1) were the plays that are most lexically proximate in a corpus of more than 1,000 texts.

As the results indicate, the closest plays were mostly written by Juan Ruiz de Alarcón, a quite prolific author born in New Spain. Thanks to this surprising result, Vega García-Luengos decided to isolate the historical and philological analysis of *La monja alférez* in a study that presents solid and variegated evidence confirming the attribution of the play to Juan Ruiz de Alarcón (2021). Vega García-Luengos and I are applying this same analytical procedure to other interesting cases that we hope to publish soon as a book-length anthology. We also use ETSO to assist scholars with specific authorship issues, such the collaborative writing in Moreto's *La adúltera penitente* (Moreto 2019) or *Empezar a ser amigos* (Demattè 2019), or the attribution dispute in *Las dos bandoleras* (Madroñal 2019).[6]

[6] For a complete description of the collaboration of ETSO with different scholars, visit http://etso.es/repercusion/.

One of the most effective applications to perform stylometric analysis is *stylo*.[7] This tool, which has to be installed as a package in R (R Core Team 2020), a programming language and free software environment for statistical computing, has been developed by the Computational Stylistics Group, a team formed by members of the universities of Krakow and Antwerp, headed by researchers Eder, Rybicki, and Kestemont (2016).

3. Corpus Selection and Homogenization

Before applying stylometry to texts that present attribution problems, it is necessary to ensure that our methods work with a corpus of controlled works. The aim of this article is to precisely verify that stylometry works with a control group of undisputed plays and determine which algorithms obtain better results.

Although it is impossible to ensure with total certainty that the plays chosen here were really written by the supposed authors—remember that they sometimes lie to us about their production—we have selected only works included in *Partes*, that is, volumes of usually twelve plays, whose publication was controlled by the authors or their relatives.[8]

To check the effectiveness of stylometry, we turn to a corpus of nine canonical authors of Golden Age theatre: Cervantes (1547–1616), Lope de Vega (1562–1635), Guillén de Castro (1569–1631), Tirso de Molina (1579–1648), Ruiz de Alarcón (1581?–1639), Calderón de la Barca (1600–1681), Pérez de Montalbán (1602–1638), Rojas Zorrilla (1607–1648), and Agustín Moreto (1618–1669). As we mentioned, we will only consider plays from *Partes de comedias* published by the authors or their relatives, which amounts to one hundred texts.

To lexically compare the texts, Vega García-Luengos and I had to homogenize the corpus. The original seventeenth-century orthography was quite unstable, and our texts come from different sources and editions, so we decided to modernize and homogenize the orthography to the current Spanish rules.[9] In addition, we removed stage directions since we consider that they may negatively affect the authorship analysis, given our understanding that they are elements that were usually not written by playwrights but by theatrical producers and therefore are not related to the unconscious style of the authors' lines.[10]

[7] Here we are using version 0.7.4.

[8] For a complete description, see Vega García-Luengos (2010).

[9] See Real Academia Española y Asociación de Academias de la Lengua Española (2010).

[10] See Rodríguez-Gallego (2018) for a study about how Vera Tassis modified the stage directions in seven plays by Calderón de la Barca when he published them.

4. Corpus Evaluation

In this section, we will check to ensure that the corpus of undisputed works responds satisfactorily to a test with different settings. Only if the corpus shows reliable results when authorship is not in doubt can stylometry be considered a useful tool for works of dubious authorship.

We have a corpus of one hundred plays. The test we will apply is known as leave-one-out cross-validation, which consists of running an analysis of all the plays within the corpus as though each were of unknown authorship, forcing the algorithm in each instance to choose an author of the text in question, and making sure that the resulting automatic classification of each play in fact corresponds to the correct one of the nine authors included in the corpus.[11] In other words, we separate a text from the rest of the corpus, the work becomes our authorship *problem*, and it must be classified. The machine is asked to classify this work among the nine possible groups according to the frequency of its words and the parameters established. Since we have carefully selected our corpus to consist only of plays with undisputed authorship, this test will help us check if the tool is accurate and, if so, to what degree. The three parameters are:

1. Classification method. There are many methods to perform the analysis, such as SVM (Support Vector Machine), NSC (Nearest Shrunken Centroids), Naive Bayes, k-NN (k-nearest neighbors), Delta[12] with different distance measures (Classic, Eder, etc.), and other distances (Canberra, Cosine, etc.).[13]
2. Most frequent words (MFW). The application will take into account the most frequent words considering all the texts for the analysis. This means that function words, such as prepositions, conjunctions or auxiliary verbal forms are usually the most frequent ones, and words with meaning are secondary to them. Thus, an MFW of 100 means that the machine uses the 100 most frequent words in the texts as a whole. An MFW of 5,000 means that the machine uses the 5,000 most frequent words.
3. Culling. Once the machine has listed all the words by their frequency, we may want to dismiss those that occur in a tiny group of texts. We probably do not want to find stylometric closeness between two plays which share characters names, places, or plots. A culling of 30 means that the machine uses words that are at least in 30 percent of the texts. A culling of 100 implies that the machine uses words that appear in all the texts, which is usually a very small number.

[11] For a step-by-step tutorial about how to program a cross-validation in *stylo*, see Cuéllar (2018).

[12] The usage of distance measurements is not a classification method comparable to the others because it usually creates a matrix of distances between texts. Nevertheless, *stylo* uses this matrix to classify the texts based on the closest neighbors, so it can be eventually considered as a classification method.

[13] For a deeper explanation, see Calvo Tello (2016) or Schöberlein (2017).

4.1 Results by Varying the Classification Method, MFW, and Culling

We can see in Figures 2, 3, 4 and 5 the classification results when varying the methods and the MFW. A result of 100 percent indicates a perfect classification: using leave-one-out cross-validation, all one hundred plays have been correctly attributed to their respective authors among the nine options given. A result of 94 percent means that the machine has correctly classified this percentage of texts.

By observing the figures, we can see that the global results are remarkably positive. Most of the methods offer classification results over 95 percent. Nevertheless, there are some methods that perform more poorly with our specific test set, such as Cosine, Euclidean and Naive Bayes. The second observation we can make is that the MFW variation between 100 MFW and 5,000 MFW does not seem to affect the classification results in a significant way.

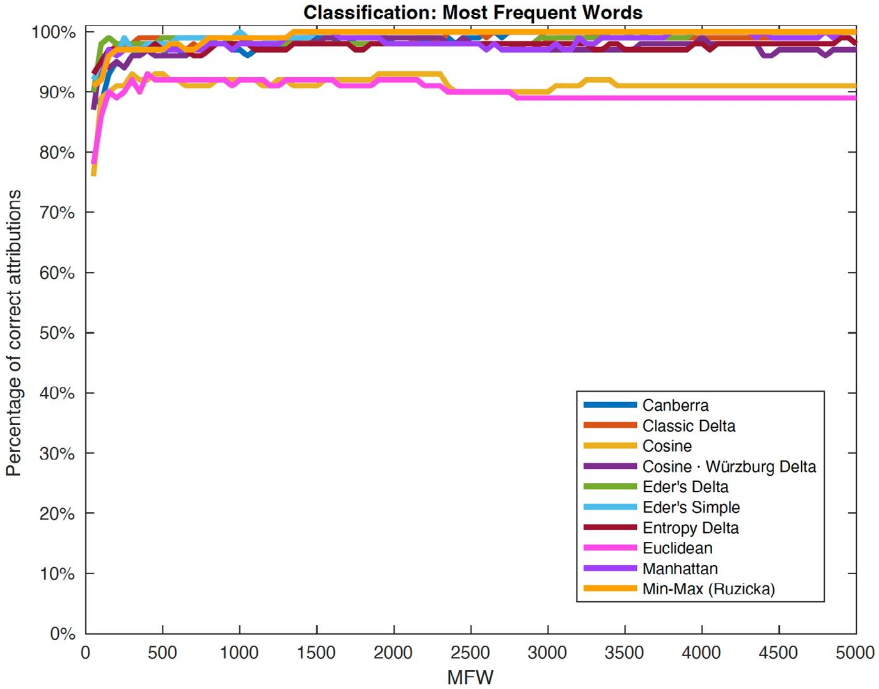


Fig. 2 Percentage of correct attributions with ten different distance methods by varying the MFW (Cuéllar, CC BY).

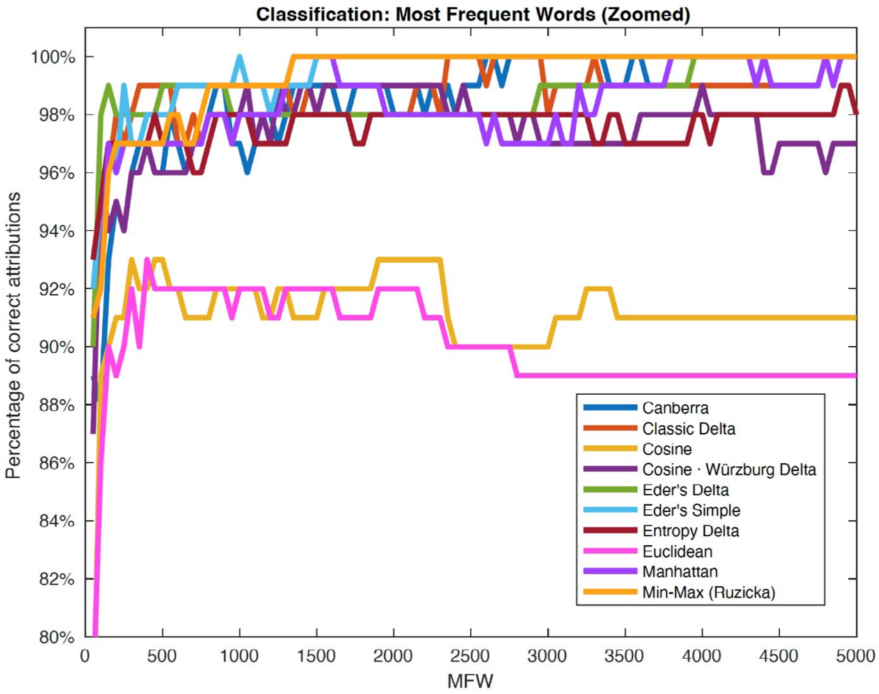


Fig. 3 Zoomed percentage of correct attributions with ten different distance methods by varying the MFW (Cuéllar, CC BY).

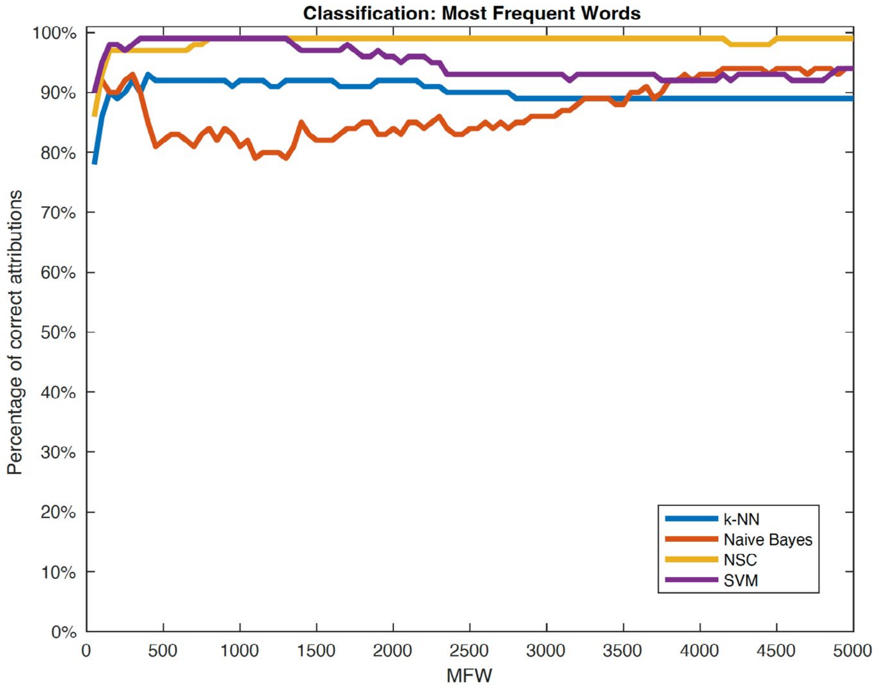


Fig. 4 Percentage of correct attributions with four different classification methods by varying the MFW (Cuéllar, CC BY).

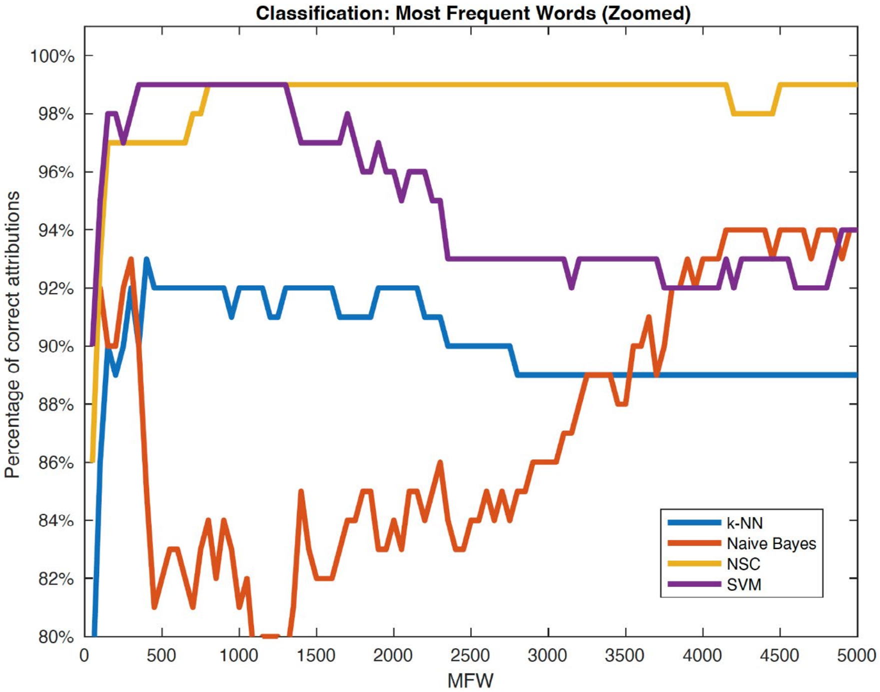


Fig. 5 Zoomed percentage of correct attributions with four different classification methods by varying the MFW (Cuéllar, CC BY).

We can now run an analysis choosing one method, such as Classic Delta, and varying the culling. We start with 5,000 MFW and a culling of 0 percent. We are going to increase the culling by steps of 0.5 percent until we reach a 100 percent of culling. This culling increment is going to affect the number of MFW available because there are not enough common words, so this number is going to reduce consequently. For example, with a 20 percent culling we perform the analysis with 3,597 words and with a 80 percent culling we perform the analysis with 603 words. It is complex, therefore, to understand if the results we obtain are due the culling or to the reduction of parameters. Nonetheless, we can see the results in Figure 6, they remain almost constant when varying the culling.[14]

The results do not seem to change significantly when we vary the MFW or the culling. It does not seem to matter if we use the 100, 500 or 5,000 most frequent

[14] Results for the other methods are quite similar: culling does not seem to significantly affect the results of the automatic classification.

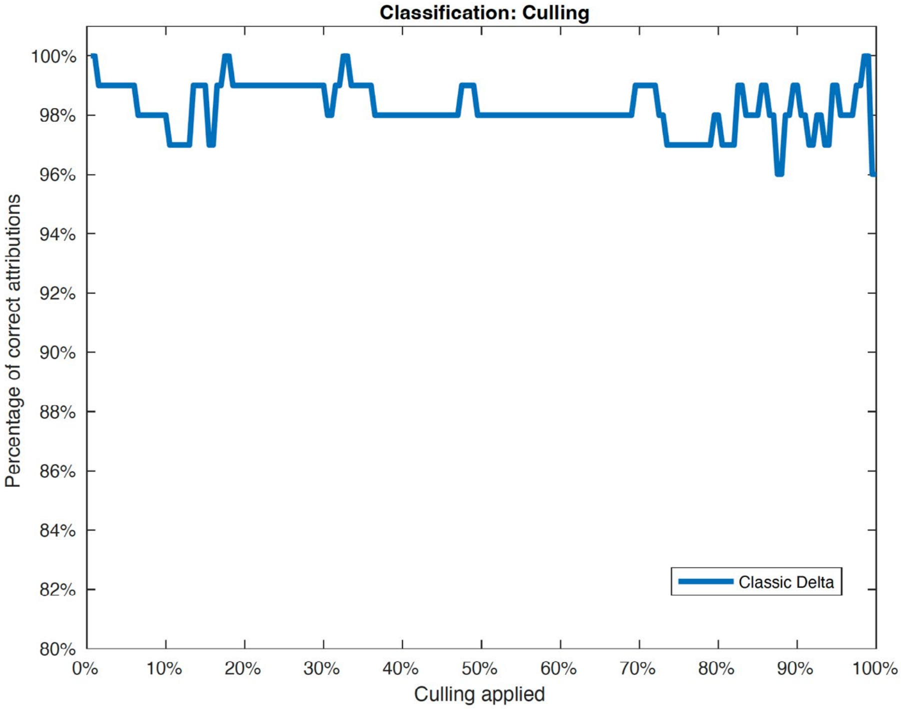


Fig. 6 Percentage of correct attributions by varying the culling. Classic Delta (Cuéllar, CC BY).

words, or if we use words that have to appear in a specific percentage of texts: the results are still extraordinary.

We can get a glimpse of the potential that these results suppose: if we now introduce a play with dubious authorship and if it is classified with one of these nine authors, we can be on the trail of an attribution. Of course, it is possible that it does not belong to any of the authors or that the system has failed, so we must be cautious and take into account additional evidence.

4.2 Results by Varying the Word N-grams

For this test, we will compare the results of using just one word (e.g. *que*), and groups of two (e.g. *lo que*), three (e.g. *ha de ser*), four (e.g. *qué es lo que*), and five words (e.g. *qué es lo que dices*). In Figure 7, we are using Classic Delta and zero percent culling. Again, the results present the percentage of success: 100 percent means a perfect classification;

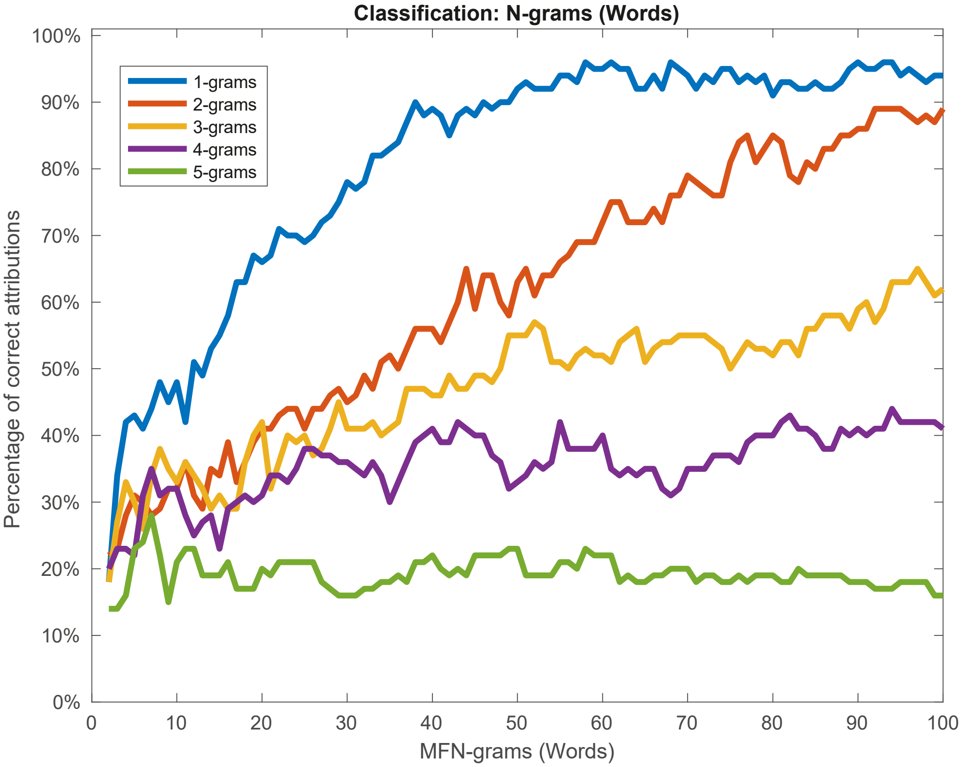


Fig. 7 Percentage of correct attributions when using different N-grams (words). Classic Delta (Cuéllar, CC BY).

77 percent means that this proportion of texts has been correctly classified among the nine groups when performing the leave-one-out cross-validation classification.

The results in Figure 7 are straightforward: they decline when the quantity of words in the n-grams increases. It seems that for authorship verification in corpora similar to ours, using words instead or bigrams, trigrams, etc., may be the best option.

4.3 Results by Varying the Text Length

In the field of Spanish Golden Age theatre, we are not limited to working with entire plays. Collaborative writing was frequent; thus, it was not uncommon for multiple writers to collaborate on the same play, each of them writing one act and then, with minimal corrections, combining them. Then there are short theatrical texts such as

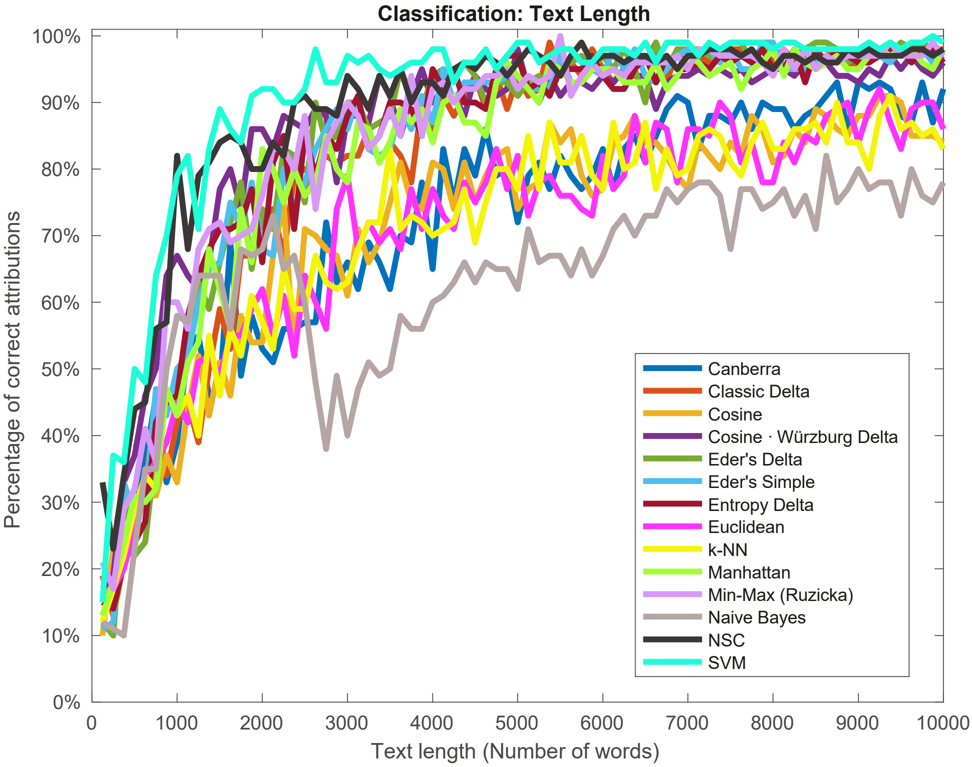


Fig. 8 Percentage of correct attributions with 14 different methods by varying the text length (Cuéllar, CC BY).

entremeses, *loas*, *bailes*, *jácaras*, *mojigangas*, etc. that also present authorship issues. Finally, some scholars are interested in exploring only passages of plays, such as a specific sonnet whose authorship is in doubt. The texts' length is therefore an element that must be taken into consideration.

In Figure 8, we are going to try to answer the length of text—how many words—where stylometry starts to work with acceptable success on my specific corpus.[15] We will compare the 14 methods, randomly sampling our corpus from 50 to 10,000 words, and we will apply 500 MFW and 0 percent culling.

The results show that trying to use stylometry on texts with less than 1,000 words is quite problematic with our methods. Around the 2,500-word mark, most of methods, such as Classic Delta, Cosine, Würzburg Delta, Eder's Delta, Eder's Simple,

[15] For a study on text length and authorial classification, see Eder (2010).

Entropy Delta, Manhattan, Min-Max (Ruzicka), NSC and very specially SVM, start working with quite a high success rate, above 80 percent. The other methods need a higher overall word count to improve their results. From 7,500 words onward, most of the methods start working as they did with the entire plays (which usually have 15,000–20,000 words). It seems clear that the larger the texts are, the better the results that we obtain.[16]

5. Conclusion

In this paper we have tested whether stylometry is an effective tool for authorship identification in Spanish Golden Age theatre with a corpus of one hundred plays of undisputed authorship. We have done so using *stylo* with 14 different methods, and parameters such MFW, culling and n-grams. The global results have been very satisfactory, reaching more than 95 percent of correct attributions. Surprisingly, results do not vary significatively when changing the method, the number of MFW or the percentage of culling, but they tend to be worse when using a number of n-grams larger than single words. We also have tested from which text length stylometry starts working effectively, obtaining results of above 80 percent of correct attributions for texts from 2,500 words with most methods. Thus, applying stylometry to shorter text can be problematic and we have to be extremely careful in those cases.

In conclusion, stylometry appears to be a promising technique for shedding light on the abundant authorship challenges that the Spanish Golden Age theatre presents. Its hypothesis is quite simple: each author uses unconsciously the words in different frequencies and, if we are able to measure these differences, we can establish relations of proximity between the texts, or classify them satisfactorily. The hypothesis seems to be supported by studies like this one, but we need to continue exploring this technique and always use it in combination with other approaches when dealing with a specific attribution.

[16] These results are close to those offered by Hernández Lorenzo (2019) on Early Modern Spanish poetry. She concluded that with texts of 2,000 words or more, stylometry starts working satisfactorily.

Acknowledgements

This article is part of the project *Sound and Meaning in Spanish Golden Age Literature* (FWF Austrian Science Fund 32563). Special thanks to Germán Vega García-Luengos (Universidad de Valladolid) for his help in the elaboration of this paper. The corpus of frequencies that allows the replication of these experiments is placed on https://github.com/AlvaroCuellar or by request to the author.

ORCID®

Álvaro Cuéllar https://orcid.org/0000-0002-9934-6321

References

Blasco, Javier. 2019. "Atribuciones cervantinas desde la estilometría: el entremés de *Los mirones*." In *Cartografía literaria en homenaje al profesor José Romera Castillo*. edited by Guillermo Laín Corona and Rocío Santiago Nogales, 151–68. Madrid: Visor Libros. http://uvadoc.uva.es/handle/10324/37760.

Calvo Tello, José. 2016. "Entendiendo Delta desde las Humanidades." *Caracteres: Estudios culturales y críticos de la esfera digital* 5 (1): 140–76.

Cerezo Soler, Juan, and José Calvo Tello. 2019. "Autoría y estilo. Una atribución cervantina desde las Humanidades Digitales. El caso de *La conquista de Jerusalén*." *Anales Cervantinos* 51: 231–50. https://doi.org/10.3989/anacervantinos.2019.011.

Coenen, Erik. 2009a. "En los entresijos de una lista de comedias de Calderón." *Revista de filología española* 89 (1): 29–56. https://doi.org/10.3989/rfe.2009.v89.i1.63.

Coenen, Erik. 2009b. "Las atribuciones de Vera Tassis." *Castilla. Estudios De Literatura*: 111–33. https://doi.org/10.24197/cel.0.2009.111-133.

Coenen, Erik. 2019. "Everett Hesse, Vera Tassis y el texto de las comedias de Calderón." *Bulletin of the Comediantes* 71 (1/2): 87–102. https://doi.org/10.1353/boc.2019.0006.

Cuéllar, Álvaro. 2018. "La necesidad de la validación cruzada en Stylo y cómo programarla en R." *Caracteres. Estudios culturales y críticos de la esfera digital* 7 (2): 301–20. https://dialnet.unirioja.es/servlet/articulo?codigo=7104985.

Cuéllar, Álvaro, and Germán Vega García-Luengos. 2017–2022. *ETSO: Estilometría aplicada al teatro del Siglo de Oro*. https://etso.es/.

De la Rosa, Javier, and Juan-Luis Suárez. 2016. "The Life of Lazarillo de Tormes and of His Machine Learning Adversities." *Lemir: Revista de Literatura Española Medieval y del Renacimiento* 20: 373–438. http://arxiv.org/abs/1611.05360.

Demattè, Claudia. 2019. "Una nueva comedia en colaboración entre ¿Calderón?, Rojas Zorrilla y Montalbán: *Empezar a ser amigos* a la luz del análisis estilométrico." *Rilce* 35 (3): 852–74. https://doi.org/10.15581/008.35.3.852-74.

Eder, Maciej. 2010. "Does Size Matter? Authorship Attribution, Small Samples, Big Problem." *Digital Scholarship in the Humanities* 30 (2): 167–82. https://doi.org/10.1093/llc/fqt066.

Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. "Stylometry with R: A Package for Computational Text Analysis." *R Journal* 8 (1): 107–21. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

Fradejas Rueda, José Manuel. 2016. "El análisis estilométrico aplicado a la literatura española: las novelas policiacas e históricas." *Caracteres. Estudios culturales y críticos de la esfera digital* 5 (2): 196–45.

Fradejas Rueda, José Manuel. 2019. *Cuentapalabras. Estilometría y Análisis de Textos Con R Para Filólogos*. http://www.aic.uva.es/cuentapalabras.

García-Reidy, Alejandro. 2019. "Deconstructing the Authorship of *Siempre ayuda la verdad*: A Play by Lope de Vega?" *Neophilologus* 103 (4): 493–510. https://doi.org/10.1007/s11061-019-09607-8.

Hernández Lorenzo, Laura. 2019. "Poesía áurea, estilometría y fiabilidad: Métodos supervisados de atribución de autoría atendiendo al tamaño de las muestras." *Caracteres. Estudios culturales y críticos de la esfera digital* 8 (1): 189–28. https://dialnet.unirioja.es/servlet/articulo?codigo=7105381.

Hernández-Lorenzo, Laura, and Joanna Byszuk. 2019. "Challenging Stylometry: The Authorship of the Baroque Play La Segunda Celestina." *Digital Humanities Conference*. https://github.com/JoannaBy/La-Segunda-Celestina.

Madroñal, Abraham. 2019. "Entre la historia y la leyenda. A propósito de *Las dos bandoleras*, comedia atribuida a Lope de Vega." *Anuario Lope de Vega* 25: 281–310. https://doi.org/10.5565/rev/anuariolopedevega.298.

Menéndez Pidal, Ramón. 1949. "Caracteres primordiales de la literatura española. Con referencias a las otras literaturas hispánicas, latina, portuguesa y catalana." In *Historia general de las literaturas hispánicas*, edited by Guillermo Díaz-Plaja, xiv–lix. Barcelona: Editorial Barna.

Moreto, Agustín. 2019. *La adúltera penitente*, edited by Fernando Rodríguez-Gallego. Alicante: Biblioteca Virtual Miguel de Cervantes. http://www.cervantesvirtual.com/nd/ark:/59851/bmc0942634.

Morley, Sylvanus Griswold, and Courtney Bruerton. 1968. *Cronología de las comedias de Lope de Vega*. Barcelona: Gredos.

R Core Team. 2020. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/.

Real Academia Española y Asociación de Academias de la Lengua Española. 2010. *Ortografía de la lengua española*. Madrid: Espasa.

Rißler-Pipka, Nanette. 2016. "Avellaneda y los problemas de la identificación del autor. Propuestas para una investigación con nuevas herramientas digitales." In *El otro Don Quijote. La continuación de Fernández de Avellaneda y sus efectos*, edited by Hanno Ehrlicher, 27–51. Augsburg: Universität Augsburg, Institut für Spanien-, Portugal- und Lateinamerika-Studien (ISLA).

Rodríguez López-Vázquez, Alfredo. 1983. “La autoría de ‘El Burlador de Sevilla’: Andrés de Claramonte.” *Castilla: Estudios de Literatura* 32 (5): 87–108. http://uvadoc.uva.es/handle/10324/16144.

Rodríguez López-Vázquez, Alfredo. 2018. “Cervantes y helecho de procusto: notas críticas al uso de la Estilometría en obras de atribución dudosa y en obras apócrifas.” *EHumanista* 41: 193–201. https://dialnet.unirioja.es/servlet/articulo?codigo=6946620.

Schöberlein, Stefan. 2016. “Poe or not Poe? A Stylometric Analysis of Edgar Allan Poe’s Disputed Writings.” *Digital Scholarship in the Humanities* 32 (3): 643–59. https://doi.org/10.1093/llc/fqw019.

Ulla Lorenzo, Alejandra, Elena Martínez Carro, and José Calvo Tello. 2021. “Las comedias de dudosa atribución de Agustín Moreto: nuevas perspectivas estilométricas.” *Neophilologus* 105: 57–73. https://doi.org/10.1007/s11061-020-09649-3.

Vega García-Luengos, Germán. 2002. “Atribución.” In *Diccionario de la comedia del Siglo de Oro*, edited by Frank P. Casa, Luciano García Lorenzo, and Germán Vega García-Luengos, 16–19. Madrid: Editorial Castalia.

Vega García-Luengos, Germán. 2008. “Consideraciones sobre la configuración del legado de comedias de Calderón.” *Criticón* 103–104: 249–71. https://doi.org/10.4000/criticon.12179.

Vega García-Luengos, Germán. 2009. “Los problemas de autoría en el teatro español de los Siglos de Oro.” In *Unidad y multiplicidad: tramas del hispanismo actual. VIII Congreso Argentino de Hispanistas*, edited by Mariana Genoud de Fourcade and Gladys Granata de Egües, 95–135. Mendoza: Zeta Editores.

Vega García-Luengos, Germán. 2010. “Sobre la identidad de las partes de comedias.” *Criticón* 108: 57–78. https://doi.org/10.4000/criticon.14266.

Vega García-Luengos, Germán. 2021. “Juan Ruiz de Alarcón recupera ‘La monja alférez’.” In *Sor Juana Inés de la Cruz y el teatro novohispano: XLII Jornadas de teatro clásico*, edited by Rafal González Cañal and Almudena García González, 89–159. Castilla-La Mancha: Ediciones de la Universidad de Castilla-La Mancha. https://ruidera.uclm.es/xmlui/handle/10578/28570.

Repetitive Research
Spitzer and Racine

Christof Schöch

Abstract This contribution attempts to retrace Leo Spitzer's (1887–1960) famous stylistic reading of the tragedies of French seventeenth-century author Jean Racine (1639–1699) using digital text collections and computational methods of analysis available today. Spitzer's analysis was first published in 1928 and richly illustrates the manifestations of a "dampening effect" which Spitzer claims is characteristic of Racine's style and at the same time functions as the signature style of the French Classical period more generally. The contribution uses a mixed-methods approach, combining corpus-based modeling and reading of stylistic patterns with statistical analyses of their distribution. The present attempt to retrace Spitzer's study not only reveals new insights into Racine's and the Classical period's style, but also serves to highlight the respective strengths and limitations of established (non-digital and/or hermeneutical) and computational (digital, algorithmic and/or quantitative) approaches to stylistic analysis and the contrasting notions of style which underpin them.

Keywords Racine, Spitzer, French literature, style, replication

> "Should we strive to obtain the same results or approach
> the process of replication as a way of knowing?"
> (Rockwell 2016)

1. Introduction

This contribution describes an attempt to retrace Leo Spitzer's (1887–1960) famous stylistic reading of the tragedies of French seventeenth-century author Jean Racine (1639–1699) using digital text collections and computational methods of analysis available today. Spitzer's analysis, titled "Die klassische Dämpfung bei Racine"

Schöch, Christof: Repetitive Research. Spitzer and Racine, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 119–148.
DOI: https://doi.org/10.17885/heiup.1157.c19369

(literally, in English: "The Classical Dampening in Racine"), was first published in 1928 and richly illustrates the manifestations of a "dampening effect" which Spitzer claims is characteristic of Racine's style and at the same time functions as the signature style of the Classical period more generally.

The approach followed here is to take a new look into the "dampening effect" described by Spitzer, that is, the ten abstract phenomena or stylistic effects into which Spitzer divided it and the roughly fifty stylistic patterns or devices he identified in relation to it. 24 of those 50 stylistic patterns could be modeled using search queries applied to richly annotated, digital versions of Racine's plays. The remaining patterns proved to be too complex and/or too much dependent on context to be modeled at this point. Despite this limitation, this approach permits us not only to reproduce and verify some of Spitzer's findings, but also to extend his investigation by retrieving additional instances of the relevant stylistic patterns and comparing them to the ones he chose to mention. In addition, the scope of the investigation is enlarged to compare the instances found in Racine's works with those found in a collection of tragedies contemporary to Racine. Finally, this approach is extended with quantitative methods, in which the quantitative prevalence of the stylistic patterns in Racine's work and in his contemporaries' work is compared.

The mixed-methods approach pursued here, combining a close reading of instances with a statistical analysis of their distribution, helps decide whether the dampening effect is characteristic of Racine—and is therefore best described as an authorial style—or, whether it is characteristic, rather, of the French Classical period more generally—and is therefore best described as a period style. Spitzer himself was ambivalent about this, as he calls the key stylistic principle *klassische Dämpfung*, with reference to a period style, but analyses it by referring exclusively to Racine, essentially conducting a one-author analysis.

The present attempt to retrace Spitzer's study not only reveals new insights into Racine's and the French Classical period's style, but also serves to highlight the respective strengths and limitations of established (non-digital and/or hermeneutical) and computational (algorithmic and/or quantitative) approaches to stylistic analysis and the contrasting notions of style which underpin them. Ultimately, by closely reenacting a previous study, the research presented here also highlights the continuities and differences between established and computational approaches to literature.

2. Context: Repetitive Research

Because it closely reenacts Spitzer's analysis using digital data and methods, this study falls into the paradigm of repetitive research (as described in Schöch 2023b; see also Schöch et al. 2020). This study repeats earlier work, "in the sense that [it] actively [seeks] to align [its] research questions or hypotheses, [its] datasets and/or [its] methods of analysis, with research practiced and published earlier. This is done with the explicit aim to approximate an earlier study, but conscious also of the fact that perfectly identical repetition is virtually impossible to achieve." (Schöch 2023b, 374). In the humanities, such repetitive research can often imply that the earlier research one is trying to repeat has been practiced within a non-digital paradigm (i.e., relying on printed sources) and often also in a non-computational paradigm (i.e. relying on digital data, but applying qualitative methods of analysis). Additionally, this study is repeatable, "in the sense that it (typically) makes all the efforts it can to provide the data, code, and explanatory information that make it possible for others, at a later point in time, to perform the same (or very similar) research again." (Schöch 2023b, 374). In this way, this kind of research is located between past and future: a (never identical) reenactment of past research, and an invitation for (never identical) further reenactments in the future. This is done with the conviction, or at least in the hope, that this cycle of repetitions is not a sterile treading in the same place, but a productive, insightful upwards spiral.

Elsewhere, I have situated this kind of repetitive research in the larger context of the reproducibility crisis that has affected not only fields like medicine and psychology, but is increasingly a concern also in artificial intelligence and the digital humanities, in particular in computational literary studies (for an introduction, see Fidler and Wilcox 2021). Also, I have attempted to structure the field of repetitive research using a simple typology that explains the various forms this kind of research can take (Schöch 2023b). In short, the typology describes the relationship between an earlier study and its repetition in terms of three key variables: the research question, the method of analysis (including the implementation of that method), and the dataset used. For each of these variables, a repetitive study can attempt to operate either in the same, a similar, a different, or an unrelated manner as the previous study. The typology is not meant to establish these distinctions in a purely categorical fashion: rather, as research questions, data or methods are never entirely identical or completely different from the earlier study, the extreme points in the typology are meant to open up a multidimensional gradient of practices. Such a typology can have a number of uses: Conceptually, it helps structure the field, establish conceptual distinctions and provide terminological clarity. Pragmatically, it offers guidance on what data, code and documentation need to be included with a given publication if one or the other mode of replication should be supported in the future.

The present study bridges both a considerable temporal gap—almost a century has passed since Spitzer first published his analysis—and a substantial methodological and technological difference: Spitzer relied on a close reading of a print edition of Racine's works, whereas the present study employs digital tools to analyze a corpus of Racine's works (in the first part of the study) as well as a larger corpus of Classical French tragedies (in the second part of the study), combining qualitative and quantitative analyses.

The first and most extensive part of the present study focuses on a computational repetition of Spitzer's previous, non-digital study. With respect to the typology of repetitive research, this part can be described as follows: as pursuing a virtually identical research question or hypothesis; as relying on very similar, though digital, data; and as using a very different, algorithmic, method of investigation. Where Spitzer uses a holistic approach, fundamentally a stylistic analysis guided by the guiding principle of the dampening effect, the present study's method is based on defining formal patterns to be identified in annotated text to identify instances of relevant stylistic patterns. In the terms proposed by the typology mentioned above, this part of the present study would best be called a *re-analysis (of data)*.

The second part of the study enlarges the dataset considerably, so that it can no longer be deemed to be similar to Spitzer's original data. In addition (and as a consequence), the research question also shifts away from Spitzer's exclusive and programmatic concern with Racine to a comparison of Racine with his contemporaries. Finally, the statistical methods of comparing the prevalence of items in two subcorpora are alien to Spitzer's study as well, so that we end up with a study that uses a different (though not unrelated) dataset, a very different method and a different research question, leading to research that, in the terms of the typology, would be called *follow-up research*.

In addition, as it is probably true of any example of repetitive research, this study of course adds the meta-level of comparing the methods and results of the previous study to the current study. I consider this to be at least as important as the findings with regards to Racine and his contemporaries. Any repetitive research helps recognize the presuppositions, strengths and weaknesses of both the original and the current study: the usefulness of their underlying datasets, the appropriateness and inherent biases of their methods and procedures; and the strength and interest of their findings with respect to the research questions at hand. I hope, therefore, that a study such as the present one is not only an occasion to think again about Spitzer, Racine, style and the Classical Age, but also a good moment for reflection about the relationship between established (hermeneutical and/or qualitative) stylistics on the one hand, and computational (algorithmic, quantitative) literary stylistics on the other hand.

3. Background: Spitzer and Racine

The period stretching roughly from 1630 to 1715 and known in French as *l'âge classique* remains one of the most prestigious periods in French cultural and literary history. Its most lasting literary legacy lies without a doubt in the *théâtre* *classique*.[1] Together with Pierre Corneille and Molière, Jean Racine is among the most famous and consecrated authors of French drama of this period.[2] Between 1664 and 1691, Racine wrote 12 plays, among which were nine tragedies, two religious plays and one comedy. This makes him noticeably less prolific than some of this contemporaries, like Pierre Corneille and Molière, who each wrote more than 30 plays. However, Racine's plays have become to be perceived as the epitome of Classical theater, with their relatively strict adherence to the *règles classiques*, i.e. the rule-based poetics of French Classicism, with their sharp focus on the tragic conflict and their measured expression of intense emotion.

To a significant degree, this perception has been created and sustained by Leo Spitzer and the many generations of his readers, despite the fact that Spitzer appears to focus primarily on literary style. It is true that the dampening effect that Spitzer postulates is at the center of Racine's work is first and foremost a principle of literary style. However, it becomes, in the eyes of Leo Spitzer, the manifestation of the aesthetic center of Racine's work. Jaubert aptly describes Spitzer's approach as follows: "dégager la forme-sens d'une écriture, le principe signifiant non seulement niché au creux du détail, mais structurant toute l'œuvre, assurant de la microlecture à la lecture globale la transitivité des niveaux" (1970, 26; "to disentangle the meaningful form of a writing style, the meaningful principle that is not only inscribed within any details, but which also structures the entire work, in this way guaranteeing, from a reading at the micro-level to a global reading, the transitivity of the levels").[3] Indeed, for Spitzer, a thorough analysis of literary style does more than discover the sum of the stylistic parts: it provides access to the underlying literary coherence and unifying principle of a great author's work. At the same time, this unifying principle guides and directs any stylistic reading of a particular passage.[4]

As noted above, and starting with the title in German—"Die klassische Dämpfung bei Racine"—, the dampening effect that Spitzer sees at work in Racine's work is ambivalent with regard to the question of whether it is a period style, a genre style, or an authorial style. However, in his analysis, Spitzer programmatically puts Racine's

[1] Several studies on the *théâtre classique* have become classics in their own right: on dramatic theory (Bray 1926) and practice (Scherer 1950). A more rounded and more recent overview is Génetiot (2005).

[2] The bibliography on Jean Racine is obviously vast. One may wish to start with a witty biography (Viala 1990) or a survey of critical work (Rohou 2005). Statistically-minded readers may enjoy consulting the very early study in lexical statistics by Bernet (1983).

[3] Note that this and all following translations are provided by the author and are purposefully literal rather than elegant.

[4] Spitzer's essay first appeared in German in 1928 (Spitzer 1931). This text is cited as *Dämpfung*, in the remainder of this study. It took four decades before it was translated into French in the volume *Études de style* (Spitzer 1970). Since then, however, the *Études de style* have become a reference for many students and researchers alike.

work front and center to illustrate this stylistic phenomenon. For example, after explaining the first stylistic pattern, he states: “Ich brauche nun nur noch dies Stilmittel als für R[acine] charakteristisch an verschiedenen Stellen der Andr[omaque] und der späteren Dramen zu belegen” (*Dämpfung*, 138; “I now only need to show, in different places of *Andromaque* and the later plays, that this stylistic device is characteristic for Racine”). Elsewhere, he describes one of the stylistic patterns as “eine ganz typische Stilfigur bei Racine” (*Dämpfung*, 232; “a stylistic device truly typical of Racine”). At the same time, Spitzer never explicitly claims that it could not also be found in other authors, whether contemporaries or predecessors.

But what exactly does he mean by the dampening effect? Spitzer explains himself that he believes it is best understood as a musical metaphor, something akin to the dampening pedal on the piano: that is, as a device that allows for a muted sound, but which really only becomes effective in its alternation and contrast with passages that are unmuted. Therefore, Spitzer describes the dampening effect in Racine as follows:

> […] das oft Nüchtern-Gedämpfte, Verstandesmäßig-Kühle, fast Formelhafte an diesem Stil, das dann oft plötzlich und unvermutet für Augenblicke in poetisches Singen und erlebte Form übergeht, worauf aber wieder rasch ein Löschhütchen von Verstandeskühle das sich schüchtern hervorwagende lyrische Sich-Ausschwelgen des Lesers niederdämpft. Racine, das eigentlich Racinesche, ist eben weder bloß Formel noch bloß lyrisches Singen, sondern die Abfolge und das Ineinandergreifen beider Elemente (*Dämpfung*, 134).
>
> “[…] the often soberly subdued, rationally cool, almost formulaic quality of this style, which then often suddenly and unexpectedly changes only for moments into poetic singing and experienced form, whereupon, however, a certain cap of rational coolness quickly subdues the reader’s shyly venturing lyrical self-expression. Racine, what is properly Racine, is neither mere formula nor mere lyrical singing, but the sequence and the interlocking of both elements”.

Again, Spitzer describes this pattern of intermittent dampening effects as the essence of Racine (“das eigentlich Racinesche”). Indeed, the dampened passages only become apparent when they contrast or alternate with other, more lively, more expressive, less formally-regulated passages. Despite this fact, Spitzer almost exclusively focuses on those general stylistic principles and those stylistic patterns that produce, in his view, a dampening effect in Racine’s plays.

How, then, does Spitzer proceed? Right from the beginning, Spitzer explicitly places the focus on Racine and Racine alone:

> Ich möchte nun im Folgenden die Dämpfungen im Stil Racines (nicht bloß in Wortstellung, Rhythmus und Reim) verfolgen, nicht bezogen auf seine Vorgänger, wie Voßler anregt, wodurch Racine zu sehr als Satellit anderer Sterne erschiene, als vielmehr als Selbst-Stern, als Sternenkosmos, den ich, wie gewöhnlich die Gegenstände meiner Stilforschung, in sich ruhend sehe (*Dämpfung*, 136).
>
> "In the following, I would like to trace the attenuations in Racine's style (not merely in word order, rhythm, and rhyme), not in relation to his predecessors, as Voßler suggests, which would make Racine appear too much as a satellite of other stars, but rather as a self-star, as a stellar cosmos, which I see, as usually the objects of my stylistic research, resting in itself".

Spitzer is very systematic in his approach: he takes passages from all of Racine's plays, except the one comedy, into account in his analysis. The examples he cites are broadly, though not evenly distributed, quantitatively, among the plays.[5] He also breaks down the overall stylistic dampening effect into ten stylistic principles and about 50 specific stylistic patterns or devices, although his grouping is not always very explicit. Each of the principles is named, and each of the patterns is described and discussed briefly before being illustrated with excerpts from Racine's tragedies, for a total of 484 examples. Effectively, one could say that his argumentative approach starts from an abstract idea (the dampening effect) and then breaks it down into ever smaller and more concrete stylistic levels: broader stylistic phenomena defined primarily by their effect and contribution to the overall dampening effect are broken down into particular stylistic patterns, defined on the level of semantics, syntax, rhetorical devices or metric structure and illustrated using individual examples from the plays. Conversely, Spitzer re-connects each individual example, each stylistic pattern and each broader phenomenon back to the overarching idea of the dampening effect, ensuring the unity not only of his argument, but also of Racine's work. This systematic approach is by no means quantitative and can only be called formalistic to a limited extent, but it does create an excellent starting point for bridging the gap between stylistic close reading, as practiced by Spitzer, and algorithmic reading, as proposed in the present study.

As mentioned before, in his extensive investigation of the stylistic devices employed by Racine to produce the dampening effect, Spitzer groups them into ten stylistic phenomena, or groups of stylistic patterns and devices, based on a shared principle or a comparable effect. The following is a list established for the purposes of the present study:[6]

[5] *Andromaque* (158 examples) and *Phèdre* (140) are most widely cited by Spitzer, followed by *Bajazet* (83) and *Athalie* (50).

[6] The alphabetic identifiers for each group have been added for convenience; page references in parentheses are to Spitzer's essay in the German edition from 1931 and mark the beginning of

- A "Die Entindividualisierung" (*Dämpfung*, 136; "the de-individualization")
- B "[Die] Dämpfende, das Unmittelbare des Empfindens abschwächend[e Wirkung]" (*Dämpfung*, 144; "the attenuating effect that weakens the immediacy of the sensation")
- C "[Das] Kühl-Abgeschwächte" (*Dämpfung*, 150; "coolly-weakened")
- D "[Dass] das Ich sich nicht zu sehr aussinge" (*Dämpfung*, 151)[7]
- E "[Die] Unpersönlich[keit] der Rede" (*Dämpfung*, 157; "the unpersonal character of the speech")[8]
- F Die "Konturverwischung" (*Dämpfung*, 163; "the blurring of the contours")
- G Die "Abkühlung der lyrischen Temperatur" (*Dämpfung*, 178; "the cooling of the lyrical temperature")[9]
- H Retardierende Elemente (*Dämpfung*, 214; "retarding elements")
- J "Gemalte Aufregung" (*Dämpfung*, 216; "painted excitement")
- K "Formeln" (*Dämpfung*, 225, term: 244; "formulae")[10]

These categories are introduced by Spitzer not as theoretically-justified or explicitly-defined categories, but in a rather casual manner. On the one hand, they serve to implicitly structure and group the analyses of the individual stylistic devices; on the other, they of course connect each device to its principle or effect and, in this manner, mark its contribution to the overall dampening effect.

4. Data and Tools

Before proceeding to a closer look at some of the stylistic patterns themselves, a description of the dataset used in the present study is in order. The dataset used is not an exact replica of the texts Spitzer used; in fact, Spitzer does not indicate the edition he used. However, the dataset does contain the same 11 plays (nine tragedies and

the discussion of each group of stylistic devices. Page references to the French edition (1970) are included in the file "overview.md" provided in the companion repository on Github at https://github.com/dh-trier/spitzer-racine (DOI: https://doi.org/10.5281/zenodo.3959974).

[7] The phrasing in the French edition is a bit clearer: "Refrèner le chant lyrique du Moi" ("To restrain the lyrical chant of the subject").

[8] Also described as: "[Das] Abrücken vom Persönlichen zum Prinzipiellen hin" (*Dämpfung,* 160; "The distancing from the personal towards the principle").

[9] Also described as: der "geformte und besänftigte Eindruck der Reden Racine'scher Figuren" or the "abdämpfende und vorbereitende Wirkung" (*Dämpfung,* 210; "the shaped and soothed impression of Racine's characters' speeches").

[10] Also described as: "die Rede an die Innerlichkeit anzuschmiegen" (*Dämpfung,* 228; "To mold the speech against the interiority").

two religious plays, but not Racine’s only comedy) that Spitzer used in his study and which were first performed between 1664 and 1691: *Alexandre, Andromaque, Athalie, Bajazet, Bérénice, Britannicus, Esther, Iphigénie, Mithridate, Phèdre* and *La Thébaïde*. An additional set of tragedies from Racine’s contemporaries was also included, for the second part of the study. It includes 38 tragedies by Claude Boyer, Jean Campistron, Pierre Corneille, Thomas Corneille, Philippe Quinault and Nicolas Pradon, all first performed between 1660 and 1695 and all written in verse, just as Racine’s plays.

Digital versions of the texts were used that are available from the excellent *Théâtre classique* platform (Fièvre 2007–2022) in an XML-TEI format (P4) with tacitly modernized spelling. Building on this basis, all texts were annotated using Freeling, NLTK and WordNet in Python, to produce a format that has a token-based annotation that covers both morphological information and semantic information.[11] All annotations have been represented in an XML-TEI format compatible with the TXM corpus analysis tool, with each token represented in a “w” (word) element and each token-level annotation represented as an attribute-value pair on the respective “w” element. See the following somewhat verbose code listing, taken from Racine’s *Bérénice*, for an illustration:

```
<s n="13">
<w n="t13.1" form="Quel" lemma="quel" tag="DT0MS0" pos="det"
type="xxx" gen="masculine" num="singular" wnsyn="xxx"
wnlex="xxx">Quel </w>
<w n="t13.2" form="fruit" lemma="fruit" tag="NCMS000" pos="noun"
type="common" gen="masculine" num="singular" wnsyn="13134947-n"
wnlex="noun.plant">fruit </w>
<w n="t13.3" form="me" lemma="me" tag="PP1CS00" pos="pron"
type="xxx" gen="xxx" num="singular" wnsyn="xxx" wnlex="xxx">me </w>
<w n="t13.4" form="reviendra" lemma="revenir" tag="VMIF3S0"
pos="verb" type="main" gen="xxx" num="singular" wnsyn="02004874-v"
wnlex="verb.motion">reviendra </w>
<w n="t13.5" form="d’" lemma="de" tag="SP" pos="prep" type="xxx"
gen="xxx" num="xxx" wnsyn="xxx" wnlex="xxx">d’ </w>
<w n="t13.6" form="un" lemma="un" tag="DI0MS0" pos="det"
type="xxx" gen="masculine" num="singular" wnsyn="xxx"
wnlex="xxx">un </w>
<w n="t13.7" form="aveu" lemma="aveu" tag="NCMS000" pos="noun"
type="common" gen="masculine" num="singular" wnsyn="06732350-n"
wnlex="noun.communication">aveu </w>
```

[11] All data (as well as the code used) is available from the companion repository mentioned above.

```
<w n="t13.8" form="téméraire" lemma="téméraire" tag="AQ0CS00"
pos="adj" type="qualif" gen="common" num="singular" wnsyn="xxx"
wnlex="xxx">téméraire </w>
<w n="t13.9" form="?" lemma="?" tag="Fit" pos="punc" type="xxx"
gen="xxx" num="xxx" wnsyn="xxx" wnlex="xxx">? </w>
</s>
```

The analyses were performed using the TXM corpus analysis tool (version 0.8.1 released in June 2020; see Heiden 2010) for the query step, while custom Python scripts were used for the comparative analysis step.

5. Formally Modeling Spitzer's Stylistic Patterns

The key methodological challenge resides in the formal modeling of the complex stylistic patterns described by Spitzer. He, for the most part, provides a brief description of the stylistic phenomenon in question and then provides further refinement, illustration and interpretation using examples from Racine's plays. Although Spitzer is quite precise in his descriptions, he does not employ formal definitions of the stylistic patterns in any systematic way, not does he limit himself to surface phenomena. Indeed, the stylistic patterns he describes often depend on semantics and metrical structure, both locally and with respect to their wider context in a given play.

As a consequence of the corpus-based, quantifying approach pursued in this replication study, Spitzer's descriptions need to be re-implemented or operationalized in a formal, machine-actionable way. The method of choice for this has been to employ the query language Corpus Query Processor (CQP; see Evert and Hardie 2011) as used by TXM, in order to be able to apply the corresponding queries on the annotated data available in the TXM corpus format. This query language can be described as relying on regular expressions operating not only on the word forms, but on all available annotations, as well as capable of taking structural cues into account (although this was not used here). The basic unit of the query is the token, and each token to be queried is described in terms of a more or less strict filter on one or several of the levels of annotation. Sequences and patterns of tokens can be defined as well. As Spitzer defines both formal and semantic constraints, the semantic annotation level was of particular importance here. Once a given stylistic pattern is modeled in this way as a CQP query matching the annotations provided in the dataset, all matching stylistic patterns can be retrieved, checked, investigated, and counted.

In order to assess the quality of these approximations of Spitzer's patterns, two parameters have been assessed: Firstly, I have checked whether all of Spitzer's examples are included in the results obtained with the query. In information retrieval terms, this very roughly corresponds to checking the recall of the query. Secondly, I have checked whether all the results identified by the query correspond to Spitzer's definition. This is very roughly equivalent to checking the precision of the query. As Spitzer does not provide an exhaustive list of relevant examples, but only a certain number of illustrative examples, it is not a sign of a bad query if more relevant examples are found relative to the number of examples given by Spitzer. As a consequence, however, no numerical accuracy score has been calculated, but rather a qualitative assessment has been derived from these two checks.

5.1 Example: Erasing of Contours

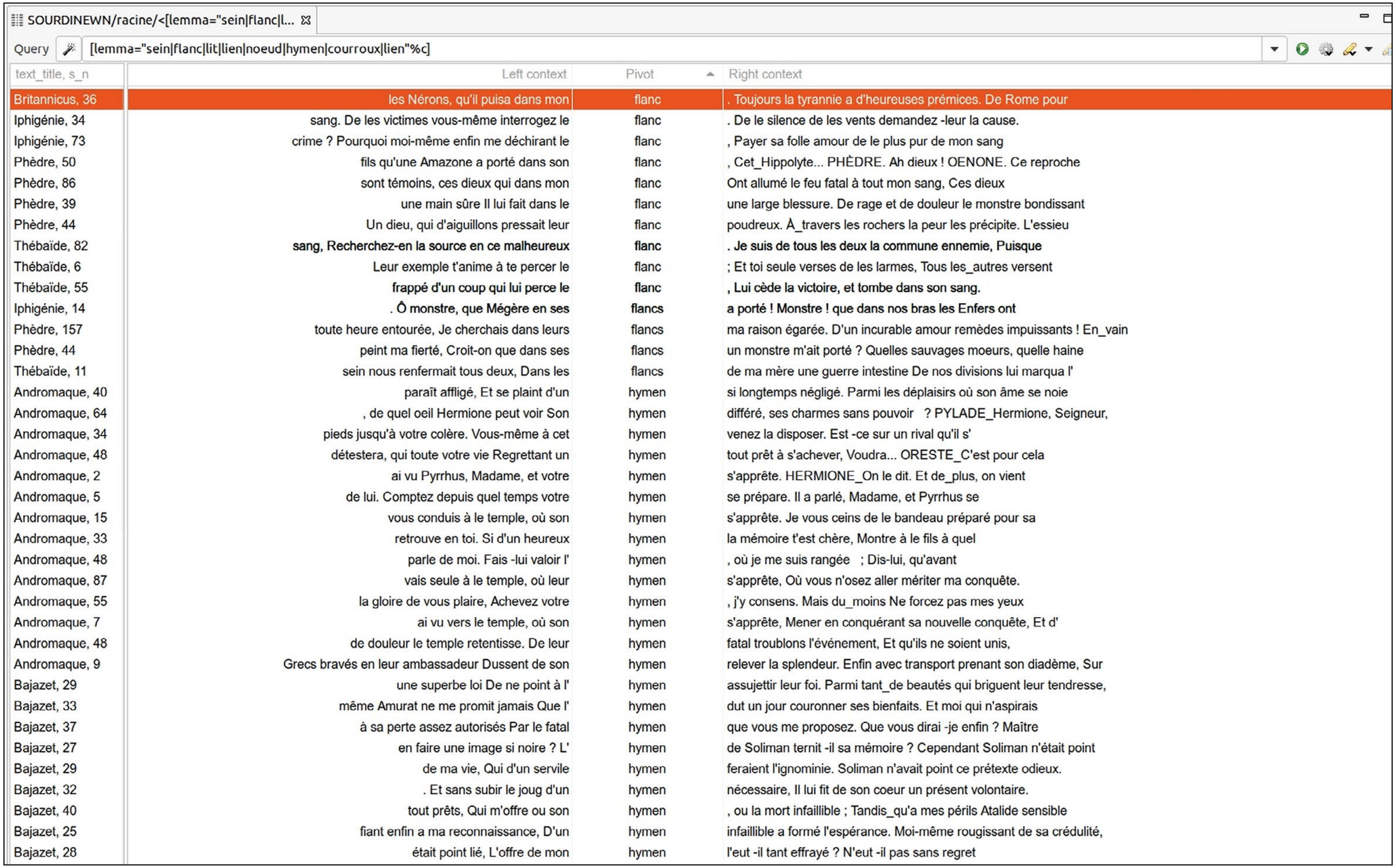
A first, and very simple example comes from the group of phenomena Spitzer calls "Kontourverwischung" (*Dämpfung,* 163 / group F; "erasing of contours"). The stylistic pattern in question is called "konturverwischende Vokabeln" (*Dämpfung,* 165; pattern F2; "contour-erasing lexical items") and can be understood as a list of lexical items that are used by Racine in place of a more direct item, as in the following two examples:

(1) Malgré tout son orgueil, ce monarque si fier
À son trône, à son lit daigna l'associer. (*Bajazet* II, 1)
'In spite of all his pride, this monarch so proud /
To his throne, to his bed deigned to associate her.'

(2) Les dieux m'en sont témoins, ces dieux qui dans mon flanc
Ont allumé le feu fatal à tout mon sang. (*Andromaque* II, 5)
'The gods are my witnesses, these gods who in my side /
Have lit the fatal fire to all my blood.'

Here, *lit* is the term chosen instead of *mariage* ('bed' instead of 'marriage') and *flanc* instead of *ventre* ('flank'/'side' instead of 'stomach'), although clearly in a metaphorical way. Other examples Spitzer cites are *sein* (again for *ventre*), *hymen* (for *mariage*) and *courroux* (for *colère*). In addition, *lien* or *nœud* can be used in the same way (for either marriage or family and/or love relationships). Spitzer notes that these are not just more noble terms, with respect to the *bienséances* that disdain excessively corporeal expressions, but also much vaguer terms. This case is interesting as well because it is an instance of an explicit stylistics of deviance from a norm, where a near-synonym is preferred, for stylistic reasons, over the more usual term.

This pattern, while simple in appearance, is not trivial to model, because some of the terms mentioned by Spitzer can not only be used in place of a more direct, usual



SOURDINEWN/racine/<[lemma="sein|flanc|l...

Query [lemma="sein|flanc|lit|lien|noeud|hymen|courroux|lien"%c]

text_title, s_n	Left context	Pivot	Right context
Britannicus, 36	les Nérons, qu'il puisa dans mon	flanc	. Toujours la tyrannie a d'heureuses prémices. De Rome pour
Iphigénie, 34	sang. De les victimes vous-même interrogez le	flanc	. De le silence de les vents demandez -leur la cause.
Iphigénie, 73	crime ? Pourquoi moi-même enfin me déchirant le	flanc	, Payer sa folle amour de le plus pur de mon sang
Phèdre, 50	fils qu'une Amazone a porté dans son	flanc	, Cet_Hippolyte... PHÈDRE. Ah dieux ! OENONE. Ce reproche
Phèdre, 86	sont témoins, ces dieux qui dans mon	flanc	Ont allumé le feu fatal à tout mon sang, Ces dieux
Phèdre, 39	une main sûre Il lui fait dans le	flanc	une large blessure. De rage et de douleur le monstre bondissant
Phèdre, 44	Un dieu, qui d'aiguillons pressait leur	flanc	poudreux. À_travers les rochers la peur les précipite. L'essieu
Thébaïde, 82	**sang, Recherchez-en la source en ce malheureux**	**flanc**	**. Je suis de tous les deux la commune ennemie, Puisque**
Thébaïde, 6	Leur exemple t'anime à te percer le	flanc	; Et toi seule verses de les larmes, Tous les_autres versent
Thébaïde, 55	**frappé d'un coup qui lui perce le**	**flanc**	**, Lui cède la victoire, et tombe dans son sang.**
Iphigénie, 14	**. Ô monstre, que Mégère en ses**	**flancs**	**a porté ! Monstre ! que dans nos bras les Enfers ont**
Phèdre, 157	toute heure entourée, Je cherchais dans leurs	flancs	ma raison égarée. D'un incurable amour remèdes impuissants ! En_vain
Phèdre, 44	peint ma fierté, Croit-on que dans ses	flancs	un monstre m'ait porté ? Quelles sauvages moeurs, quelle haine
Thébaïde, 11	sein nous renfermait tous deux, Dans les	flancs	de ma mère une guerre intestine De nos divisions lui marqua l'
Andromaque, 40	paraît affligé, Et se plaint d'un	hymen	si longtemps négligé. Parmi les déplaisirs où son âme se noie
Andromaque, 64	, de quel oeil Hermione peut voir Son	hymen	différé, ses charmes sans pouvoir ? PYLADE_Hermione, Seigneur,
Andromaque, 34	pieds jusqu'à votre colère. Vous-même à cet	hymen	venez la disposer. Est -ce sur un rival qu'il s'
Andromaque, 48	détestera, qui toute votre vie Regrettant un	hymen	tout prêt à s'achever, Voudra... ORESTE_C'est pour cela
Andromaque, 2	ai vu Pyrrhus, Madame, et votre	hymen	s'apprête. HERMIONE_On le dit. Et de_plus, on vient
Andromaque, 5	de lui. Comptez depuis quel temps votre	hymen	se prépare. Il a parlé, Madame, et Pyrrhus se
Andromaque, 15	vous conduis à le temple, où son	hymen	s'apprête. Je vous ceins de le bandeau préparé pour sa
Andromaque, 33	retrouve en toi. Si d'un heureux	hymen	la mémoire t'est chère, Montre à le fils à quel
Andromaque, 48	parle de moi. Fais -lui valoir l'	hymen	, où je me suis rangée ; Dis-lui, qu'avant
Andromaque, 87	vais seule à le temple, où leur	hymen	s'apprête, Où vous n'osez aller mériter ma conquête.
Andromaque, 55	la gloire de vous plaire, Achevez votre	hymen	, j'y consens. Mais du_moins Ne forcez pas mes yeux
Andromaque, 7	ai vu vers le temple, où son	hymen	s'apprête, Mener en conquérant sa nouvelle conquête, Et d'
Andromaque, 48	de douleur le temple retentisse. De leur	hymen	fatal troublons l'événement, Et qu'ils ne soient unis,
Andromaque, 9	Grecs bravés en leur ambassadeur Dussent de son	hymen	relever la splendeur. Enfin avec transport prenant son diadème, Sur
Bajazet, 29	une superbe loi De ne point à l'	hymen	assujettir leur foi. Parmi tant_de beautés qui briguent leur tendresse,
Bajazet, 33	même Amurat ne me promit jamais Que l'	hymen	dut un jour couronner ses bienfaits. Et moi qui n'aspirais
Bajazet, 37	à sa perte assez autorisés Par le fatal	hymen	que vous me proposez. Que vous dirai -je enfin ? Maître
Bajazet, 27	en faire une image si noire ? L'	hymen	de Soliman ternit -il sa mémoire ? Cependant Soliman n'était point
Bajazet, 29	de ma vie, Qui d'un servile	hymen	feraient l'ignominie. Soliman n'avait point ce prétexte odieux.
Bajazet, 32	. Et sans subir le joug d'un	hymen	nécessaire, Il lui fit de son coeur un présent volontaire.
Bajazet, 40	tout prêts, Qui m'offre ou son	hymen	, ou la mort infaillible ; Tandis_qu'a mes périls Atalide sensible
Bajazet, 25	fiant enfin a ma reconnaissance, D'un	hymen	infaillible a formé l'espérance. Moi-même rougissant de sa crédulité,
Bajazet, 28	était point lié, L'offre de mon	hymen	l'eut -il tant effrayé ? N'eut -il pas sans regret

Fig. 1 Results for query F2 in Racine's plays, using TXM (Schöch, CC BY).

term; in particular, *lit* and *lien* are also used with the literal meaning by Racine and the interpretation of the words strongly depend on their particular context. The following very simple CQP query can be used to retrieve these instances:

```
[lemma="sein|flanc|lit|lien|noeud|hymen|courroux|lien"%c]
```

This query relies only on the isolated lemmata for a list of words. First of all, we can see that these lexical items are all but rare in Racine's works, with a total of 318 instances (Figure 1).[12] The most frequent term is *courroux* (119 instances), followed by *hymen* (87 instances) and *sein* (57 instances).

All of Spitzer's examples are included in the results, but there is quite a number of cases where it is debatable whether Spitzer would have included the instances among his examples: Indeed, it appears from a close reading of the various instances of *flanc*, for example, that this appears to be used in two ways: in a metaphorical way (as in the above example), but also in a literal, physical sense (as in the example below):

(3) Le roi frappé d'un coup qui lui perce le flanc,
Lui cède la victoire, et tombe dans son sang. (*Thébaïde,* V, 3)
'The king struck by a blow which pierces his side, /
Gives up the victory, and falls in his blood.'

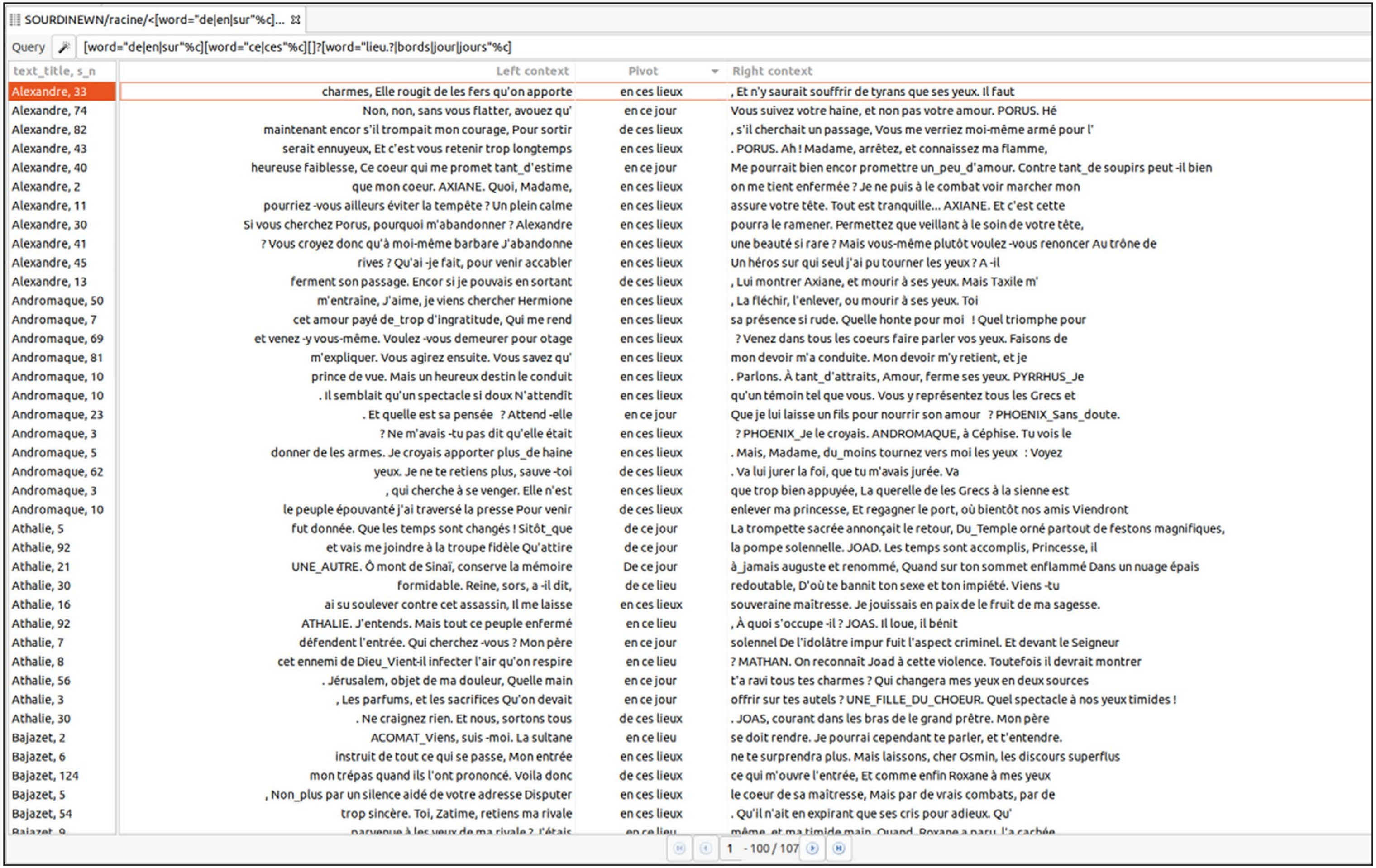
5.2 Example: Spatial and Temporal Paraphrases

Another example concerns the group of stylistic patterns Spitzer describes as creating a weakened immediacy of sensibility (*Dämpfung,* 216 / group B). The stylistic pattern in focus here consists of spatial and temporal paraphrases (pattern B2), notably using the paraphrases *en ces lieux*, *en ce lieu* or *sur ces bords* instead of the direct locative adverb *ici* as well as the paraphrase *en ce jour* instead of the direct temporal adverb *aujourd'hui* (*Dämpfung*, 147).

(4) Vous savez qu'en ces lieux mon devoir m'a conduite. (*Andromaque*, III, 4)
'You know that in these places my duty led me.'

(5) Depuis que sur ces bords les Dieux ont envoyé
La fille de Minos et de Pasiphaé. (*Phèdre*, I, 1)

[12] For a closer look, see the CSV file corresponding to the F2 (= group F, pattern 2) query's results in the folder "analysis" of the project repository.



SOURDINEWN/racine/<[word="de|en|sur"%c]... ☒

Query [word="de|en|sur"%c][word="ce|ces"%c][]?[word="lieu.?|bords|jour|jours"%c]

text_title, s_n	Left context	Pivot	Right context
Alexandre, 33	charmes, Elle rougit de les fers qu'on apporte	en ces lieux	, Et n'y saurait souffrir de tyrans que ses yeux. Il faut
Alexandre, 74	Non, non, sans vous flatter, avouez qu'	en ce jour	Vous suivez votre haine, et non pas votre amour. PORUS. Hé
Alexandre, 82	maintenant encor s'il trompait mon courage, Pour sortir	de ces lieux	, s'il cherchait un passage, Vous me verriez moi-même armé pour l'
Alexandre, 43	serait ennuyeux, Et c'est vous retenir trop longtemps	en ces lieux	. PORUS. Ah ! Madame, arrêtez, et connaissez ma flamme,
Alexandre, 40	heureuse faiblesse, Ce coeur qui me promet tant_d'estime	en ce jour	Me pourrait bien encor promettre un_peu_d'amour. Contre tant_de soupirs peut -il bien
Alexandre, 2	que mon coeur. AXIANE. Quoi, Madame,	en ces lieux	on me tient enfermée ? Je ne puis à le combat voir marcher mon
Alexandre, 11	pourriez -vous ailleurs éviter la tempête ? Un plein calme	en ces lieux	assure votre tête. Tout est tranquille... AXIANE. Et c'est cette
Alexandre, 30	Si vous cherchez Porus, pourquoi m'abandonner ? Alexandre	en ces lieux	pourra le ramener. Permettez que veillant à le soin de votre tête,
Alexandre, 41	? Vous croyez donc qu'à moi-même barbare J'abandonne	en ces lieux	une beauté si rare ? Mais vous-même plutôt voulez -vous renoncer Au trône de
Alexandre, 45	rives ? Qu'ai -je fait, pour venir accabler	en ces lieux	Un héros sur qui seul j'ai pu tourner les yeux ? A -il
Alexandre, 13	ferment son passage. Encor si je pouvais en sortant	de ces lieux	, Lui montrer Axiane, et mourir à ses yeux. Mais Taxile m'
Andromaque, 50	m'entraine, J'aime, je viens chercher Hermione	en ces lieux	, La fléchir, l'enlever, ou mourir à ses yeux. Toi
Andromaque, 7	cet amour payé de_trop d'ingratitude, Qui me rend	en ces lieux	sa présence si rude. Quelle honte pour moi ! Quel triomphe pour
Andromaque, 69	et venez -y vous-même. Voulez -vous demeurer pour otage	en ces lieux	? Venez dans tous les coeurs faire parler vos yeux. Faisons de
Andromaque, 81	m'expliquer. Vous agirez ensuite. Vous savez qu'	en ces lieux	mon devoir m'a conduite. Mon devoir m'y retient, et je
Andromaque, 10	prince de vue. Mais un heureux destin le conduit	en ces lieux	. Parlons. À tant_d'attraits, Amour, ferme ses yeux. PYRRHUS_Je
Andromaque, 10	. Il semblait qu'un spectacle si doux N'attendit	en ces lieux	qu'un témoin tel que vous. Vous y représentez tous les Grecs et
Andromaque, 23	. Et quelle est sa pensée ? Attend -elle	en ce jour	Que je lui laisse un fils pour nourrir son amour ? PHOENIX_Sans_doute.
Andromaque, 3	? Ne m'avais -tu pas dit qu'elle était	en ces lieux	? PHOENIX_Je le croyais. ANDROMAQUE, à Céphise. Tu vois le
Andromaque, 5	donner de les armes. Je croyais apporter plus_de haine	en ces lieux	. Mais, Madame, du_moins tournez vers moi les yeux : Voyez
Andromaque, 62	yeux. Je ne te retiens plus, sauve -toi	de ces lieux	. Va lui jurer la foi, que tu m'avais jurée. Va
Andromaque, 3	, qui cherche à se venger. Elle n'est	en ces lieux	que trop bien appuyée, La querelle de les Grecs à la sienne est
Andromaque, 10	le peuple épouvanté j'ai traversé la presse Pour venir	de ces lieux	enlever ma princesse, Et regagner le port, où bientôt nos amis Viendront
Athalie, 5	fut donnée. Que les temps sont changés ! Sitôt_que	de ce jour	La trompette sacrée annonçait le retour, Du_Temple orné partout de festons magnifiques,
Athalie, 92	et vais me joindre à la troupe fidèle Qu'attire	de ce jour	la pompe solennelle. JOAD. Les temps sont accomplis, Princesse, il
Athalie, 21	UNE_AUTRE. Ô mont de Sinaï, conserve la mémoire	De ce jour	à_jamais auguste et renommé, Quand sur ton sommet enflammé Dans un nuage épais
Athalie, 30	formidable. Reine, sors, a -il dit,	de ce lieu	redoutable, D'où te bannit ton sexe et ton impiété. Viens -tu
Athalie, 16	ai su soulever contre cet assassin, Il me laisse	en ces lieux	souveraine maîtresse. Je jouissais en paix de le fruit de ma sagesse.
Athalie, 92	ATHALIE. J'entends. Mais tout ce peuple enfermé	en ce lieu	, À quoi s'occupe -il ? JOAS. Il loue, il bénit
Athalie, 7	défendent l'entrée. Qui cherchez -vous ? Mon père	en ce jour	solennel De l'idolâtre impur fuit l'aspect criminel. Et devant le Seigneur
Athalie, 8	cet ennemi de Dieu_Vient-il infecter l'air qu'on respire	en ce lieu	? MATHAN. On reconnaît Joad à cette violence. Toutefois il devrait montrer
Athalie, 56	. Jérusalem, objet de ma douleur, Quelle main	en ce jour	t'a ravi tous tes charmes ? Qui changera mes yeux en deux sources
Athalie, 3	, Les parfums, et les sacrifices Qu'on devait	en ce jour	offrir sur tes autels ? UNE_FILLE_DU_CHOEUR. Quel spectacle à nos yeux timides !
Athalie, 30	. Ne craignez rien. Et nous, sortons tous	de ces lieux	. JOAS, courant dans les bras de le grand prêtre. Mon père
Bajazet, 2	ACOMAT_Viens, suis -moi. La sultane	en ce lieu	se doit rendre. Je pourrai cependant te parler, et t'entendre.
Bajazet, 6	instruit de tout ce qui se passe, Mon entrée	en ces lieux	ne te surprendra plus. Mais laissons, cher Osmin, les discours superflus
Bajazet, 124	mon trépas quand ils l'ont prononcé. Voila donc	de ces lieux	ce qui m'ouvre l'entrée, Et comme enfin Roxane à mes yeux
Bajazet, 5	, Non_plus par un silence aidé de votre adresse Disputer	en ces lieux	le coeur de sa maîtresse, Mais par de vrais combats, par de
Bajazet, 54	trop sincère. Toi, Zatime, retiens ma rivale	en ces lieux	. Qu'il n'ait en expirant que ses cris pour adieux. Qu'
Bajazet, 9	parvenue à les yeux de ma rivale ? J'étais	en ce lieu	même, et ma timide main. Quand Roxane a paru, l'a cachée

Fig. 2 Results for query B2 in the Racine plays, using TXM (Schöch, CC BY).

‘Since on these shores the Gods sent /
The daughter of Minos and Pasiphae.’

(6) […] et je puis, dès ce jour,
Accomplir le dessein qu’a formé mon amour. (*Bajazet*, II, 1)
‘and I can, from this day /
Accomplish the plan which my love formed.’

The corresponding CQP query is only slightly more complex than in the first example, and similarly doesn’t even require any linguistic or semantic annotation, due to the limited range of word forms involved:

```
[word="de|en|sur"%c][word="ce|ces"%c][]?[word="lieu.?|bords|jour|
jours"%c]
```

Essentially, this query defines a sequence of four tokens, each delimited by square brackets, corresponding to either *en ces lieux*, *sur ces bords* or *en ces jours*, with some flexibility for one or several words to appear in front of the final noun. This latter element is added because Spitzer gives one example where an adjective is placed in that position. This query corresponds to 107 instances in all 11 plays (Figure 2).

The most frequent paraphrase is clearly *en ces lieux* (52 instances). The quality assessment test shows that recall is perfect, with this pattern: all 16 of Spitzer’s examples are included in the instances found. However, precision is not flawless, with a small number of instances where the literal sequence does not form a paraphrase for *here* or *now*, as in the verses: “Et l’aspect de ces lieux où vous la retenez, N’a rien dont mes regards doivent être étonnés” (*Britannicus*, III, 8; ‘And the look of these places where you hold her, has nothing of which my glances must be astonished’).[13]

Spitzer claims that the literal *ici* is rather rare in Racine’s plays. A quick search shows 152 instances of *ici* in Racine’s plays, clearly outnumbering the synonymous paraphrases, although not all of them are amenable to a replacement by one of the paraphrases mentioned by Spitzer. Whether or not this should be considered rare, however, cannot be ascertained without a deeper qualitative analysis as well as a comparison of the ratios between paraphrase and direct expression in Racine’s and his contemporaries’ works.

[13] Again, the full table of results can be found in the “analysis” folder in the project repository (B2.csv).

5.3 Example: Das entgrenzende *où*

As a final example, I would like to discuss another stylistic pattern from the group of contour-erasing devices (group F) that Spitzer calls "das entgrenzende *où*" (F4; 'the de-bordering *where*') and that he describes as follows:

> Zur Konturverwischung trägt auch bei das entgrenzende 'où' ("wo"), das besonders gern bei Abstrakten eintritt, bei denen man sich schwer eine Örtlichkeit, einen umzirkten Raum denken kann, in den das Wo eindringen könnte (*Dämpfung*, 168).[14]
>
> "The delimiting "où" ('where') also contributes to the blurring of contours, which occurs especially readily in abstract nouns for which it is difficult to imagine a locality, a circumscribed space, into which the 'where' could penetrate."

Among the examples for this pattern cited by Spitzer are the following:

(7) […] pour avancer cette mort où je cours (*Andromaque,* II, 2)
'[…] this death, where I aspire'

(8) Ô toi qui vois la honte où je suis descendue. (*Phèdre,* III, 2)
'[…] this humiliation, where I have sunken to'

(9) Parmi les déplaisirs où son âme se noie
Il s'élève en la mienne une secrète joie. (*Andromaque,* I, 1)
'[…] this pain, where his soul is drowning'

The key marker mentioned by Spitzer is the relative pronoun *où*, combined with a noun expressing an abstract concept. Looking at the examples, we can note that these abstract nouns are often emotion words (such as shame, melancholy, joy, happiness, or worry). We can also note that they are always followed by a verb phrase (*where I run*, *where I have descended to*, *where my soul drowns*, in the examples above). This pattern is referenced as pattern F4 in the documentation for this study. Translating this stylistic pattern into a CQP query, we can formulate the following:

[14] Engl.: "The delimiting 'où' ('where') also contributes to the blurring of contours, which occurs especially readily in abstract nouns for which it is difficult to imagine a locality, a circumscribed space, into which the 'where' could penetrate."

```
[wnlex="noun.feeling"|lemma="coeur|honte|pudeur|mélancolie|
déplaisir|penchant|chagrin|mort|hymen|trouble|mal|désespoir|
malheurs|joie|bonheur|malheur| ennui|horreur|douleur|erreur|noeud|
peine|rage|sacrifice|transport|colère|courroux|crainte|hyménée"]
[word=","]{0,1}[word="où"%c][]{0,10}[pos="verb"]
```

This again defines a sequence of tokens: first, a token defined as either a noun that is an emotion word according to the WordNet annotation or that corresponds to one in a list of emotions words missing from the WordNet list but important in the context of seventeenth-century tragedy; then, an optional token: a comma; then, a token formed by the word form *où*, whether capitalized (for example at the beginning of a verse line) or not, followed by an optional number of unspecified words; finally, a verb. The list of possible alternative nouns is so long because the coverage of WordNet is very limited, especially with regard to seventeenth-century vocabulary such as *hymen*, *nœud*, *corroux* or *hyménée*.

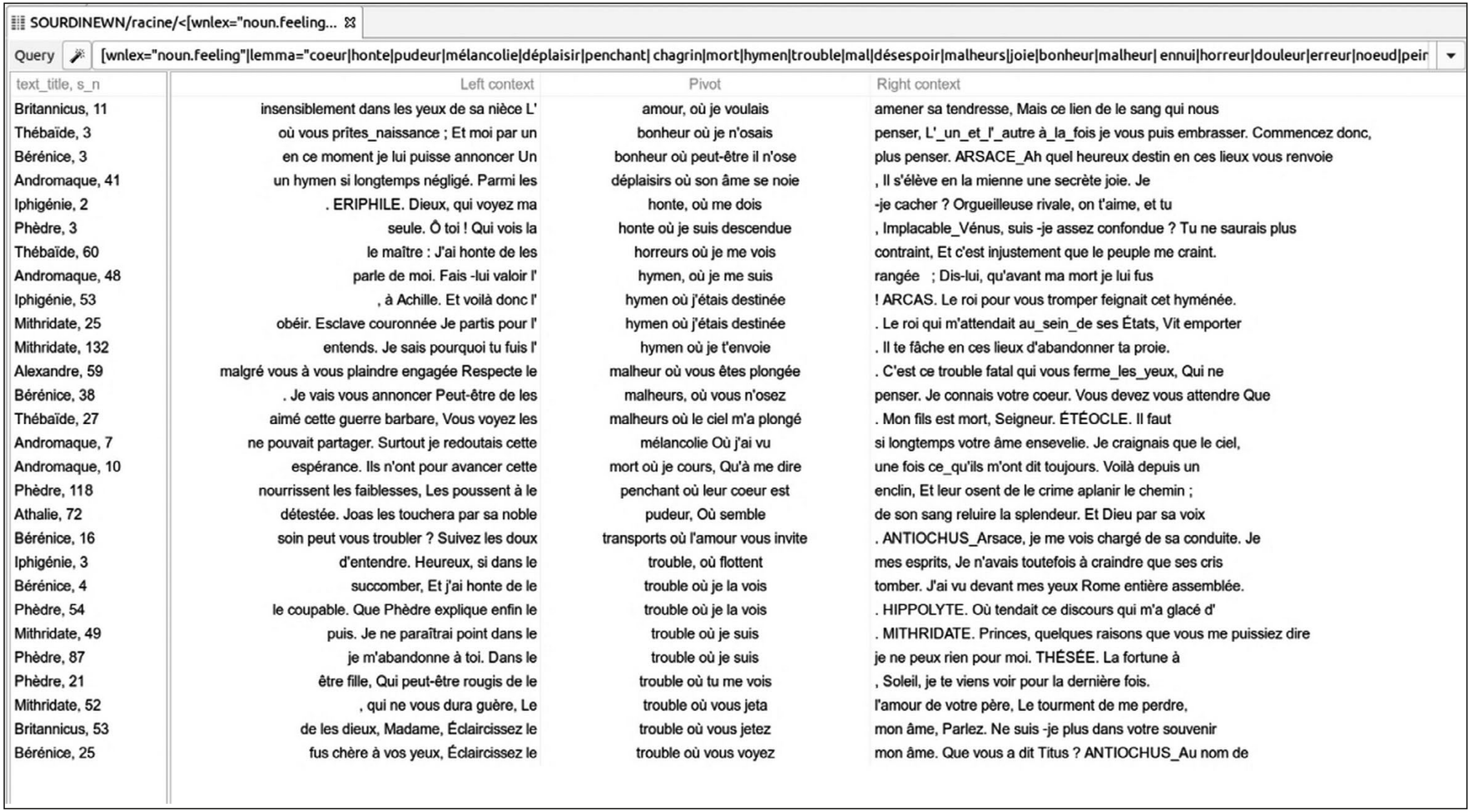
While Spitzer quotes ten instances of this pattern, this query, when applied to the eleven plays by Racine, yields 29 instances in nine different plays and depending on 14 different supporting nouns (Figure 3). The quality assessment for this query shows that the recall test is perfect, with all of Spitzer's examples being part of the retrieved set, and that the precision test is equally perfect, with no instance found using the pattern that cannot be considered to match Spitzer's description of the pattern.[15]

Something that Spitzer does not assess is that the most frequently used supporting noun is *trouble* (nine instances), followed by *hymen* (four instances). The most common personal pronoun is *je* (13 instances), followed by *vous* (six instances). Spitzer mentions that this pattern can be combined with the "de-bordering plurals" (found in patterns A1 and A2), and the supporting noun is indeed in the plural form in five out of the 28 cases, although it could be debated whether these are really all de-bordering plurals.

Due to limitations of space, it is not possible to show more examples from Spitzer, CQP queries, and their results for all 50 stylistic patterns.[16] When taking more of a bird eye's view, it can be noted that out of the 50 stylistic patterns that Spitzer distinguished, only 24 can be modeled with good or satisfactory accuracy. Several other patterns are too reduced in their formal expression to be identified with any precision, as in the case of the pattern Spitzer describes as the "Entindividualisierung durch den unbestimmten Artikel" (pattern A1; 'the dis-individualization by the indefinite article'). One instance of this patterns is the following:

[15] For a closer look, see the CSV file corresponding to the F4 query's results in the folder "analysis" of the project repository.

[16] For the remaining stylistic patterns, the corresponding queries and results can be found in the project repository.



SOURDINEWN/racine/<[wnlex="noun.feeling...

Query [wnlex="noun.feeling"|lemma="coeur|honte|pudeur|mélancolie|déplaisir|penchant| chagrin|mort|hymen|trouble|mal|désespoir|malheurs|joie|bonheur|malheur| ennui|horreur|douleur|erreur|noeud|peir

text_title, s_n	Left context	Pivot	Right context
Britannicus, 11	insensiblement dans les yeux de sa nièce L'	amour, où je voulais	amener sa tendresse, Mais ce lien de le sang qui nous
Thébaïde, 3	où vous prîtes_naissance ; Et moi par un	bonheur où je n'osais	penser, L'_un_et_l'_autre à_la_fois je vous puis embrasser. Commencez donc,
Bérénice, 3	en ce moment je lui puisse annoncer Un	bonheur où peut-être il n'ose	plus penser. ARSACE_Ah quel heureux destin en ces lieux vous renvoie
Andromaque, 41	un hymen si longtemps négligé. Parmi les	déplaisirs où son âme se noie	, Il s'élève en la mienne une secrète joie. Je
Iphigénie, 2	. ERIPHILE. Dieux, qui voyez ma	honte, où me dois	-je cacher ? Orgueilleuse rivale, on t'aime, et tu
Phèdre, 3	seule. Ô toi ! Qui vois la	honte où je suis descendue	, Implacable_Vénus, suis -je assez confondue ? Tu ne saurais plus
Thébaïde, 60	le maître : J'ai honte de les	horreurs où je me vois	contraint, Et c'est injustement que le peuple me craint.
Andromaque, 48	parle de moi. Fais -lui valoir l'	hymen, où je me suis	rangée ; Dis-lui, qu'avant ma mort je lui fus
Iphigénie, 53	, à Achille. Et voilà donc l'	hymen où j'étais destinée	! ARCAS. Le roi pour vous tromper feignait cet hyménée.
Mithridate, 25	obéir. Esclave couronnée Je partis pour l'	hymen où j'étais destinée	. Le roi qui m'attendait au_sein_de ses États, Vit emporter
Mithridate, 132	entends. Je sais pourquoi tu fuis l'	hymen où je t'envoie	. Il te fâche en ces lieux d'abandonner ta proie.
Alexandre, 59	malgré vous à vous plaindre engagée Respecte le	malheur où vous êtes plongée	. C'est ce trouble fatal qui vous ferme_les_yeux, Qui ne
Bérénice, 38	. Je vais vous annoncer Peut-être de les	malheurs, où vous n'osez	penser. Je connais votre coeur. Vous devez vous attendre Que
Thébaïde, 27	aimé cette guerre barbare, Vous voyez les	malheurs où le ciel m'a plongé	. Mon fils est mort, Seigneur. ÉTÉOCLE. Il faut
Andromaque, 7	ne pouvait partager. Surtout je redoutais cette	mélancolie Où j'ai vu	si longtemps votre âme ensevelie. Je craignais que le ciel,
Andromaque, 10	espérance. Ils n'ont pour avancer cette	mort où je cours, Qu'à me dire	une fois ce_qu'ils m'ont dit toujours. Voilà depuis un
Phèdre, 118	nourrissent les faiblesses, Les poussent à le	penchant où leur coeur est	enclin, Et leur osent de le crime aplanir le chemin ;
Athalie, 72	détestée. Joas les touchera par sa noble	pudeur, Où semble	de son sang reluire la splendeur. Et Dieu par sa voix
Bérénice, 16	soin peut vous troubler ? Suivez les doux	transports où l'amour vous invite	. ANTIOCHUS_Arsace, je me vois chargé de sa conduite. Je
Iphigénie, 3	d'entendre. Heureux, si dans le	trouble, où flottent	mes esprits, Je n'avais toutefois à craindre que ses cris
Bérénice, 4	succomber, Et j'ai honte de le	trouble où je la vois	tomber. J'ai vu devant mes yeux Rome entière assemblée.
Phèdre, 54	le coupable. Que Phèdre explique enfin le	trouble où je la vois	. HIPPOLYTE. Où tendait ce discours qui m'a glacé d'
Mithridate, 49	puis. Je ne paraîtrai point dans le	trouble où je suis	. MITHRIDATE. Princes, quelques raisons que vous me puissiez dire
Phèdre, 87	je m'abandonne à toi. Dans le	trouble où je suis	je ne peux rien pour moi. THÉSÉE. La fortune à
Phèdre, 21	être fille, Qui peut-être rougis de le	trouble où tu me vois	, Soleil, je te viens voir pour la dernière fois.
Mithridate, 52	, qui ne vous dura guère, Le	trouble où vous jeta	l'amour de votre père, Le tourment de me perdre,
Britannicus, 53	de les dieux, Madame, Éclaircissez le	trouble où vous jetez	mon âme, Parlez. Ne suis -je plus dans votre souvenir
Bérénice, 25	fus chère à vos yeux, Éclaircissez le	trouble où vous voyez	mon âme. Que vous a dit Titus ? ANTIOCHUS_Au nom de

Fig. 3 Results for query *F4* in Racine's plays, using TXM.

(10) Le croirai-je, seigneur, qu'un reste de tendresse
Vous fasse ici chercher une triste princesse? (*Andromaque*, II, 2)
'Will I believe him, Lord, that a remainder of tenderness /
makes you seek here a sad princess?'

Whether or not the (naturally very widespread) use of *un(e)* or *des* in a given verse corresponds to what Spitzer has in mind here is difficult to ascertain formally, as it depends very much on the context, notably on who are the speaker and adressee of the phrase in question. Other patterns depend strongly on the varying degree of metric and rhetorical complexity of verses, something which is also out of scope for the current annotation schema, but which would of course be an interesting expansion for future work. An example of this is the pattern described by Spitzer as "ganz einfache Verse oder Halbverse, die auf eine hochrhetorische Versreihe folgen" (pattern K2, *Dämpfung* 228; "very simple verses or half-verses following a series of highly rhetorical verses"). Both metrical structure and rhetorical complexity would need to be identified and annotated with a high level of reliability in order to make such patterns automatically identifiable.[17]

What also becomes clear, however, is that the procedure so far has already some advantages. For example, instead of relying only on the selection of examples given by Spitzer, the procedure used here allows us to identify many more (if not all) matching instances in Racine's work and as a consequence, get a more precise sense of the range of variants of a pattern and the distribution of the pattern in the plays. However, the key question raised in the introduction, namely: whether the dampening effect is an authorial style (proper only to Racine) or a period style (characteristic of Classical tragedy as a whole) cannot be answered when just looking at Racine's plays: this requires making a comparison, both qualitatively and quantitatively, between the instances found in Racine with the instances found in tragedies written by Racine's contemporaries. This is what the next section attempts to do.

6. Expansion: Racine and His Contemporaries

As we have already seen in section 3 of this paper, Spitzer programmatically describes his approach as describing Racine's style not in relation to other authors, but as a self-sufficient object, to which he adds:

[17] For the description of an automatic tool for metrical annotation of French Classical verse, see Beaudouin and Yvon (1996). For a survey of the state of the art in metaphor detection, see Rai and Chakraverty (2021).

> Die Aufgabe, diese sozusagen absoluten Dämpfungen in ihrer relativen Stärke gegenüber den Vorbildern darzustellen und historisch zu begreifen, bleibt anderen Mitforschern vorbehalten. (*Dämpfung,* 136)
>
> "The task of presenting these, so to speak, absolute attenuations in their relative strength compared to the models and to understand them historically is left to other fellow researchers."

So far, this study has respected Spitzer's focus on Racine alone. However, I would now like to expand the focus of the study beyond the Racine cosmos not so much to his predecessors, but to his direct contemporaries. The text collection used for this is an expanded dataset including the eleven plays by Racine as well as the 38 plays by his contemporaries, as described above. The approach will be primarily comparative.[18]

Applying the same search queries to this expanded dataset requires some adjustments. Not all patterns can be transposed directly to the larger dataset, with its wider range of authors and plays. For this reason, most queries were enhanced or 'generalized' in this step, for example with additional terms being included among the alternatives. However, these expanded queries were then also used for a second round of analyses on Racine alone and have been incorporated in the results presented above. The aim of this procedure is to avoid, as much as possible, any bias that would lead to results favoring higher number of instances in Racine compared to the contemporaries (any remaining bias is most likely in his favor, though). Also, in this constellation, there is no possibility to do the same kind of quality checks as in the first part of this study, as there is no reference any more to establish the recall. However, the same check for false positives (precision) can of course be made manually.

Using the same queries as above on Racine alone for illustration, we can observe some relevant effects. Taking the case of pattern F4 from the analysis above as an example, it can first of all be noted that there are indeed some additional nouns of emotion that appear in relevant patterns, notably *maux* and *désespoir*, but also *joie* and *ennuis* and several others. However, there is also a small number of false positives.

[18] Spitzer argued against this kind of approach, which is incompatible with his contention that Racine's work should be treated as a unique and unified stylistic object: "Es läßt sich also streng genommen kein Zug der Sprache des Dichters isolieren und mit parallelen Zügen der Sprache anderer Dichter, die ebenfalls aus ihrem Kontext isoliert werden, vergleichen; die einzelnen Züge einer Dichtung sind vorerst miteinander zu vergleichen, als Glieder, Elemente, Träger eines Systems, einer in sich ruhenden Einheit." (*Dämpfung,* 257; "Strictly speaking, therefore, no trait of the poet's language can be isolated and compared with parallel traits of the language of other poets, which are also isolated from their context; the individual traits of a poem must first be compared with one another, as links, elements, carriers of a system, a unity at rest in itself").

In the following example, the sequence of words fits the pattern, but the underlying construction is different:

> (11) M'offre un sujet de joie où j'en voyais d'ennui (Pierre Corneille, *Agesilas*, 1666)
> 'Offers me a subject of joy / where I used to see boredom.'

A second observation is that there is a large number of additional instances, with 166 instances found among the contemporaries in addition to the 29 instances found in Racine, for a total of 195. Relative to the lengths of the plays, it appears there are 1.7 instances of this pattern per 10,000 tokens in Racine's plays, but 2.1 instances per 10,000 tokens in the contemporaries' tragedies.[19]

This kind of comparative analysis can be performed for all 24 patterns that were successfully modeled for retrieval in the corpus (see Figure 4 for an overview). It turns out that eleven out of the 24 patterns are over-represented in Racine, but only in one case is that difference statistically significant. The remaining 13 patterns are under-represented in Racine, but only in four cases is that difference statistically significant. This means that 19 out of 24 patterns (or 79 percent) do not vary in prevalence between Racine and his contemporaries with any statistical significance.[20]

From this second step of the analysis, it can be concluded that the stylistic patterns producing the dampening effect are by no means exclusive to Racine.[21] Rather, it seems that the patterns Leo Spitzer identified in Racine's style are actually quite typical of many tragedies written and performed in the second half of the seventeenth century. In this sense, they appear to be a genre style at the very least. Whether or not they are also prevalent in literary genres other than tragedy in verse, during the same period, or whether they are also similarly prevalent in other periods, is something I need to leave to further research at this stage.

However, if the stylistic patterns that Leo Spitzer detected in Racine's plays are not specific to his style and vocabulary, what is? There is a substantial tradition in the fields of information retrieval, corpus or computational linguistics, and computational literary studies regarding the extraction of words or other features that are distinctive or characteristic of one group of texts when compared to another group of texts. These features are then often called *keywords* or *key features*. Such distinctive features can be

[19] This means the pattern is clearly under-represented in Racine (with Racine using the pattern only at 66 percent of the level of his contemporaries). However, a Wilcoxon rank-sum test shows that this difference is not statistically significant (p-value = 0.30).

[20] Further details can be found in the corresponding folder of the companion repository.

[21] Of course, this can only be a statement about the 24 patterns modeled here and does not preclude that some or even many of the remaining 26 patterns turn out to be strongly distinctive of Racine.

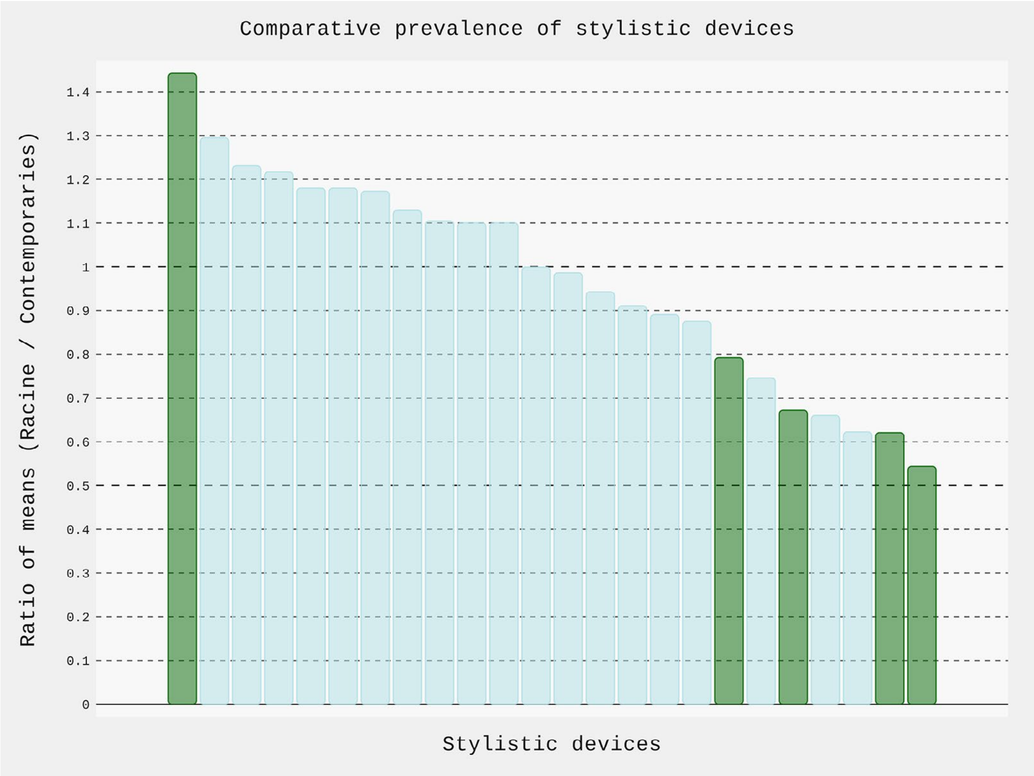


Fig. 4 Comparative prevalence of stylistic patterns in Racine's work and the works by his contemporaries. Ordered by ratio of mean relative frequency. Statistically significant differences are highlighted in green. An interactive version of this graph is available in the companion repository (Schöch, CC BY).

extracted for various kinds of tokens (for example word forms, lemmata or part-of-speech) as well as for unigrams, bigrams or trigrams, etc.[22]

[22] A good introduction to several such measures of keyness is provided by Paquot and Bestgen (2009). An important, more technical evaluation study is Lijffijt et al. (2016). Since John Burrows introduced it in 2007, a keyness measure he called *Zeta* has received quite a lot of attention in computational literary studies; see Burrows (2007), Craig and Kinney (2009), Hoover (2012), Schöch et al. (2018) and Schöch (2018). This last measure, implemented in the *pyzeta* tool, has been used in the following analysis. See: https://github.com/cligs/pyzeta. The code actually used corresponds to release v0.5.0, 2017 of *pyzeta* (DOI: https://doi.org/10.5281/

This approach basically turns the approach pursued in the last section on its head: instead of specifying a certain number of stylistic patterns and checking whether they are characteristic of Racine or not, compared to his contemporaries, we now ask: given a certain type of feature definition (like part-of-speech bigrams), which ones among all such features are the most distinctive of Racine, and which ones are the most distinctive of his contemporaries? This also means that we do not start out anymore with a certain stylistic hypothesis and interpretive focus (of the kind: assuming Racine's preference for a muted expression of emotions, which relevant stylistic patterns can we identify in his works?). Instead, we now start only with the assumption of the existence of some difference, and of the applicability of a statistical method, but without a clear hypothesis of what it will bring to the fore. This is all the more true as the results up to now have not suggested a very clear contrast between the two groups of texts under study.

When applying the contrastive analysis using the *pyzeta* implementation, the results for unigrams of word forms, certainly the simplest case, are as described below when focusing on just the nouns (Figure 5).[23]

As can be seen from the visualization, Racine has a preference for several words related to family relationships (*mère* 'mother', *aieux* 'ancestor', *frère* 'brother'). Others can be related to a characteristic vision of the tragic plot (*autel* 'altar', *nuit* 'night', *larme* 'teardrop', *cri* 'cry', but also *silence* 'silence'). At least one of these distinctive nouns can be related to a stylistic pattern described by Spitzer, namely *pas* ('steps'): it is a key element in pattern F7, the periphrases with a verb like *porter ses pas* or *guider ses pas* that are somewhat over-represented in Racine (see above).

Similarly, such a contrastive analysis can be done focusing on the adjectives (Figure 6). In the list of the Racinian adjectives, the balance is clearly on the side of the negative (*ennemi* 'hostile', *sévère* 'severe', *infortuné* 'unfortunate', *étranger* 'stranger', *homicidal* 'homicide', *perfide* 'treacherous', *farouche* 'fierce', *odieux* 'hateful', *funeste* 'fatal', *impuissant* 'impotent') rather than the positive adjectives (*superbe* 'superb', *jeune* 'young', *éternel* 'eternal', *content* 'content', *sacré* 'sacred', *libre* 'free', *tranquille* 'quiet', *immortel* 'immortal'). Overall, it should be noted that these lists of words are rather hard to interpret. More abstract features, such as part-of-speech trigrams, tend to be even harder to interpret.

zenodo.844555), but is also included with all input and output in the project repository linked above.

[23] The parameters used are the following: The Zeta variant sd2 has been used, as it has proven to be particularly robust in Schöch et al. (2018). This Zeta variant differs from the standard Zeta calculation in that a log transformation is applied to the document proportions. Text segments of a length of 3,000 words have been used, which is rather short but appears appropriate given that Classical tragedies typically only have 14–15,000 words.

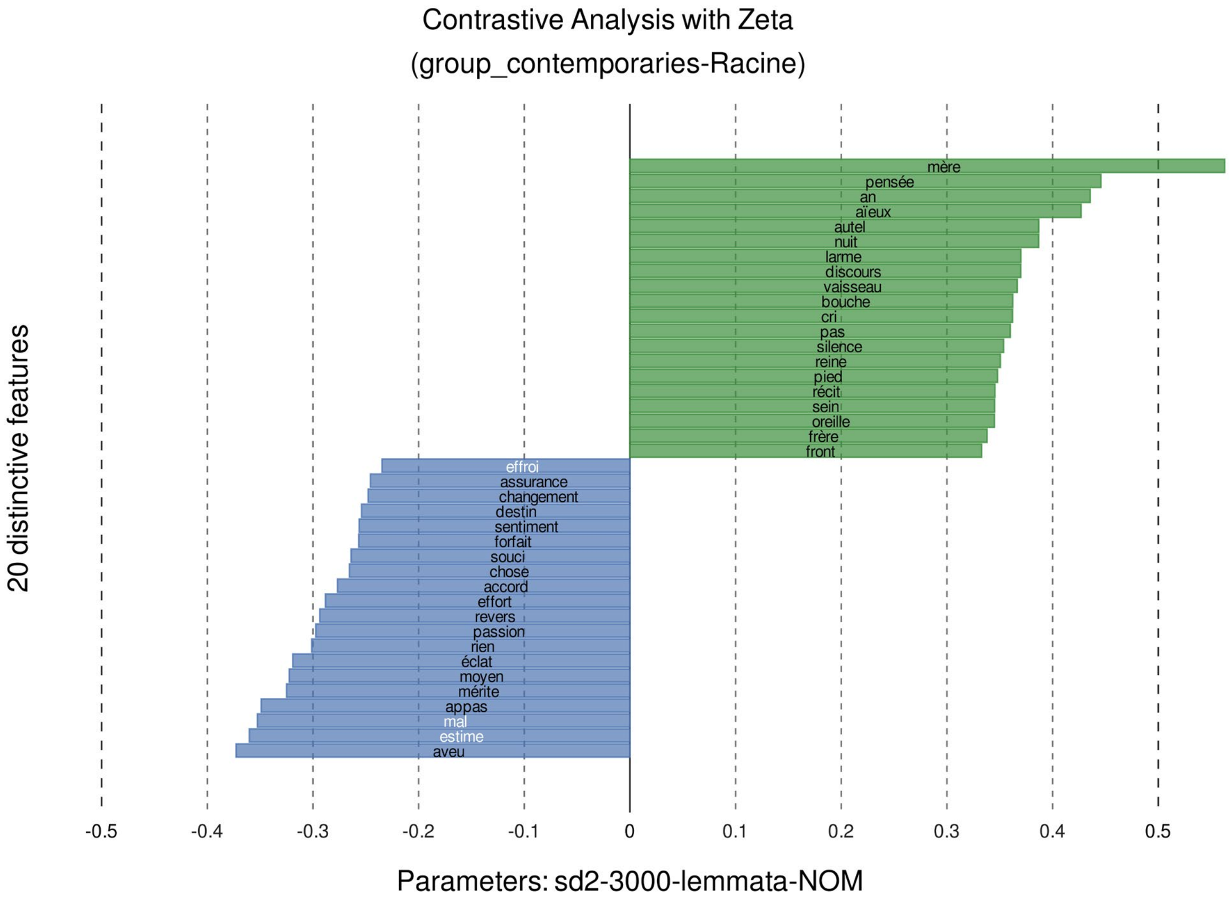


Fig. 5 Distinctive nouns for Racine (right) compared with his contemporaries (left). Implementation: https://github.com/cligs/pyzeta (Schöch, CC BY).

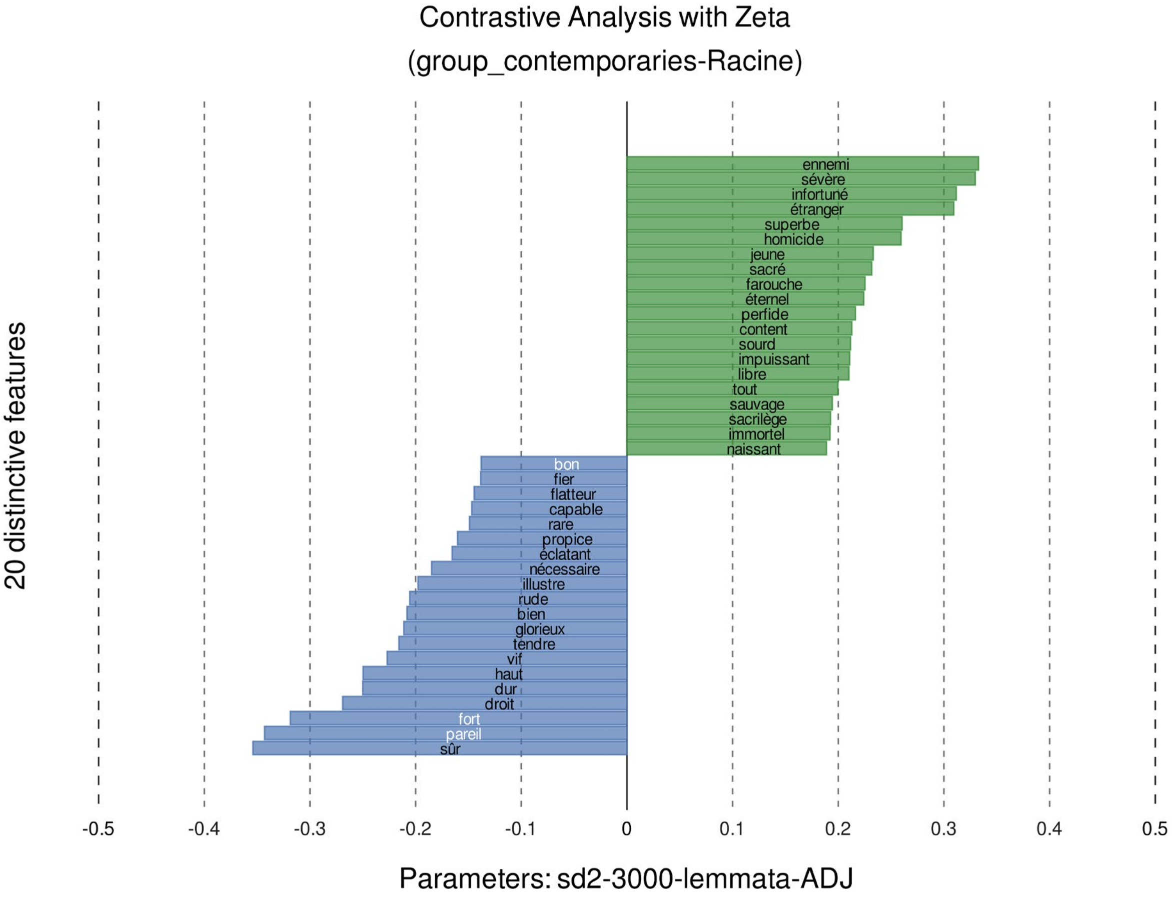


Fig. 6 Distinctive adjectives for Racine (right) compared with his contemporaries (left). Implementation: https://github.com/cligs/pyzeta (Schöch, CC BY).

7. Conclusion and Future Work

The initial idea of a digital replication of Spitzer's study of Racine's style has led from formal operationalization of a close stylistic reading using TXM to rather more quantitative approaches, including comparative and contrastive analyses of tragedies by Racine and his contemporaries.

Regarding Racine and the Classical Age, it has become apparent that the stylistic patterns identified by Leo Spitzer, and which in the aggregate produce a dampening effect, again according to Spitzer, are even more frequent in Racine's work than Spitzer showed. In addition, the analysis was able to show that these patterns are present in varying degrees in Racine's work and in the work of his contemporaries, but without very strong (or statistically significant) differences, for the most part. The dampening effect, if understood as the aggregate use of individual stylistic patterns, is therefore clearly not an authorial signature style, but the mark of an entire literary genre during a certain period, at the very least.

In terms of the notion of style that underpins the original analysis of Spitzer and the current reenactment, it can be seen that the two notions are rather different.[24] For Spitzer, style is the unifying principle of Racine's work that is both a general principle and is manifest in a multiplicity of details that each are seen as deriving from, and at the same time reinforcing, the postulated unifying principle. Spitzer's analysis is, in this sense, inherently interpretative. At the same time, but more implicitly, the notion of style underlying Spitzer's analysis is also one of style as a deviation from a norm: what is notable, for a stylistic analysis, is what is different from general expectations. By contrast, the definition of style in quantitative stylistics, as evidenced in the last section of this paper, starts with a statistical operationalization of distinctiveness, in our case via the Zeta measure, and is explicitly contrastive, but does not have a strong hypothesis as to how the stylistic features identified as being characteristic of the target group of text can be interpreted. Another difference is that Spitzer's notion of style supports and even programmatically requires the focus on a single author. An algorithmic approach to style, and particularly any quantification, however, requires explicit target and comparison domains in order to give meaning to the varying degrees of prevalence of stylistic phenomena.

With regard to the formal modeling of the stylistic patterns identified by Leo Spitzer, this has been successful in only 24 out of 50 cases. To a considerable degree, this is due to limitations in the annotations that I was able to create. For example, it would probably be helpful if the semantic annotation had a much larger coverage,

[24] For a more general investigation into notions of literary style and how they have shifted between 1950 and the advent of computational stylistics, see Herrmann, van Dalen-Oskam, and Schöch (2015).

either through an expansion of WordNet or through the use of an appropriate word embedding model.[25] Similarly, a syntactically parsed text would probably allow for more nuanced analyses of stylistic patterns. Finally, it would be of use for several patterns if metrical annotation could also be used. Providing such improved or additional annotations for the challenging language of seventeenth-century verse tragedies was outside the scope of this study. In part, the limitations are also due to the fact that some patterns defined by Spitzer depend on interpretation in the larger context and are probably hard to model even with more annotations. A more controlled approach to the generalization of the patterns, with additional checks for the appropriate coverage, would also be important. Finally, the extraction of key features has used a rather simple approach based on single word forms. Recent, more sophisticated approaches in phraseology, as exemplified for example in the *Phraseorom* project, would certainly be worth pursuing in order to achieve the goal of an algorithmic identification of complex and significant stylistic patterns.[26] An obvious upside of the digital approach to the stylistic patterns is the possibility to transfer a given analysis, with just a little adaptation, to a comparison corpus for further analysis.

These strengths and limitations of the digital approach are the mirror image, in a way, of Spitzer's original study. He would probably not have agreed neither with the rough operationalizations of the stylistic devices, nor with the quantitative comparison of Racine to his contemporaries that I have practiced here. Spitzer is much more nuanced, flexible and mindful of the semantic and pragmatic context of a stylistic device, when identifying and interpreting relevant passages, compared to the many rough and ultimately often imprecise operationalizations used here. However, Spitzer's analysis would also be extremely time-consuming to transfer to a corpus of different authors. Not just because all of the material needs to be united, but more importantly, because a principle of unity different from Racine's would most likely need to be postulated and pursued, if one wanted to do justice both to Spitzer's approach and the contemporary authors. Finally, Spitzer's study is not easy to replicate: he never even mentions the edition of Racine's work that he used, he sometimes describes patterns with precision, but in other instances describes rather their purported effect on readers or provides individual examples in order to explain what he means by a given pattern; finally, he refrains from formulating hypotheses that could be tested algorithmically and statistically.

Taking up this last point, I would like to also return to my initial description of this study as repetitive research. In terms of the repeating aspect, I believe that attempting to repeat or reenact Spitzer's earlier study has been partially successful, in the sense

[25] For an introduction to word embedding models used in computational literary studies, see Schöch (2023a).

[26] See e.g. Kraif (2016), Novakova and Siepmann (2020) and Jacquot, Vidotto, and Gonon (2023, this volume).

that (a) roughly 50 percent of the stylistic patterns Spitzer defined could be transposed into the digital and algorithmic paradigm and (b) a close investigation of all textual instances of these patterns, both in Racine and in the works of his contemporaries, could be conducted. In terms of the repeatable aspect, things are more challenging. Certainly, with the complete set of annotated texts available exactly as used for this study, an important part of the project documentation is present. However, documenting the queries and their results based on these texts in an entirely transparent manner is difficult, because copying the query from the documentation into TXM and saving the results to a file needs to be done manually, something which can easily introduce accidental inconsistencies. Similarly, creating the comparison statistics involves another break in the toolchain, as this is not done directly in TXM but in Python, based on data exported manually from TXM. The whole process would be more transparent, and more directly replicable and executable, if it was all done in the same environment, either by integrating the subsequent analyses into TXM using a Groovy-script and automatically exporting all results and intermediary data, or by performing the initial analyses directly using a script-controlled CQP instance. Such an approach has not been used in the present study, but at least the tools used are either well-established in the community (TXM) or available in the long term (*pyzeta* / *pydistinto*).

Ultimately, I hope and believe that the obvious shortcomings of this study do not induce the conclusion that computational literary studies is a futile enterprise. Rather, I hope that in the near future, someone with the required technical expertise in literary and linguistic annotation and bringing to bear new statistical methods of identifying key stylistic devices will further bridge the gap between the nuanced stylistic close readings of a Leo Spitzer with the scope, speed and scalability of algorithmic approaches.

ORCID®

Christof Schöch https://orcid.org/0000-0002-4557-2753

References

Beaudouin, Valérie, and François Yvon. 1996. "The Metrometer: A Tool for Analysing French Verse." *Literary and Linguistic Computing* 11 (1): 23–31. https://doi.org/10.1093/llc/11.1.23.

Bernet, Charles. 1983. *Le Vocabulaire des tragédies de Jean Racine: analyse statistique*. Genève: Slatkine-Champion. http://catalog.hathitrust.org/api/volumes/oclc/11357992.html.

Bray, René. 1926. *La Formation de la doctrine classique en France.* Paris: Nizet.

Burrows, John. 2007. "All the Way Through: Testing for Authorship in Different Frequency Strata." *Literary and Linguistic Computing* 22 (1): 27–47. https://doi.org/10.1093/llc/fqi067.

Craig, Hugh, and Arthur F. Kinney, eds. 2009. *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511605437.

Evert, Stefan, and Andrew Hardie. 2011. "Twenty-First Century Corpus Workbench: Updating a Query Architecture for the New Millennium." In *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham: University of Birmingham. https://eprints.lancs.ac.uk/id/eprint/62721.

Fièvre, Paul, ed. 2007–2022. *Théâtre classique*. http://theatre-classique.fr/.

Fidler, Fiona, and John Wilcox. 2021. "Reproducibility of Scientific Results." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: Metaphysics Research Lab. https://plato.stanford.edu/archives/sum2021/entries/scientific-reproducibility/.

Génetiot, Alain. 2005. *Le Classicisme*. Paris: Presses universitaires de France. http://catalog.hathitrust.org/api/volumes/oclc/57526639.html.

Heiden, Serge. 2010. "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme." November 4, 2010. http://halshs.archives-ouvertes.fr/halshs-00549764/en.

Herrmann, Berenike, Karina van Dalen-Oskam, and Christof Schöch. 2015. "Revisiting Style, a Key Concept in Literary Studies." *Journal of Literary Theory* 9 (1): 25–52. https://zenodo.org/record/50893.

Hoover, David L. 2012. "The Tutor's Story: A Case Study of Mixed Authorship." *English Studies* 93 (3): 324–39. https://doi.org/10.1080/0013838X.2012.668791.

Jacquot, Clémence, Ilaria Vidotto, and Laetitia Gonon. 2024. "Digital Stylistic Analysis in 'PhraseoRom': Methodological and Epistemological Issues in a Multidisciplinary Project." In *Digital Stylistics in Romance Studies and Beyond*, edited by Robert Hesselbach, José Calvo Tello, Ulrike Henny-Krahmer, Daniel Schlör, and Christof Schöch, 261–278. Heidelberg: heiUP.

Jaubert, Anna. 1996. "Le Style et la vision. L'héritage de Léo Spitzer." *L'Information grammaticale* 70: 25–30.

Kraif, Olivier. 2016. "Le Lexicoscope : Un outil d'extraction des séquences phraséologiques basé sur des corpus arborés." *Cahiers de Lexicologie* 1: 91–106.

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila. 2016. "Significance Testing of Word Frequencies in Corpora [2014]." *Digital Scholarship in the Humanities* 31 (2): 374–97. https://doi.org/10.1093/llc/fqu064.

Novakova, Iva, and Dirk Siepmann, eds. 2020. *Phraseology and Style in Subgenres of the Novel: A Synthesis of Corpus and Literary Perspectives*. Cham: Palgrave Macmillan.

Paquot, Magali, and Yves Bestgen. 2009. "Distinctive Words in Academic Writing: A Comparison of Three Statistical Tests for Keyword Extraction." In *Corpora: Pragmatics and Discourse*, edited by Andreas H. Jucker, Daniel Schreier, and Marianne Hundt, 247–269. Brill Rodopi. https://doi.org/10.1163/9789042029101_014.

Rai, Sunny, and Shampa Chakraverty. 2021. "A Survey on Computational Metaphor Processing." *ACM Computing Surveys* 53 (2): 24:1–24:37. https://doi.org/10.1145/3373265.

Rockwell, Geoffrey. 2016. "Replication as a Way of Knowing in the Digital Humanities." Lecture given at Julius-Maximilians-Universität, Würzburg.

Rohou, Jean. 2005. *Jean Racine: Bilan critique*. Paris: Colin.

Scherer, Jacques. 1950. *La Dramaturgie classique en France*. Paris: Nizet.

Schöch, Christof. 2018. "Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie." In *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*, edited by Toni Bernhart, Sandra Richter, Marcus Lepper, Marcus Willand, and Andrea Albrecht, 77–94. Berlin: de Gruyter. https://www.degruyter.com/view/books/9783110523300/9783110523300-004/9783110523300-004.xml.

Schöch, Christof. 2023a. "Quantitative Semantik: Word Embedding Models für literaturwissenschaftliche Fragestellungen." In *Digitale Literaturwissenschaft. Beiträge des DFG-Symposiums 2017*, edited by Fotis Jannidis, 535–62. Stuttgart: Metzler.

Schöch, Christof. 2023b. "Repetitive Research. A Conceptual Space and Terminology of Replication, Reproduction, Re-Implementation, Re-Analysis, and Re-Use in Computational Literary Studies." *International Journal of Digital Humanities* 5, 373–403. https://doi.org/10.1007/s42803-023-00073-y.

Schöch, Christof, Daniel Schlör, Albin Zehe, Henning Gebhard, Martin Becker, and Andreas Hotho. 2018. "Burrows' Zeta: Exploring and Evaluating Variants and Parameters." In *Book of Abstracts of the Digital Humanities Conference*. Mexico City: ADHO. https://dh2018.adho.org/burrows-zeta-exploring-and-evaluating-variants-and-parameters/.

Schöch, Christof, Karina van Dalen-Oskam, Fotis Jannidis, Maria Antoniak, and David Mimno. 2020. "Panel: Replication and Computational Literary Studies." In *Digital Humanities 2020: Book of Abstracts*. Ottawa: ADHO. https://doi.org/10.5281/zenodo.3893427.

Spitzer, Leo. 1931. "Die Klassische Dämpfung bei Racine [1928]." In *Romanische Stil- und Literaturstudien I*, 135–268. Marburg: Elwert.

Spitzer, Leo. 1970. "L'effet de sourdine dans le style classique: Racine [1928/1931]." In *Études de Style*, translated by Alain Coulon, 208–335. Paris: Gallimard.

Viala, Alain. 1990. *Racine: La stratégie du caméléon*. Paris: Seghers.

Family Resemblance in Genre Stylistics
A Case Study with Nineteenth-Century Spanish-American Novels

Ulrike Henny-Krahmer

Abstract Family resemblance is a concept that has been introduced into genre theory as an analogy to describe partial and overlapping similarities between different works of the same genre. The concept aims to capture continuities and shifts in historically changing realizations of genres. In this article, family resemblance is applied in a digital genre stylistics analysis of subgenres of nineteenth-century Spanish-American historical and sentimental novels. A formal implementation of the concept is developed, and its usefulness in a corpus-based and quantitative setup is tested. For this, networks of nearest neighbors are built on topic features, and subgroups in the networks are identified with community detection. As a result, the concept of family resemblance itself undergoes change, but the digital methods also appear in a new light as tools that can be engaged for soft categorization.

Keywords genre stylistics, genre theory, categorization, family resemblance, nineteenth century, Spanish-American novel, subgenres, network analysis, topic modeling

1. Genre Categories, Family Resemblance, and Digital Genre Stylistics

A central aspect of literary genre theory is the discussion about the ways in which genres can be defined. More precisely, this includes the question of what kind of categories genres can be understood and conceived as. In this context, it is assumed that the function of generic terms is to represent a group of texts associated with them adequately. The question about the categorial status of genres is of importance both

Henny-Krahmer, Ulrike: Family Resemblance in Genre Stylistics. A Case Study with Nineteenth-Century Spanish-American Novels, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 149–171. DOI: https://doi.org/10.17885/heiup.1157.c19370

on a systematic level as well as from a historical perspective. The system is paramount, for example, when literary scholars define generic terms to capture recurrent characteristics inside groups of texts and the differences between them. On the other hand, the historical perspective prevails when scholars analyze and reproduce how historical genre labels related to groups of texts that they were associated with by writers, readers, and critics of the time.[1] Dependent on the dominant perspective, different concepts of genres as categories prevail. Two basic ways to categorize can be distinguished: classificatory and typological categorization. In their purest form, classifications lead to disjunct groups of texts and do not allow for overlaps. Types, on the other hand, form the basis for fuzzy categories because a text can be more or less similar to an ideal type at the center of a category. Both ways to categorize have been advocated for and applied in genre theory and history (Müller 2010, 21; Strube 1993, 59–65; Tophinke 1997). Furthermore, the tension between strict classificatory terms and historical terms that are collectively coined, anchored in time, and hence unsharp has led to the development of flexible approaches to genre definitions. These combine necessary and optional features that texts need to have in order to be considered instances of a genre (Fricke 2010, 7–10).

An approach that completely abandons the idea of necessary common features is the semantic concept of family resemblance, which the philosopher of language Ludwig Wittgenstein developed:

> Consider, for example, the activities that we call "games". I mean board-games, card-games, ball-games, athletic games, and so on. What is common to them all?—Don't say: "They must have something in common, or they would not be called 'games'"—but look and see whether there is anything common to all. For if you look at them, you won't see something that is common to all, but similarities, affinities, and a whole series of them at that [...]. And the upshot of these considerations is: we see a complicated network of similarities overlapping and criss-crossing [...]. I can think of no better expression to characterize these similarities than "family resemblance"; for the various resemblances between members of a family—build, features, colour of eyes, gait, temperament, and so on and so forth − overlap and criss-cross in

[1] Ultimately, these two perspectives cannot be strictly separated from each other because literary scholars setting up definitions of literary genres will not completely ignore historical conventions of the terms, as these are at least in part also transmitted through literary history. On the other hand, historical labels are not entirely free of any systematic relationship to textual patterns, either. At the same time, both complexes are confronted with historical changes affecting the relationship between the terms and the subjects. However, there has been much discussion about the distinction and mediation of genre theory and genre history, see Neumann and Nünning (2007, 9–10).

> the same way.—And I shall say: “games” form a family (Wittgenstein 2009, §66–67).

In Wittgenstein’s work, the concept serves as an analogy to describe linguistic activities involving the use of the same word for phenomena with partial and indirect similarities, like the ones seen between different family members. This analogy was adopted in literary genre theory in the 1960s and it became popular because it allowed for more open definitions of genre. According to the family resemblance concept, not all genre-relevant features need to be present in all literary works assigned to it. However, the concept was also criticized as being too loose, mainly because the boundaries between different categories are not defined sharply (Fishelov 1993, 53–68; Fricke 2010, 8–9).[2] Nonetheless, its potential value for genre theory has been emphasized: “I believe that genre theory within literary studies can, on the basis of the concepts of family resemblance and prototypes, manage to realign key questions, especially those arising from the polysemy and historicity of genre concepts” (Hempfer 2014, 414).[3]

The categorization of literary texts by genre is also one of the key concerns of digital stylistics but the discussion about the categorical status of genres that took place in literary genre theory is usually not directly addressed in digital approaches. In many literary stylistic papers, classificatory approaches are used, focusing on which features are most suitable to capture differences between genres. As features, such studies include aspects of style and content, but also structural characteristics of the texts such as text order or representation of speech (Calvo Tello 2018; Gianitsos et al. 2019; Henny-Krahmer 2018; Hettinger et al. 2016; Schöch 2017; Schöch et al. 2016; Underwood 2015). Other studies concentrate on differences between authors, not genres (Calvo Tello et al. 2017; Schöch 2013). Digital studies on the classification of genres, in the strict sense of logical classes, are relevant for a number of reasons. They help to model and interpret textual cues that are crucial to recognizing genres. They also contribute to assessing established methods of text mining, machine learning, and natural language processing (NLP) regarding their value for genre classification. When tested empirically on corpora of different languages, periods, cultural contexts, and genres,

[2] Hempfer (2014, 409–10) points out that, on the other hand, it has been misinterpreted by several genre theorists trying to lead it back to require overall fundamental traits for genres.

[3] As an example of the application of the family resemblance concept, Hempfer (2014, 416–17) describes the history of the elegy, a genre that was originally only identifiable metrically and later by several other traits, amongst other things, intertextual references and motifs. He concludes: “The diachrony of the genre can best be represented as a synchronic network of relations, in which each individual text or epochal version of the genre is linked to other historical versions through common features. […] The genre identity, then, is not produced by a single trait but by the entirety of all relations among their historical versions” (2014, 419). For an application of the family resemblance concept to genre theory, see also Strube (1993, 21–25), who interprets a definition of the novella set up by Seidler in that way.

they expand knowledge about the extent to which the methods are sensitive to the kind of data. In addition, they also help to solve practical problems such as, for example, indexing large collections of texts.

Even so, it is important to question the composition of the categories examined. If the results are low classification rates, for instance, they cannot only indicate problems with the selection of features or classification parameters and methods, but may also be due to the underlying category not having the structure that is assumed. Alternative categorization methods can therefore be a good complement to the *classic* classification methods. It is not that only traditional classification methods have been used in digital genre stylistics. Thoughts about prototypicality, historical variability, and social embedding of literary genres have been expressed and tested as well. Underwood (2016; 2019, 34–67), for example, approaches genre via the history of reception, comparing competing definitions of detective fiction, science fiction, and the Gothic with statistical methods. Henny-Krahmer et al. (2018) use measures of similarity to determine the distances of novels belonging to different subgenres to predefined prototypes. In his contribution to this volume, Schröter (2023) proposes applying machine learning methods to reconstruct the historical change of *disordered* genres such as the German *Novelle*. Such experiments link key discussions of literary genre theory to digital genre stylistics. They contribute to a deeper confrontation of research results in the areas involved, potentially increasing their interest in each other and also challenging the findings achieved in one of the disciplines. For example: are the computational methods really suited for the analysis of genre concepts beyond classification, which is the prevalent method for categorization in computer science? The mentioned papers indicate that they are, but that the methods need to be adapted or creatively used to that end. On the other hand, what happens to the literary theoretical concepts of genre if they are tested for applicability in empirical digital studies? They may need to be modified, and furthermore, genre theory can receive new input from the results of digital genre stylistics. Of course, there is a similar relationship of exchange between literary genre theory and historical and empirical research in literary studies, but the difference to digital stylistics lies in the methodological apparatus. In the latter case, it is highly influenced by computational linguistics and computer science, disciplines with different scientific-theoretical backgrounds, which makes this kind of exchange equally promising and challenging.

This article aims to contribute to the described interface between literary genre theory, literary historical research, and digital genre stylistics. To that end, the concept of family resemblance is formalized and applied in a case study on subgenres of nineteenth-century Spanish-American novels. The choice of the corpus is due to the disciplinary orientation of this article's author and is explained further in the next section. The approach chosen to map the idea of family resemblance to a digital text analysis environment is network analysis. This way of formalization appears highly suitable, given the explanation of family resemblance formulated by Wittgenstein.

2. Subgenres of Spanish-American Nineteenth-Century Novels

Novels lend themselves well to an analysis based on the concept of family resemblance because the novel as a genre can hardly be defined uniquely and in formal terms. The necessary formal conditions are not sufficient to distinguish novels from other fictional narrative prose texts of considerable length (Hempfer 2014, 410), nor can additional formal or content-related criteria serve to capture all types of novels (Fludernik 2009, 627). As Fowler (1982, 112–13) points out, subgenres, while usually sharing the formal features of the genre, are often determined by subject matter or motifs and can be specified from a whole range of perspectives and on increasing levels of detail. He continues:

> Subgenres also threaten to defy subdivision in that they are extremely volatile. To determine the features of a subgenre is to trace a diachronic process of imitation, variation, innovation—in fact, to verge on source study. At the level of subgenre, innovation is life. Here, simple resemblance hardly produces a new work: at the very least there is elegant variation. And from time to time quite fresh subgenres will be invented, enlarging the kind in new directions altogether. It may be the conventionality of subgenres that strikes the beginner. But in reality they are the common means of renewal (Fowler 1982, 114).

As not even the genre itself is unified, the degree of variation applies even more to the subgenres of the novel (ibid, 118–126).

Early Spanish-American novels were influenced by European models, including types of the romantic novel such as sentimental novels and historical novels.[4] Often, the subgenres were adjusted to better serve the needs of expression of the Spanish-American authors. For example, in the nineteenth-century novels, political issues were often mixed up with a love story. The most prominent example is the novel *Amalia* (1851–1855) by the Argentine José Mármol, which tells the story of a group of resistance fighters against the dictatorship of Rosas[5] and of the protagonist's tragic love relationship (Dill 1999, 127). This novel is also an example of a special type of historical novel practiced in nineteenth-century Spanish-America because the contemporary

[4] Nineteenth-century novels are called *early* here because, in the Spanish-American colonies and countries, the genre only took off considerably in the course of that century (see Gálvez 1990, 15–25; Lindstrom 2004).

[5] Juan Manuel de Rosas (1793–1877) was a governor of the province of Buenos Aires who established a dictatorial system marked by repressive measures that lasted between 1829 and 1852 and that enforced a political and economic hegemony of Buenos Aires over the other provinces.

political events are presented as if they were historical. These *prospectively historical* novels (Molina 2011, 285–312) helped the authors to conceal that they were actually criticizing current political regimes and circumstances, and they served to inscribe events of the present or recent past into the history of the country. Then again, there were also conventional types of sentimental and historical novels, the latter, for example, dealing with the history of America's conquest, colonial times, and also European history (see, for instance, Read 1939).

Although more types of subgenres of the novel were practiced by Spanish-American writers in the nineteenth century and interpreted by literary historians later on, this article focuses on novels that have been primarily described as having a sentimental or historical theme. These descriptions may have been made either explicitly or implicitly, by contemporary authors or by modern critics. No distinction will be made here between different historical perspectives on the works, mainly because there is not much dissent for these subgenres and not enough data. Not all the novels were categorized by their authors, and especially lesser-known works have not been discussed extensively by literary historians. Novels from three countries, Argentina, Mexico, and Cuba, were chosen to cover different geographical and sociocultural areas of Spanish-America. The novels selected were first published between 1840 and 1910.[6]

The first aim of the family resemblance analysis is to find out how the two subgenres are organized internally: by looking at the network of similarities between the individual novels of a subgenre, do subgroups, i.e., *families*, emerge? Which traits hold them together? Can prospectively historical novels, for example, be distinguished from other types of historical novels? Or are there differences by country or over time? No preliminary assumption is made here as to the kind of connections that the family resemblance analysis might reveal. There could be diachronic shifts but also synchronic variations in the subgenres. In a second step, the sentimental and historical novels are analyzed together to see how the two subgenres are connected when no strict boundaries are applied, testing whether pure and mixed types become visible.

[6] Argentine, Mexican, and Cuban novels are understood as (1) novels written by authors having that nationality or having their center of life in these countries and (2) novels first published in the countries (that might have been written by authors with another nationality). During the nineteenth century, Argentina became independent in 1816, Mexico in 1821, and Cuba only in 1898. Even though Cuba was still a colony until the end of the century, the Cuban novel developed earlier and contributed to forming a national identity (Ferrer 2018, 11–19). Nevertheless, before independence, it is most convenient to call the authors of that country Cuban-Spanish.

3. Analysis

3.1 Corpus

The corpus of novels used for the analysis includes 83 works first published between 1840 and 1910. Of these, 40 are historical (*novela histórica*), and 43 sentimental novels (*novela sentimental*). 32 of the works are Argentinian, 35 are Mexican, and 16 are Cuban. They were written by 74 different authors, 9 of them female and 65 male. Only one work per author was chosen for each subgenre to prevent the authorial signal from interfering too much with the genre signal. However, if authors wrote in both subgenres, one novel of each subgenre is included. In the corpus, there are nine authors to whom this applies. Figure 1 shows the distribution of the novels in the corpus by decade and subgenre.

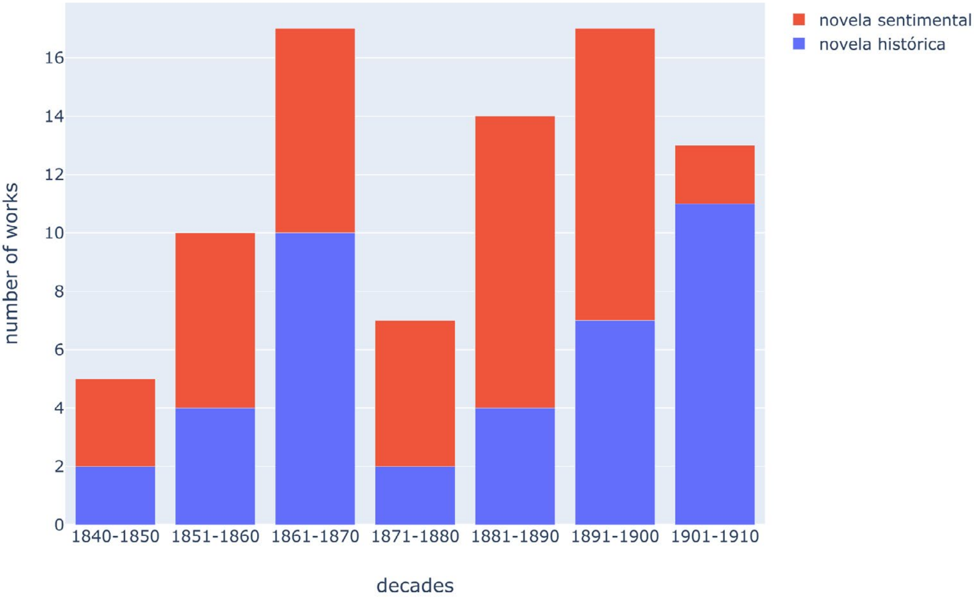


Fig. 1 Number of works in the corpus (Henny-Krahmer, CC BY).

The collection of texts used for this analysis is a subset of a larger corpus of 256 Spanish-American novels created in the project Computational Literary Genre Stylistics (CLiGS) at the University of Würzburg (Germany), which also includes novels of other subgenres. As explained below, the larger corpus is used as a basis for the generation of features.[7]

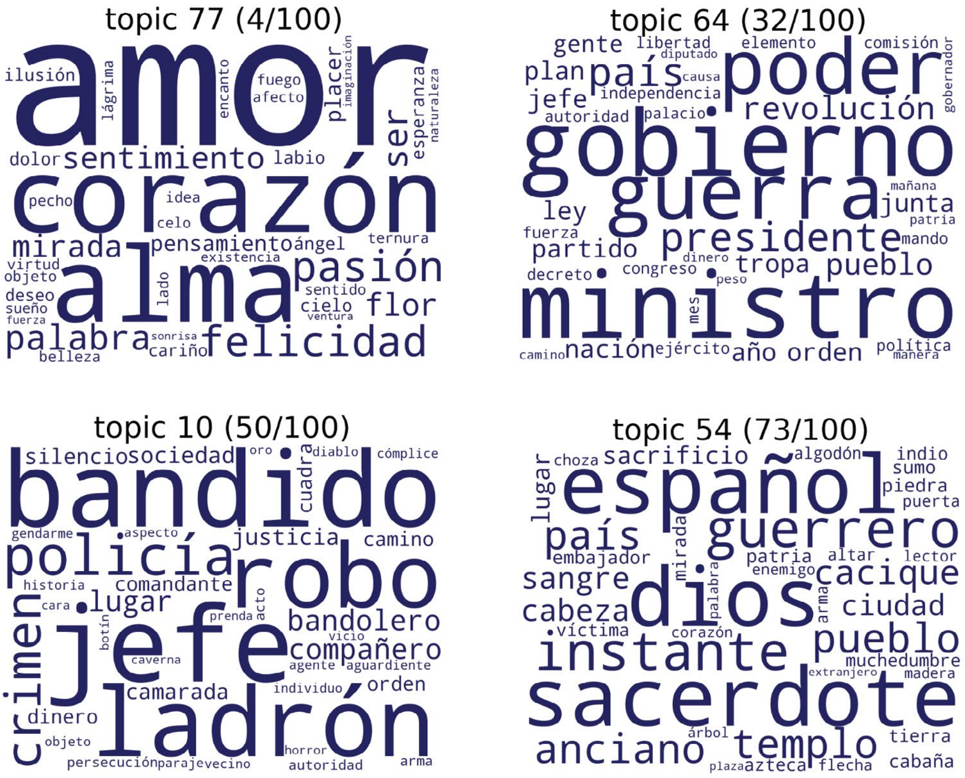
3.2 Features

It was decided to use topics as features for the analysis because they represent the themes developed in the novels and are, therefore, suitable to analyze subgenres primarily defined on a thematic basis. Topics have been tested successfully for the classification of novels by subgenre (Hettinger et al. 2016). The number of topics was set to 100, which is considered a medium degree of specification, given that the overall corpus contains 256 novels. The feature set was generated for the larger corpus with the goal of having more stable topics that represent the novel of the time in a better way than if they had been based on the smaller subcorpus. For the network analysis, only the features for the novels in the subcorpus are used. The topic model was built with the tool MALLET (McCallum 2002) and pre- and post-processed with *tmw* (Schöch and Schlör 2017). The texts were lemmatized with TreeTagger (Schmid 1995), using the Spanish parameter file, and keeping only nouns.[8]

In Figure 2, the top 40 words of four of the resulting 100 topics are visualized. They exemplify the range of themes covered in the novels. The first topic is about love and feelings: *amor* ('love'), *corazón* ('heart'), *alma* ('soul'), *pasión* ('passion'); the second topic is dominated by politics: *gobierno* ('government'), *ministro* ('minister'), *guerra* ('war'), *poder* ('power'); the third is about crime and banditry: *bandido* ('bandit'), *jefe* ('chief'), *ladrón* ('thief'), *robo* ('robbery'); and the fourth about religion and colonization: *sacerdote* ('priest'), *dios* ('god'), *español* ('Spaniard'), *guerrero* ('warrior'). The first number in parentheses indicates the rank of the topic by its probability in the whole corpus. The lower the number, the more important the topic is. Hence, the love topic

[7] The metadata of the corpus used for this analysis is available as "metadata.csv" in the folder "corpus_metadata" at https://github.com/hennyu/family_resemblance_dsrom19 and the metadata of the larger corpus as "metadata_full.csv". The whole corpus is called "Corpus de novelas hispanoamericanas del siglo XIX (conha19)" and is published at https://github.com/cligs/conha19. Besides the corpus metadata, the first GitHub repository mentioned also includes other scripts, data, and figures used in this article. Both links were accessed on May 26, 2021.

[8] In addition, a list of stop words was prepared based on the 50 most frequent nouns and adapted manually. To this, some more stop words were added manually after inspecting the results of the topic model (e.g., proper names or very general nouns). Before running the topic modeling, the texts were first lemmatized and then segmented into chunks with a length of 1000 tokens. Besides the number of topics, the topic model was created with 5,000 iterations and a hyperparameter optimization interval of 100. The feature matrices, both for the full and the reduced corpus, can be viewed on GitHub (see footnote 7).


topic 77 (4/100)

topic 64 (32/100)

topic 10 (50/100)

topic 54 (73/100)

Fig. 2 Examples for topics (Henny-Krahmer, CC BY).

(rank 4) is a very general one, the politics (rank 32) and crime (rank 50) topics are still common, and the colonization topic (rank 73) is more special.

3.3 Network Analysis

To create the network for the family resemblance analysis, first, the similarities between all the individual novels were calculated for the feature set using cosine similarity.[9] After that, the resulting textual similarities were mapped onto a network structure. The nodes in the network are constituted by the novels themselves. The network relationships (or *edges*) were determined using the three nearest neighbors of each text, which

[9] Cosine similarity measures the cosine of the angle between two text vectors, see Singhal (2001).

were selected from a ranking of the text similarities. The strength (or *weight*) of the edges was calculated by summing up the similarity values of the neighbors.[10]

In the overall similarity matrix, there are relationships between all the texts. Selecting only the connections of the three nearest neighbors reduces the network's complexity and makes the closest relationships more salient. This, in turn, enhances the interpretability of the network. The choice of three is arbitrary and could be varied. However, using more than one nearest neighbor makes the results of the network more stable, as Eder (2017, 56–60) has shown. He introduced the idea of visualizing nearest neighborships based on textual similarities in a network structure to make the results of stylometric cluster analysis more reliable. This technique is adapted here with a different aim: to formalize the family resemblance concept for genre analysis.

In addition to creating the basic network structure, community detection was used to explore *families* of novels in the network. *Communities* are sets of nodes in a network that are more densely connected to each other than to nodes outside (Javed et al. 2018, 87–90). Different algorithms for the detection of network communities exist. It was decided to use the Louvain modularity algorithm (Blondel et al. 2008) because it is a suitable algorithm for detecting disjoint communities in static networks, is comparatively efficient, and has been implemented in Python, which is used to create and visualize the network here.[11]

Reflecting on how the concept of family resemblance is formulated by Wittgenstein and in literary genre theory, on the one hand, and how it is implemented here, on the other, the following observations can be made. First, using similarity relationships between the novels based on feature distributions means that it is not the presence or absence of a trait which determines the connection between members of a family and the difference to other families, but the numerical strengths of the features in combination. This transfers the idea of partial and overlapping similarities to a quantitative approach.[12] Second, when distinct communities are calculated and interpreted as families, the boundaries of the categories are sharpened retroactively. This is an advantage that balances out the looseness of the original family resemblance concept. Nevertheless, there is a significant difference between the families based on communities and conventional classes because the former emerge from a network of similarities and not from shared common features. The communities mark a boundary between one group

[10] Because the closest neighborship depends on the perspective, it was calculated for each node. When two nodes are mutually closest, the strength of the edge increases.

[11] See https://github.com/taynaud/python-louvain (accessed April 15, 2021).

[12] Of course, zero values are also possible in the feature matrices and could be interpreted as *absent*, but it would not be proportionate to consider all values that are greater than zero as *present*. A possibility to model the features in a different way would be to define a threshold value and convert all values below it to zero and all values above it to one to get a binary distinction. Still, good reasons would have to be given for the value at which to set the threshold.

of dense relationships and another, they cut off the family at a certain point, but they do not lever out the basic idea of family resemblance.[13]

3.4 Results

With the approach outlined in the previous section, three kinds of networks were produced, two for the individual subgenres and one for the two subgenres combined, as shown in Table 1.[14]

Table 1 Overview of the family resemblance networks produced.

shortcut	subgenre(s)	number of novels	number of communities (*families*)
HIST	historical novels	40	6
SENT	sentimental novels	43	6
HIST-SENT	historical and sentimental novels	83	8

The last column indicates how many *families,* that is, clusters based on the communities in the network, were produced. The number of clusters is identical for both historical and sentimental novels when they are analyzed separately. Given that the number of novels doubles when the two subgenres are combined, the number of resulting clusters does not grow proportionally, indicating that there is an overlap between the subgenres. Due to the lack of space, only some of the results can be discussed in detail in this article, and the discussion focuses on historical novels.[15] Figure 3 shows the first network for historical novels and topics (HIST). The communities detected are indicated by the different colors of the nodes.

An important question for the interpretation of the network is which kinds of novels constitute the different families. Before looking at different clusters in detail,

[13] So far, many decisions have been taken to formalize the family resemblance concept for the case study in this article. It becomes clear that variants of this approach are possible. For example, the similarity measure used, the number of nearest neighbors considered, the way to determine the strength of the edges, and the kind of community detection algorithm could be varied. As with feature-based categorization in general, also here, the selection of the features and their modeling and parametrization are subject to choice. Further empirical studies and serial analyses are needed to test the effects of such variation on the results.

[14] The script calling the various functions of the network analysis for the different setups is available at https://github.com/hennyu/family_resemblance_dsrom19/blob/main/analysis/run_scripts.py (accessed May 26, 2021).

[15] The overall results can be inspected on GitHub, though (see footnote 7), where also an analysis with the most frequent words (MFW) instead of topics is included.

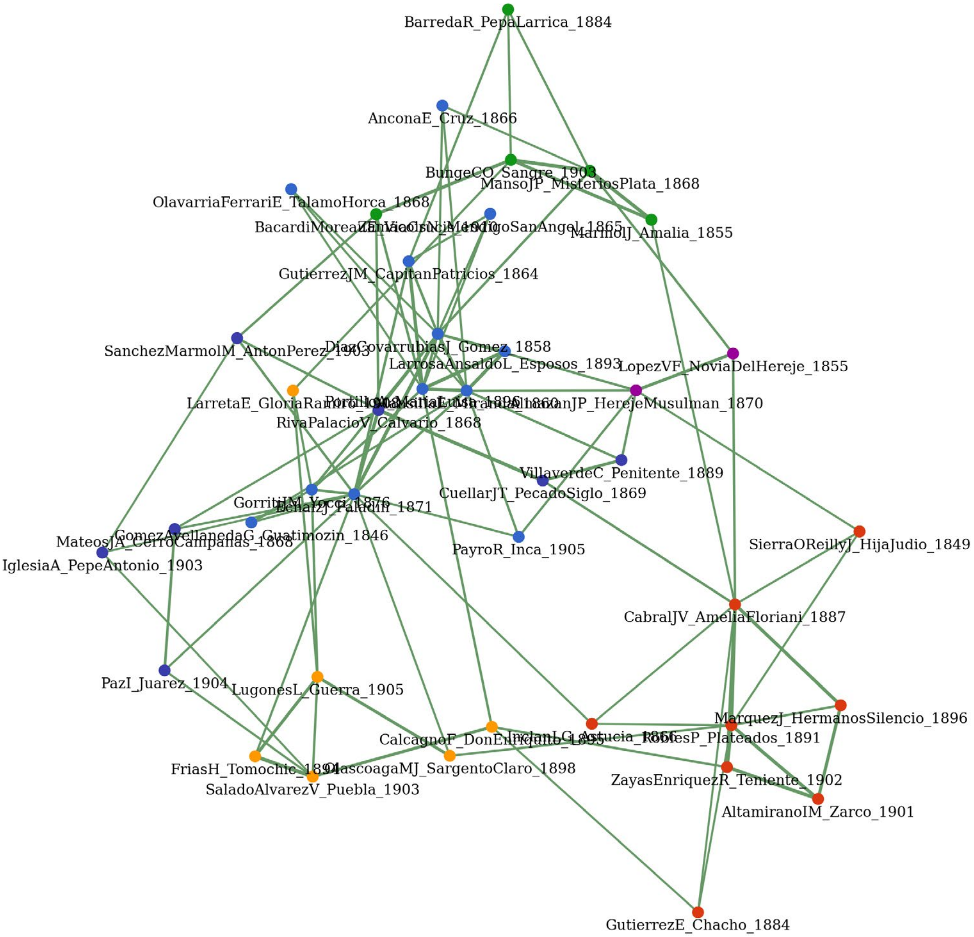

BarredaR_PepaLarrica_1884
AnconaE_Cruz_1866
BungeCO_Sangre_1903
MansoJP_MisteriosPlata_1868
OlavarriaFerrariE_TalamoHorca_1868
MarmolJ_Amalia_1855
GutierrezJM_CapitanPatricios_1864
SanchezMarmolM_AntonPerez_1903
DiazCovarrubiasJ_Gomez_1858
LarrosaAnsaldoL_Esposos_1893
LopezVF_NoviaDelHereje_1855
HerejeMusulman_1870
RivaPalacioV_Calvario_1868
VillaverdeC_Penitente_1889
CuellarJT_PecadoSiglo_1869
GorritiJM_Yocci_1876
EchaizJ_Paladin_1871
GomezAvellanedaG_Guatimozin_1846
MateosJA_CerroCampanas_1868
IglesiaA_PepeAntonio_1903
PayroR_Inca_1905
SierraOReillyJ_HijaJudio_1849
CabralJV_AmeliaFloriani_1887
PazI_Juarez_1904
LugonesL_Guerra_1905
MarquezJ_HermanosSilencio_1896
RoblesP_Plateados_1891
FriasH_Tomochic_1894
OlascoagaMJ_SargentoClaro_1898
SaladoAlvarezV_Puebla_1903
ZayasEnriquezR_Teniente_1902
AltamiranoIM_Zarco_1901
GutierrezE_Chacho_1884

Fig. 3 Network of historical novels based on topics (HIST), (Henny-Krahmer, CC BY).

an overview of the cluster sizes was generated, and the possible influence of some text-external and -internal factors on the clusters was calculated, as displayed in Figures 4 and 5.

Four of the resulting six clusters are evenly sized, with 8 novels each; the other two are smaller. Cuban novels are only contained in clusters 1, 2, 3, and 5. Clusters 1, 4, and 5 are dominated by Mexican novels, and clusters 2 and 3 by Argentine novels. Cluster 3 is a Argentine–Cuban cluster, and cluster 5 a Mexican–Cuban cluster. Even

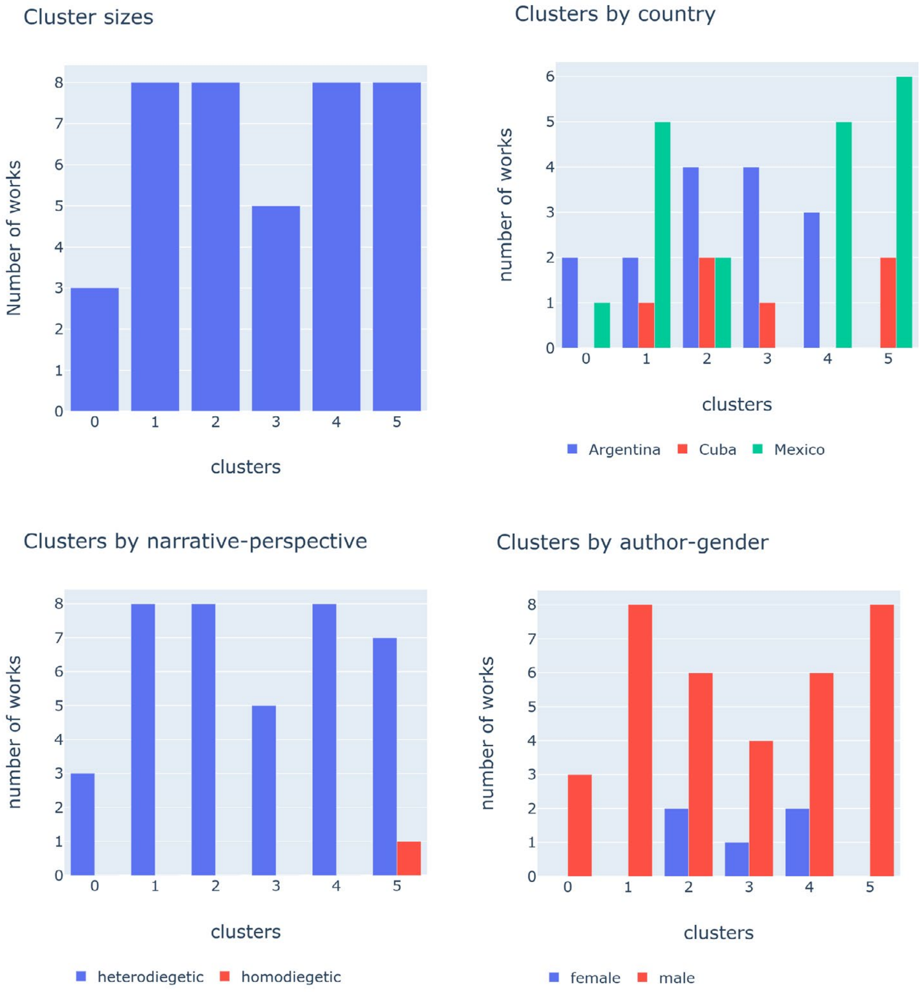


Fig. 4 Overview of cluster metadata in the network HIST (Henny-Krahmer, CC BY).

if there are some tendencies regarding the distribution of novels by country in the different clusters, there is no cluster consisting only of novels from one country. It should also be kept in mind that the overall number of novels in the individual clusters is quite small. The narrative perspective is not significant for the historical novels because there is only one novel with a homodiegetic narrator, the others all have a heterodiegetic narrator. The five historical novels written by female authors are distributed over the three clusters 2, 3, and 4, so there is no clear female cluster. Regarding the distribution

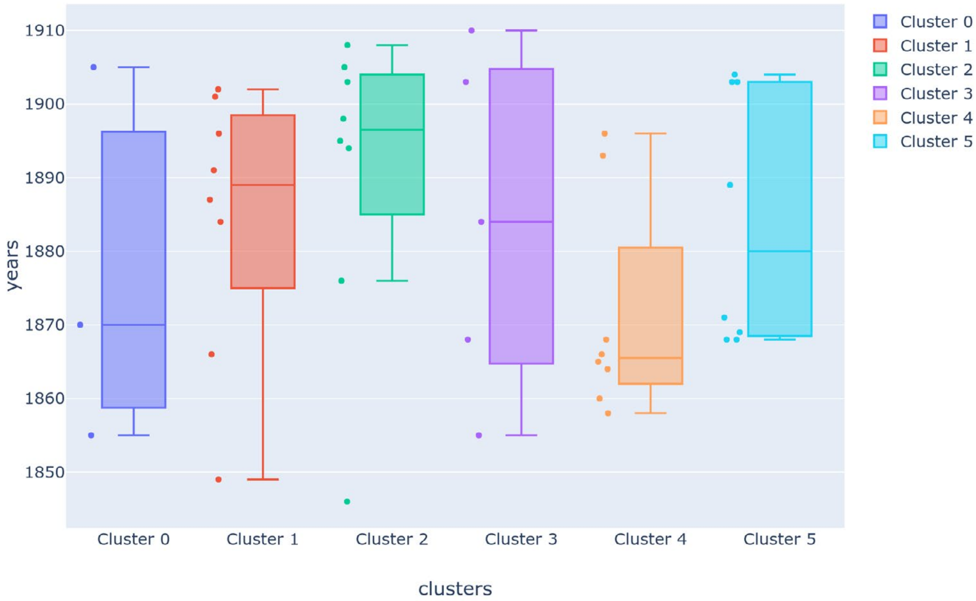


Fig. 5 Clusters by year in the network HIST (Henny-Krahmer, CC BY).

of the novels over the years, there is also much overlap, as all the clusters have earlier and later novels. Apart from one outlier, cluster 2 is rather late, and cluster 4 is mostly filled with earlier works.

Looking at one cluster in detail, it is possible to retrace the family resemblance relationships. In Table 2, the novels contained in cluster 3 are listed together with their nearest neighbors (N-1, N-2, N-3), including the weight of the edge to the respective neighbor. The strongest relationship exists between *La novela de la sangre* (Arg., 1903) by Carlos Octavio Bunge and *Los misterios del Plata* (Arg., 1868) by Juana Manso de Noronha because they are mutually closest to each other. Other bilateral nearest neighborships between novels in the cluster are highlighted in lighter orange. The novel *Pepa Larrica* (Arg., 1884) by Rafael Barreda has two nearest neighbors in the cluster, but the relationships are only unilateral. Boxes highlighted in gray show which nearest neighbors are outside of the current cluster. It becomes clear that some novels are central members of the family while others are rather distant relatives.

The topic distributions for the five historical novels are visualized in Figure 6 below to see what topics are decisive for the relationships in this cluster. The axis on the top shows the absolute value that the topic achieved in each novel, and the axis to the

Table 2 Nearest neighbors in cluster 3 of the network HIST.

idno	author	title	N-1		N-2		N-3	
nh0017	Mármol	Amalia	Misterios	1.4	Sangre	1.2	Cl 1	0.3
nh0081	Bunge	La novela de la sangre	Misterios	1.5	Crucis	1.2	Amalia	1.2
nh0094	Manso	Los misterios del Plata	Sangre	1.5	Amalia	1.4	Cl 4	0.6
nh0160	Barreda	Pepa Larrica	Cl 4	0.4	Misterios	0.4	Sangre	0.4
nh0166	Bacardí Moreau	Vía Crucis	Sangre	1.2	Cl 4	0.6	Cl 5	0.6

left shows the individual 100 topics.[16] In addition to the lines for the five novels in the cluster, a black dashed line indicating the mean topic values for all the historical novels in the network is added. The topics are ordered by importance in the whole corpus of 256 novels from top to bottom so that more general topics are at the top and more special topics are further down. Some topics of interest are labeled, the black ones being particularly important for this cluster and the red ones less important when compared to all the historical novels in the corpus.

The family approach is visible because not all decisive topics are equally relevant for the individual novels in the cluster. For example, the topics *sacerdote-dios-español* ('priest-god-Spaniard') and *fortaleza-batería-plaza* ('fortress-battery-square') are underrepresented in the whole cluster, but *amor-corazón-alma* ('love-heart-soul') and *soldado-fuego-columna* ('soldier-fire-column') are only partly less relevant. The first corresponds to the mean for the novel *Amalia*, and the second reaches almost the mean for *Vía Crucis*. Topics that are overrepresented in several novels in the cluster are *voz-palabra-brazo* ('voice-word-arm'), *idea-espíritu-instante* ('idea-spirit-moment'), *pueblo-ley-país* ('people-law-country'), *calle-puerta-voz* ('street-door-voice'), *agua-cuerpo-sangre* ('water-body-blood'), *gobierno-ministro-guerra* ('government-minister-war'), *puerta-espíritu-cabeza* ('door-spirit-head') and *cabeza-rosa-asesino* ('head-rose-assassin'). They stand for the general characteristics of the family: historical novels that are not mixed with love stories so much, not focused on military actions and not about the conquest or colonial history, but about political ideas and conditions, (inter)personal contacts and states, voices, words, and bodies. However, as specific topic values are not necessary conditions, some of the novels have their own special topics. The topic *mar-buque-puerto* ('sea-boat-harborur') is specific for *Los misterios del Plata*, *negro-esclavo-amo* ('black-slave-lord') for *Vía Crucis*, the only Cuban novel in this cluster, and *capitán-voz-revolución* ('captain-voice-revolution') for *Pepa Larrica*.

[16] In a strict sense, the topics are categories and not numerical values and should be visualized as bars rather than lines. The line plot was chosen here because it facilitates seeing the differences between the data series.

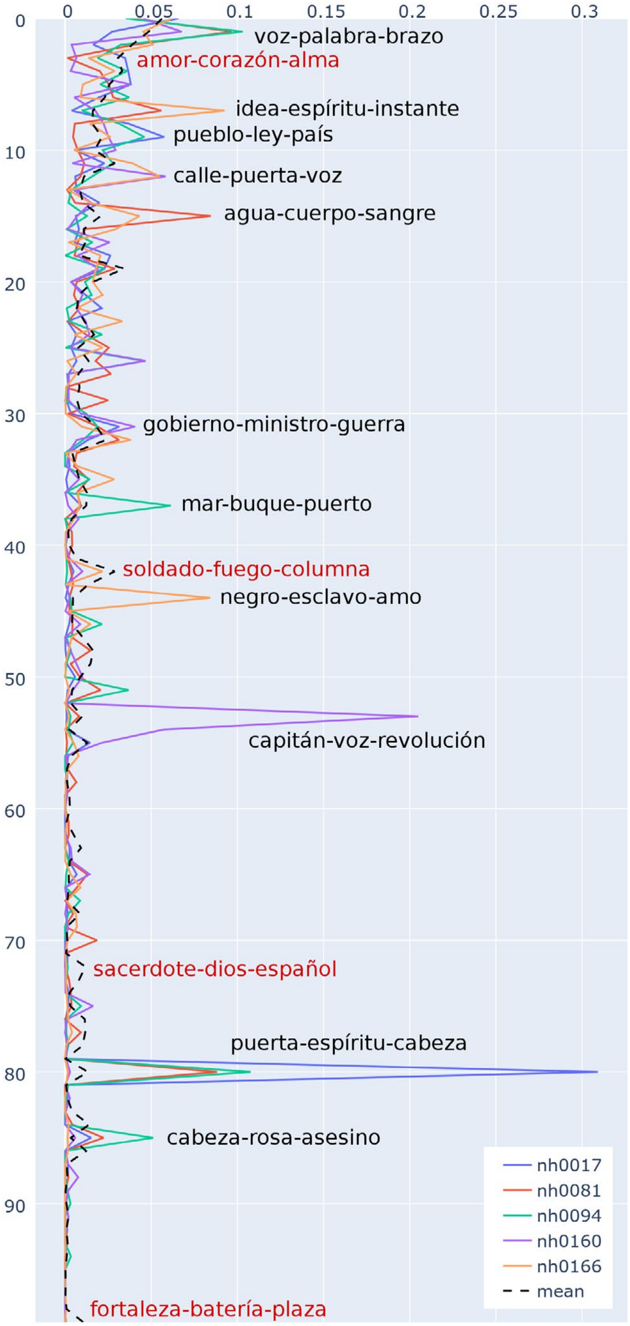


Fig. 6 Topic scores for cluster 3 in the network HIST (Henny-Krahmer, CC BY).

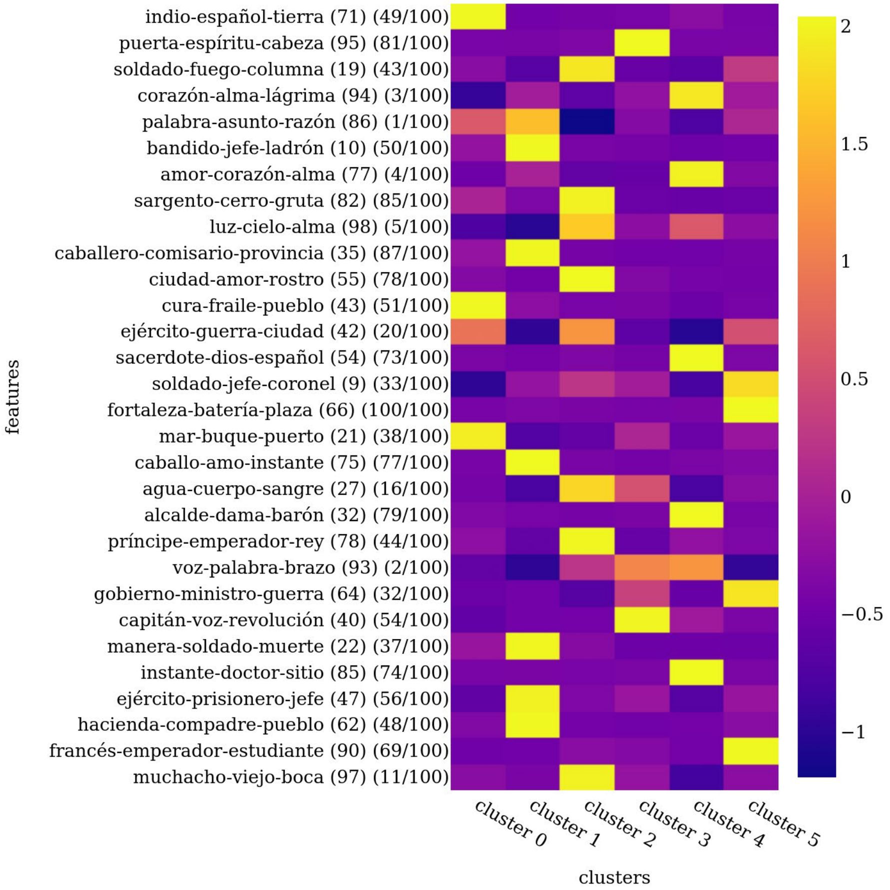


Fig. 7 Top distinctive topics in the clusters of the network HIST (Henny-Krahmer, CC BY).

A more general overview of the topics that are distinctive for the different clusters in the network of historical novels is given in Figure 7. In the heatmap, the yellower the boxes, the more important the topics are for the cluster, and the bluer, the less important they are. The distinctiveness was calculated by normalizing the topic values to z-scores (Oakes 2003, 7–8). Here, only the top 30 most distinctive topics are shown. The values in parentheses at the end of the topic labels indicate the ranks of the topics in the whole corpus (by probability), so the topic *voz-palabra-brazo* ('voice-word-arm') with rank 2, for example, is much more general than *fortaleza-batería-plaza* ('fortress-battery-square') with rank 100.

The distinctive topics of cluster 3 that were already discussed can be recognized in the heatmap. The smallest cluster 0 seems to be about the conquest and colonial history, as the most distinctive topics are *indio-español-tierra* ('Indian-Spaniard-land'), *cura-fraile-pueblo* ('priest-friar-village'), and *mar-buque-puerto* ('sea-ship-harbor'). In cluster 1, topics about military campaigns and rural life prevail, making one think about internal struggle, bandits and *gauchos*: *palabra-asunto-razón* ('word-matter-reason'), *bandido-jefe-ladrón* ('bandit-chief-thief'), *caballero-comisario-provincia* ('gentleman-inspector-province'), *caballo-amo-instante* ('horse-lord-moment'), *manera-soldado-muerte* ('manner-soldier-dead'), *ejército-prisionero-jefe* ('army-prisoner-chief'), *hacienda-compadre-pueblo* ('estate-godfather-village'). Cluster 2 is not so easy to interpret. It is about military action (*soldado-fuego-columna* ('soldier-fire-column'), *sargento-cerro-gruta* ('sargeant-hill-grot'), *ejército-guerra-ciudad* ('army-war-city'), but there are other, individual topics. Cluster 4 is clearly romantic with colonial and aristocratic elements: *corazón-alma-lágrima* ('heart-soul-tear'), *amor-corazón-alma* ('love-heart-soul'), *sacerdote-dios-español* ('priest-god-Spaniard'), *alcalde-dama-barón* ('mayor-lady-baron'), *instante-doctor-sitio* ('moment-doctor-place'). This fits well with the observation that it contains mostly earlier novels published during the main phase of the Romantic current in Spanish-America in the first half of the nineteenth century. The last cluster is politico-historical with the top topics *soldado-jefe-coronel* ('soldier-chief-colonel'), *fortaleza-batería-plaza* ('fortress-battery-square'), *gobierno-ministro-guerra* ('government-minister-war'), and *francés-emperador-estudiante* ('French-emperor-student').

For reasons of space, the results for the network of sentimental novels and the one containing both types of subgenres are only summarized briefly here. Regarding the metadata, the cluster sizes vary more for the sentimental novels, with the biggest cluster having 14 novels and the smallest one only two. All the clusters are mixed by country. Among the sentimental novels, there are more with autodiegetic and homodiegetic narrators, and the narrative perspective has an influence on the results. The smallest cluster, for instance, consists solely of autodiegetic texts featuring topics related to inner life and landscape. For the sentimental novels, there are also clearer tendencies of topic changes over time, as Figure 8 shows. The early cluster 0 is romantic with letters, dance, aristocracy, and much emotionality. The three later clusters are the ones dominated by interiorization, and the mid-century cluster 5 is worldly about food, marriage, business, and money.

The most important point to note when both subgenres are analyzed together is that the subgenres are not neatly sorted into the different families. As can be seen in Figure 9, there are clusters dominated by one subgenre—clusters 1, 3, and 6 by sentimental novels and clusters 2, 4, and 7 by historical novels—but there is no cluster containing only novels of one subgenre. The clusters 0 and 5 are entirely mixed. This indicates that it could be difficult to classify these novels by the primary type to which they have been assigned by their authors and by critics.

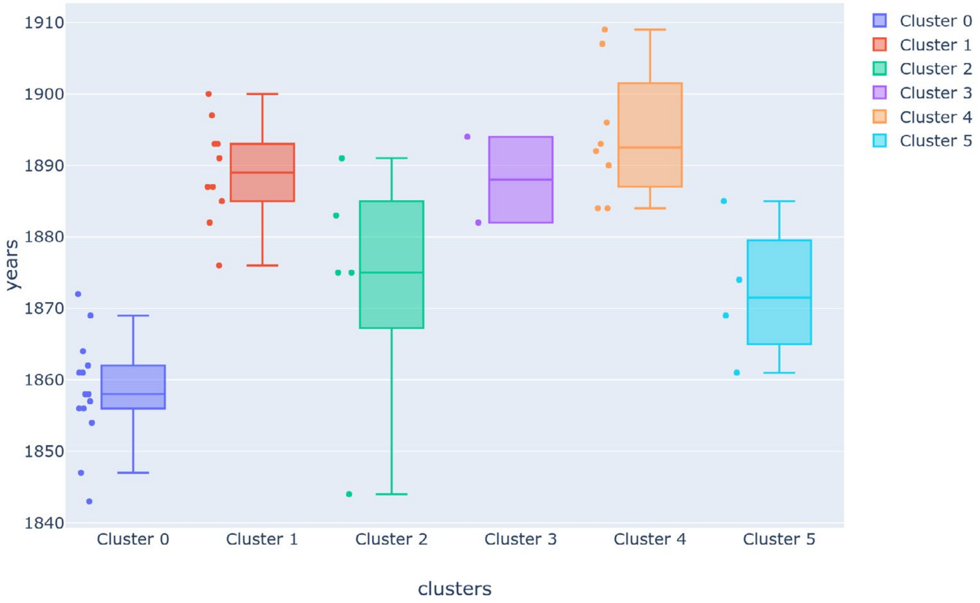


Fig. 8 Clusters by year in the network SENT (Henny-Krahmer, CC BY).

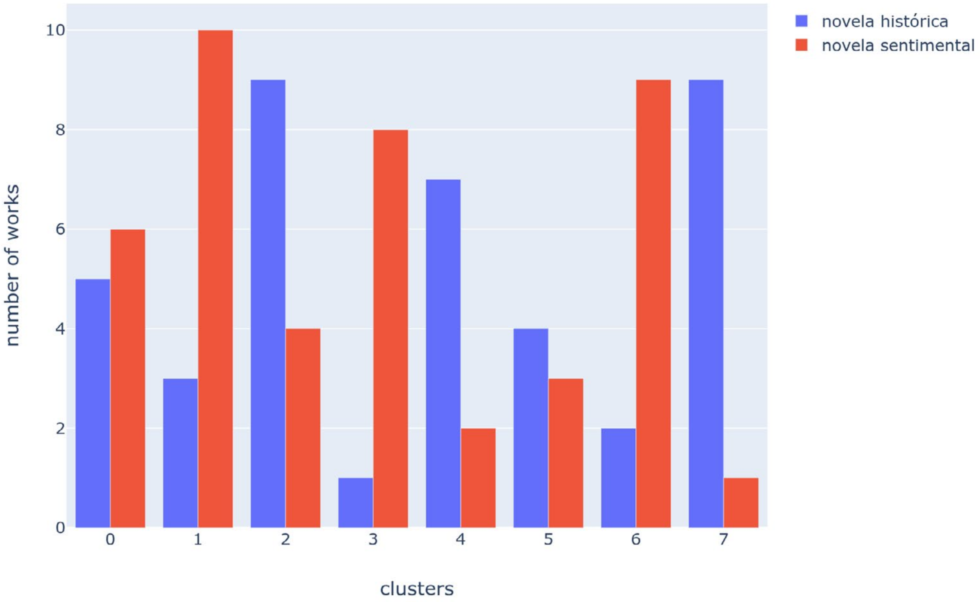


Fig. 9 Clusters by subgenre in the combined network (Henny-Krahmer, CC BY).

In the combined network, the cluster sizes vary moderately from seven to 13 novels. Here again, there is no clear tendency for countries. Different narrative perspectives are not concentrated in single clusters, so this aspect, which was observed for the sentimental novels alone, disappears when they are analyzed in the more general setup. Regarding the distribution by years, cluster 1 is early, clusters 3, 4, and 6 are late, and the others are mixed. The topics that are distinctive for the different families reflect the relative purity or mixture of subgenres as well as the preferences of the early versus the late nineteenth century.

4. Conclusion

Here a proposal was made for how the concept of family resemblance, which was introduced into genre theory in the 1960s and also argued for by several genre theorists recently, can be applied in a digital genre stylistics approach. With the analysis of topics in nineteenth-century Spanish-American historical and sentimental novels, the proposal was empirically tested in a network-based approach. If one looks at the current strategies to categorize genres in digital stylistics, the majority focus on classificatory groupings based on the assumption of features that are common to all members of a class. However, there are also alternative ways to analyze genres in digital stylistics, some of which have explicitly addressed genre theoretical questions, while others have not. In particular, stylometric network analyses implicitly contain the idea of overlapping similarities and unsharp boundaries, which is characteristic of the family resemblance approach. In this article these two scenarios were brought together. With the chosen approach of comparing the feature distributions of the novels and organizing the resulting network of similarities into communities interpreted as families, the original idea of family resemblance is adapted for digital analysis. First, rather than the presence or absence of individual textual features, the degree of their joint presence is decisive. Second, communities or clusters found in the similarity network constitute a way to delimit the families retroactively without changing the underlying concept of intertwining shared characteristics of individual members of the groups.

For the Argentine, Mexican, and Cuban historical and sentimental novels, the analysis confirmed that there are subtypes of the subgenres that have already been described in literary historical approaches. Such subtypes are, for example, a novel with a historical setting and a sentimental plot, or a historical novel focusing on contemporary political conditions, or a historical novel about events of colonial times. In addition, influences of the narrative perspective on subtypes of the sentimental novel became visible. Analyzing both types of subgenres together resulted in mixed groups and some that are dominated by one subgenre. While the country in which the novels were published or the authors' nationalities do not have a clear impact on the resulting families of novels, the year of publication does, in some cases, when the preferred and avoided topics reflect the literary development in the nineteenth century. All in all, the results show that features common to all novels of a subgenre cannot be expected and that the factors that influence the subgroups or families of subgenres are diverse. There is not one decisive factor, each family has its own traits that hold it together, and inside of each one, there are additional individual traits as well as connections to other families.

To conclude, the algorithm producing the family resemblance network and the resulting data offer an empirical ground on which literary historians can look for sense in genre historical terms. It does not say anything about the historico-cultural and communicative relevance of the connections, but it might reveal previously unrecognized textual similarities in addition to confirming known ones on a broader textual basis. By not presupposing strict uniformity inside and strict boundaries between the genre categories, it comes closer to the open genres that the novel and its subgenres form in terms of theme and style.

ORCID®

Ulrike Henny-Krahmer https://orcid.org/0000-0003-2852-065X

References

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 10. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Calvo Tello, José. 2018. "Genre Classification in Novels: A Hard Task for Humans and Machines?" Paper presented at *EADH 2018: Data in Digital Humanities, Galway, 7–9 December 2018*. Galway: National University of Ireland. https://eadh2018.exordo.com/programme/presentation/82 (Accessed April 15, 2021).

Calvo Tello, José, Daniel Schlör, Ulrike Henny, and Christof Schöch. 2017. "Neutralizing the Authorial Signal in Delta by Penalization: Stylometric Clustering of Genre in Spanish Novels." In *Digital Humanities 2017. Conference Abstracts, Montréal, Canada, 8–11 August 2017*, 181–184. Montréal: McGill University and Université de Montréal.

Dill, Hans-Otto. 1999. *Geschichte der lateinamerikanischen Literatur im Überblick*. Stuttgart: Reclam.

Eder, Maciej. 2017. "Visualization in Stylometry. Cluster Analysis Using Networks." *Digital Scholarship in the Humanities* 32 (1). https://doi.org/10.1093/llc/fqv061.

Ferrer, José Luis. 2018. *La invención de Cuba: Novela y nación (1837–1846)*. Madrid: Editorial Verbum.

Fishelov, David. 1993. *Metaphors of Genre: The Role of Analogies in Genre Theory*. University Park, PA: Pennsylvania State University Press.

Fludernik, Monika. 2009. "Roman." In *Handbuch der literarischen Gattungen*, edited by Dieter Lamping and Sandra Poppe, 627–45. Stuttgart: Kröner.

Fowler, Alastair. 1982. *Kinds of Literature. An Introduction to the Theory of Genres and Modes*. Oxford: Clarendon Press.

Fricke, Harald. 2010. "Definitionen und Begriffsformen." In *Handbuch Gattungstheorie*, edited by Rüdiger Zymner, 7–10. Stuttgart, Weimar: J.B. Metzler.

Gálvez, Marina. 1990. *La novela hispanoamericana (hasta 1940)*. Madrid: Taurus.

Gianitsos, Efhimios Tim, Thomas J. Bolt, Pramit Chaudhuri, and Joseph P. Dexter. 2019. "Stylometric Classification of Ancient Greek Literary Texts by Genre." In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Minneapolis, MN, USA, 7 June, 2019*, 52–60. Minneapolis: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-2507.

Hempfer, Klaus W. 2014. "Some Aspects of a Theory of Genre." In *Linguistics and Literary Studies: Interfaces, Encounters, Transfers*, edited by Monika Fludernik and Daniel Jacob, 405–22. Berlin: De Gruyter.

Henny-Krahmer, Ulrike. 2018. "Exploration of Sentiments and Genre." In *Digital Humanities 2018. Puentes–Bridges. Book of Abstracts, Mexico City, 26–29 June 2018*, 399–403. Mexico City: Red de Humanidades Digitales.

Henny-Krahmer, Ulrike, Katrin Betz, Daniel Schlör, and Andreas Hotho. 2018. "Alternative Gattungstheorien. Das Prototypenmodell am Beispiel hispanoamerikanischer Romane." In *DHd 2018. Kritik der digitalen Vernunft. Konferenzabstracts, Köln, 26 February–2 March 2018*, 105–12. Köln: Universität zu Köln.

Hettinger, Lena, Isabella Reger, Fotis Jannidis, and Andreas Hotho. 2016. "Classification of Literary Subgenres." In *DHd2016. Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Konferenzabstracts, Leipzig, 7–12 March 2016*, 160–64. Duisburg: nisaba verlag.

Javed, Muhammad Aqib, Muhammad Shahzad Younis, Siddique Latif, and Adeel Baig. 2018. "Community Detection in Networks: A Multidisciplinary Review." *Journal of Network and Computer Applications* 108: 87–111. https://doi.org/10.1016/j.jnca.2018.02.011.

Lindstrom, Naomi. 2004. *Early Spanish American Narrative*. Austin: University of Texas Press.

McCallum, Andrew Kachites. 2002. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu (Accessed April 15, 2021).

Molina, Hebe Beatriz. 2011. *Como crecen los hongos. La novela argentina entre 1838 y 1872*. Buenos Aires: Teseo.

Müller, Ralph. 2010. "Kategorisieren." In *Handbuch Gattungstheorie*, edited by Rüdiger Zymner, 21–23. Stuttgart, Weimar: J. B. Metzler.

Neumann, Birgit, and Ansgar Nünning. 2007. "Einleitung: Probleme, Aufgaben und Perspektiven der Gattungstheorie und Gattungsgeschichte." In *Gattungstheorie und Gattungsgeschichte*, edited by Marion Gymnich, Birgit Neumann, and Ansgar Nünning, 1–28. Trier: WVT.

Oakes, Michael P. 2003. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Read, John Lloyd. 1939. *The Mexican Historical Novel*. New York: Instituto de las Españas en los Estados Unidos.

Schmid, Helmut. 1995. "Probabilistic Part-of-Speech Tagging Using Decision Trees." *Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK, 1994*. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf. (Accessed April 15, 2021).

Schöch, Christof. 2017. "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html (Accessed April 15, 2021).

Schöch, Christof. 2013. "Fine-tuning Our Stylometric Tools: Investigating Authorship and Genre in French Classical Theater." In *Digital Humanities 2013. Conference Abstracts, Lincoln, USA, 16–19 July 2013*, 383–86. Lincoln: Center for Digital Research in the Humanities.

Schöch, Christof, and Daniel Schlör. 2017. "tmw – Topic Modeling Workflow." *GitHub*. https://github.com/cligs/tmw (Accessed April 15, 2021).

Schöch, Christof, Ulrike Henny, José Calvo Tello, Daniel Schlör, and Stefanie Popp. 2016. "Topic, Genre, Text. Topics im Textverlauf von Untergattungen des spanischen und hispanoamerikanischen Romans (1880–1930)." In *DHd2016. Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Konferenzabstracts, Leipzig, 7–12 March 2016*, 235–39. Duisburg: nisaba verlag.

Schröter, Julian. 2024. "Machine-Learning as a Measure of the Conceptual Looseness of Disordered Genres: Studies on German *Novellen*." In *Digital Stylistics in Romance Studies and Beyond*, edited by Robert Hesselbach, José Calvo Tello, Ulrike Henny-Krahmer, Christof Schöch, and Daniel Schlör, 173–195. Heidelberg: heiUP.

Singhal, Amit. 2001. "Modern Information Retrieval: A Brief Overview." In *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24 (4): 35–43. http://sites.computer.org/debull/A01dec/singhal.ps (Accessed April 15, 2021).

Strube, Werner. 1993. *Analytische Philosophie der Literaturwissenschaft. Untersuchungen zur literaturwissenschaftlichen Definition, Klassifikation, Interpretation und Textbewertung*. Paderborn: Schöningh.

Tophinke, Doris. 1997. "Zum Problem der Gattungsgrenze – Möglichkeiten einer prototypentheoretischen Lösung." In *Gattungen mittelalterlicher Schriftlichkeit*, edited by Barbara Frank, Thomas Haye, and Doris Tophinke, 161–82. Tübingen: Narr.

Underwood, Ted. 2015. *Understanding Genre in a Collection of a Million Volumes*. White Paper Report. Urbana,-Champaign: University of Illinois. http://dx.doi.org/10.17613/M6W07V.

Underwood, Ted. 2016. "The Life Cycles of Genres." *Journal of Cultural Analytics*. 2 (2) https://doi.org/10.22148/16.005.

Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change*. Chicago: The University of Chicago Press.

Wittgenstein, Ludwig. 2009. *Philosophical Investigations*. Translated by G. E. M. Anscombe, P. M. S. Hacker, and Joachim Schulte. Edited by P. M. S. Hacker and Joachim Schulte. New York: Wiley.

Machine Learning as a Measure of the Conceptual Looseness of Disordered Genres
Studies on German *Novellen*

Julian Schröter

Abstract This essay discusses how procedures of computational and literary genre stylistics can be built and implemented in order to reconstruct the ways in which disordered, heterogeneous, and historically discontinuous genres undergo historical change. The discussion consists of three parts. In the first part, I will introduce a distinction between classificatory and aesthetic interest and show that genre stylistics should in general be based on aesthetic interest. This is an argument against the common claim that genre concepts should be defined prior to historiographical work. The second part outlines the specific historical situation of the German *Novelle* as well as the basis of an aesthetic historiography of this disordered genre. The third part gives an operationalization of one fundamental task in computational and literary genre stylistics—namely, the psychological question of the extent to which historical readers were able to gain an understanding of different genres, based on texts that I grouped according to genre labels. This issue is processed here as a supervised learning task. With this method, low accuracy scores in classification tasks are usually interpreted as methodic flaws, and optimal accuracy scores are regarded as the criterion for adequate modeling. This essay argues, however, that if the question concerns the aesthetic history of disordered genres, then any adequate classification task based on supervised machine learning is expected to yield low accuracy scores. In the next step, the criterion of adequateness must therefore be revised so that low accuracy scores can be attributed to a specific genre itself and not to inappropriate selection of features, problematic parametrization, and other aspects of modeling. Finally, accuracy scores are not to be optimized, but they can be interpreted, in the case of disordered genres. Machine learning tasks can be integrated into a psychological framework so that accuracy scores can be interpreted as a measure of the semantic looseness of the concept of a specific genre within historical literary communities.

Schröter, Julian: Machine Learning as a Measure of the Conceptual Looseness of Disordered Genres. Studies on German *Novellen*, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 173–195. DOI: https://doi.org/10.17885/heiup.1157.c19371

Keywords supervised machine learning, German *Novelle*, computational and literary genre stylistics

1. Introduction

In terms of conceptual extension, genre can be thought of as a group or cluster of texts that can be separated from other clusters. Clustering texts by means of their features is one major task of computational stylistics (Schöch 2017). Herrmann et. al. (2015, 46) emphasize that "style can be associated with categories such as genre, epoch, author, and many more." For quantitative text analysis, it is not the only option but, I believe, essential that style be investigated as *style of one or several categories* such as author, period, or genre (Jockers 2013; Evert et. al. 2017). The practice of investigating the style of a certain genre in the realm of computational and literary genre stylistics (CLGS) can be characterized by a common premise: we presume that genres behave like logical classes insofar as we presume that texts of the same genre share a set of characteristic and specific features. This set yields the style of that genre and can be described "in terms of a quantitative profiling of formal features" (Herrmann et al. 2015, 46). However, for the German *Novelle*, recent research indicates quite a different situation: over the last decades, studies have increasingly questioned an understanding of the *Novelle* as a genre in literary studies that had largely escaped critical scrutiny.[1] Earlier studies of the nineteenth and early twentieth century (Borcherdt 1926; Klein 1960; Himmel 1963; Kunz 1970) as well as recently published introductions to the genre (Freund 2009; Rath 2008; Füllmann 2010; Meier and Vrckovski 2014) claim that the German *Novelle* is a very stringent and clearly delineated type of text that must be distinguished from the genre of *Erzählungen*.[2] In contradiction to this position, several studies have suggested that, from a historical rather than a normative perspective, the label 'Novelle'

[1] As 'Novelle' and 'novella' do not have the same meaning, I will use, or rather refer to, the German term. Single quotation marks are used in this paper according to conventions in analytic philosophy. If a term or a sentence is addressed as the object of a statement, it is covered by single quotation marks, for example in statements such as: "The term 'novella' is used by person *X* in situation *S*." Both the whole sentence and the term 'novella' within this sentence are addressed as different observations. In contrast, if a person *X* claims, "Though included to Heyses's *Deutscher Novellenschatz*, Goethe's *Die neue Melusine* is not a novella but a tale," X uses the terms 'novella' and 'tale' and thus makes use of his concepts of novella and tale. Within the quoted claim, these terms are not covered by single quotation marks.

[2] See Baker (1999, 34): "The *Novelle* is regarded as one of the most stringent forms of prose as it deals with a closed, narrow segment of reality, where one particular event is important, and where there is one central conflict around which the whole is organized. Few characters are involved in the *Novelle* so the effect is heightened, and the action is generally of short duration."

does not signify any clear-cut literary genre (Polheim 1965; Meyer 1987; 1998; Lukas 1998). This paper asks who is right: the proponents of a strict genre of *Novellen* or their more recent and skeptical opponents? For CLGS, and more precisely, for analyses based on supervised machine learning, this endeavor presents a methodological challenge. If the skeptical position is right, which means that the *Novelle* does not behave like a category, we would expect that classification tasks will yield bad results. Usually, bad results require the learning or clustering algorithm to be optimized. This paper must thus elaborate a strategy to prove that bad results are due to the semantics of the genre concepts and not to methodic flaws. This essay accordingly develops such a strategy, outlining a psychological framework that will enable us to investigate the idiosyncratic structure of genres by means of classification and clustering algorithms.[3] This will include a clear notion of a baseline that makes sense of the distinction between bad, good, and expectedly weaker or stronger accuracy scores. In order to make clear why the looseness of genre concepts is worth being studied at all, this essay starts with taking one step back from common presuppositions to clarify our general interest in genres.

2. Why Genres Matter—Classificatory Versus Aesthetic Interest

In the case of loose genre concepts, scholarly work usually begins by offering a definition in order to get rid of any vague use of language that might threaten conceptual clarity (for the *Novelle*, see Lukas 1998, 252; Rath 2008). This first section of my essay will reject this default strategy and offer a better alternative. This requires distinguishing between two common interests: classificatory interest and aesthetic interest. Classificatory interest aims at a systematic concept of the metalanguage that is used to describe a specific object language.[4] Thus, the metalanguage is also called description language. In communities that stand in the tradition of the Vienna Circle or the analytic philosophy of science, the terms of our description language should be unambiguous,

[3] Underwood (2016 and 2019) proposed the idea of perspectival modeling to analyze the historical change or the "life spans" of historically discontinuous genres. This paper owes much to Underwood's idea but suggests a more psychological interpretation of the results of classification tasks.

[4] The distinction between metalanguage and object language relates to Tarski (1935). The notion of metalanguage shall roughly refer to a more or less strictly organized field of a regulated semantics, which is operatively used to conduct research. The notion of object language refers here to the semantics of historical use of genre concepts.

which can most suitably be achieved by definitions based on necessary and sufficient conditions. Thus, a definition of any term in a description language should meet the following requirement:

> (R1) A definition of a term '*c*' within a language *L* has to enable the users of *L* to determine whether any item *T* of any kind (in case of literary studies, mostly texts) is an instance of the concept which is expressed by the term '*c*.'

I call this the criterion of *decidability*. Fricke (2010b, 19) regards this requirement as essential to classificatory genre concepts. It is worth noting that he abandons another requirement postulating taxonomic simplicity so that any item *A* would be an instance of only one category on each categorical level (for example, level 1: epic, drama, poetry; level 2: novel, novella, short story, tale, tragedy, comedy, sonnet, elegy, etc.; level 3: adventure novel, education novel, historical novel, etc.). One reason why Fricke dispenses with the requirement of systematic simplicity is that another requirement obtains priority. This is the requirement of continuing ordinary language use:

> (R2) One important task of establishing an academic description language is to clarify vague ordinary language use. Carnap (1950) developed the procedure of clarifying vague concepts as explication.

This requirement (R2) is not a logical one but a formalization of a pragmatic aim of scholarly work. Physics and chemistry, for example, should help people better understand what water, carbon combustion, lightning, etc. actually are. By analogy, literary studies should help people come to a better understanding of what sonnets, elegies, novels, novellas, comedies, tragedies, and so on are. Explications clarify what the ordinary language words 'water' and 'novella' mean if described within scholarly description language. Most theorists of genre postulate that the concepts of concrete literary genres must necessarily be established as terms that serve the classificatory interest (Fricke 1981, Fricke 2010a, Fricke 2010b, Zymner 2003). At the same time, these theorists advocate for the second requirement. In short, most theories of genre claim that genre concepts must be *explicated as classificatory terms*.

This essay will question the claim this classificatory interest makes to such a general scope. The shortcomings of the classificatory stance can be inferred from arguments made by its prominent opponents. From the 1960s on, an approach oriented toward natural language and based on the notion of family resemblance as elaborated in Wittgenstein's *Philosophical Investigations* (1953) became more and more influential in theories of literary genre. Among the first supporters of this approach was Fishelov (1991). Hempfer and Strube refined this Wittgensteinian approach (Hempfer 2010b; 2014; Strube 1986). Their reasoning is built upon the presumed structure of ordinary language. Several genre concepts (such as elegy, *Novelle*, comedy) do not rest on the structure of classificatory concepts but on the structure of family resemblance, or on the structure of prototype theory, which is also based on Wittgenstein's thoughts (Rosch 1978). Although their Wittgensteinian line of reasoning is moving in the right direction, Hempfer and Strube run into a variant of the *is-ought-problem*.[5] As classificatory terms fulfill the function of more specifically describing and designating objects, classificatory definitions can be accepted as a pragmatically solid instrument. Thus, the empirical finding that everyday or even scholarly readers are using a concrete genre concept in a rather vague way does not entail that the concept of that genre must be used in the same vague manner in every possible analytical situation. Strube and Hempfer make the mistake of starting dogmatically from Wittgenstein's results when they assert that all concepts and, a fortiori, all genre concepts are vague. Instead, they should have started with Wittgenstein's way of asking questions. The proper question would be: which function is the concrete genre concept supposed to fulfill in a concrete analytical practice?

My claim is that for the task of interpreting texts and of reconstructing a concrete genre history, the concept of the respective genre need not fulfill any classificatory function but rather an aesthetic one. The aesthetic function is delineated by Fishelov, too: "In order to understand and to evaluate the writer's work, we are expected to take into account the generic background against which he operates" (1991, 135). The best reason why "we are expected to take into account the generic background against which" the author operates is provided by Walton (1970). At the core of Walton's argument is the "thesis that what aesthetic properties a work seems to have [...] often depends (in part) on which of its features are standard, which are variable, and which are contra-standard for us" (ibid., 343). Which features are standard, variable, or contra-standard is regulated by categorical expectations. One of the best examples of a category of art might be the sonata-allegro form. Walton shows that judgments on aesthetic properties, in particular on novelty, originality, and perfection, require that we know which features would be expected to be realized in that respective work. In Walton's sense, a work is perfectly shaped relative to a form, and innovative relative to the techniques and procedures that were operative at the time the work was produced. Walton is consequently able to formulate four requirements that must be satisfied if a work *W* is to be ascribed to a certain category of art *C*:

[5] I refer here to Hume's classic notion of the *is-ought-problem* in his *Treatise of Human Nature* (1739), which addresses the fallacy of unreasonably and often implicitly changing from descriptive statements (such as 'scholars do use vague concepts') to 'ought'-claims with the same propositional content (such as 'scholars ought to use vague concepts') without further reasoning (Hume, *A Treatise of Human Nature* [1739] 469–70, URL: https://en.wikisource.org/wiki/Treatise_of_Human_Nature/Book_3:_Of_morals/Part_1/Section_1.

i) The work *W* fits in with the category *C* inasmuch as "it has a minimum of contra-standard features" of *C* (ibid., 357).

ii) "*W* is better, or more interesting or pleasing aesthetically, or more worth experiencing when perceived in *C* than it is when perceived in alternative ways" (ibid.).

iii) "[T]he artist who produced *W* intended or expected it to be perceived in *C*, or thought of it as a *C*" (ibid).

iv) "*C* is well established in and recognized by the society in which *W* was produced" (ibid.).

Point (i) is a quasi-classificatory requirement, (ii) expresses the rule of maximizing aesthetic value, and (iii) and (iv) are historicizing requirements. In a previous study, I elaborated in more detail on how overemphasizing the maxim (ii) is very common to literary criticism because (ii) regulates the degree of originality and relevance of an interpretation but leads to ahistorical or anachronistic constructions of genre concepts (Schröter 2019, 232).

It might seem that referring to an aesthetic category requires a classificatory definition of that category. The distinction between *use* and *mention*, which is fundamental to analytic philosophy, comes into effect here. Genre theorists such as Fricke and Zymner claim that in scholarly work and interpretive practice we necessarily make use of the concepts of concrete genres whenever we designate and describe a concrete text as a novel, comedy, sonnet, etc. Hempfer's claim (2010a, 16) that whenever we *use* a concept, we use it within our description language, is an analytic statement. I agree with Fricke (2010a, 7) that the concepts of a description language should at least in part be defined as classificatory terms. Fricke's line of reasoning suggests that interpreting a text would be an instance of ascribing a work to a category. However, I consider this latter assumption to be a serious error.[6] Scholarly interpreters who are reading texts as *Novellen* seek to find out in what way these texts refer to certain historical understandings of 'Novelle.' Hence we have to take into account different kinds of reference: a text can be an instance of a genre, a provocation, or a challenge to a genre concept. Moreover, each text can be a typical or a very innovative instance of the respective genre and, finally, the reference to the genre can be merely ironic. This latter aspect has

[6] The same seems to apply when we are dealing with the history of a concrete genre because talking about the historical changes of a certain object presupposes that the object is determined. This affects this essay itself: if this text is about the history of German *Novellen*, the word *Novelle* seems to be used to designate the object of research. Therefore it appears that this essay should define the concept of *Novelle* in its description language. I do agree that scholarly work has to clarify the way it refers to the object which is to be investigated. However, I maintain that such a clarification does not have to be composed of a classificatory definition. Nor does it have to be composed of an explication in Carnap's sense. I offered a solution in another essay (Schröter 2019).

already been emphasized, or rather over-emphasized, by Derrida (1980, 59) and, for the *Novellen*, by Kiefer (2010, 14).[7] According to Walton's requirement (i), interpreting a text with reference to its genre seems at first glance to be an act of *using* a genre concept. However, whenever we really try to interpret a text in a historical way according to the author intentional requirement (iii) or the reader response requirement (iv), we do not actually *make use* of the genre concept. Instead, we are establishing one of several distinguishable connections between a work and a historical category. Establishing a connection between an artfact and a historical category can be regarded as a somewhat odd kind of *mentioning* a historical genre concept. We refer to the historical use of the genre concept and not to our systematic use. [8] However, this kind of reference usually entails a description of the historical use within our systematic language because we need to reconstruct the historical semantics of the concept.

In several cases of clear-cut genres such as the sonnet, the ode in the Asclepiadean style, or—perhaps—the joke, the classificatory interest matches the aesthetic interest by and large. Mentioning and using a genre concept amount to the same thing if the respective intensions of both the genre understanding and the classificatory term are consistent. Clear-cut genres might even be a regular case because poetological definitions mostly had the conceptual structure of classificatory concepts until the nineteenth century. This essay is concerned with the case of problematic or disordered genres where the aesthetic and the classificatory interest do not align. This leads me to the following two conclusions:

1. Reading literary works as an aesthetic endeavor that seeks to avoid crude anachronisms requires reference to historical understandings of genre. This implies that Walton's requirements (iii) and (iv) are of higher priority than (i) and (ii). In terms of cultural studies, this insight was already stated by Ryan: "The significance of generic categories thus resides in their cognitive and cultural value, and the purpose of genre theory is to lay out the implicit knowledge of the users of genres" (1981, 112). However, it is not only a matter of prioritization but also a logical issue. Whereas the academic practice of reconstructing historical concepts uses the same techniques as the practice of explication in Carnap's sense, its aim is quite different from the practices of explication and definition (Pawlowski 1980, 18–28). Reconstructing conceptual history tries to elucidate historical language

[7] Derrida might have overstated the possibility of ironic references because such references are very seldom. Neither the period of Biedermeier, nor that of poetic Realism rely on aesthetic deviance.

[8] In the German-speaking world, a distinction has been proposed between text type (*Textsorte*) and genre; the first designates the metalinguistic classificatory transhistorical concept, the latter refers to the respective historical understanding (Fricke 1981; 2010a). As I have demonstrated, this distinction is neither necessary nor useful to historiographical methodology (see Schröter 2019).

use, whereas explication and definition aim at regulating ordinary language use. My intention here is to make clear that neither scholarly nor lay readers need a classificatory concept of the respective genre defined in their description language when they interpret a literary text. In case the classificatory interest threatens to distort the reconstruction of a historical genre concept from an aesthetic point of view, the classificatory interest should be abandoned.

2. One of the main tasks of a literary history that supports reading as an aesthetic endeavor is to supply readers with all the various understandings of genre that are historically relevant. Of course, each historical understanding of genre itself has a conceptual structure, but this structure is not necessarily classificatory. It could be the structure of a concept of prototype or of loose family resemblance. In contrast to Strube (1986), I do not introduce the theory of conceptual structure on the level of our academic description language but on the level of historical semantics. The main constituent of each historical understanding of a concrete genre is the respective set of standard, variable, and contra-standard features that historical readers and authors linked with the genre labels. The following section elaborates the general requirements for investigating the genre of the *Novelle* with historical interest, beginning with the situation of the research specific to that genre.

3. The Situation of the *Novelle* and the Basic Outline of Its Aesthetic Historiography

The controversial situation of the German *Novelle* as outlined in the introduction, namely that of the *Novelle* as a clear-cut genre in contrast to skeptical opponents, raises several problems. First, this situation demands some clarification of both positions because this kind of contradiction raises suspicion that both parties are dealing with quite different objects. I assume that the older position is right according to its own terms, if applied exclusively to historical genre poetics. Indeed, historical poetics postulates, in a prescriptive manner, that the genre of the Novelle *should* be established as a highly valuable literary genre. However, the older position misses two central facts. First, that the actual text production did not meet the prescriptive claims. And second, that historical poetics of the *Novelle* mostly included a dialectical view of the situation and recognized that text production was widely heterogeneous and unstructured. A historiography of the German *Novellen* should therefore test five hypotheses as a first step:

1. German literary communication developed two labels, ‘Novelle’ and ‘Erzählung’ for mid-length prose fiction.

2. Whereas there was no poetics of the 'Erzählung,' the concept of 'Novelle' was subjected to intense theorizing and was included within cultural political semantics. Thus, the latter concept was invested with claims to high literary value.
3. The labels 'Novelle' and 'Erzählung' do not mark any substantial generic differences on the level of textual features (Polheim 1981; Lukas 1998).
4. The historical concepts of the '*Novelle*' differ synchronically between literary groups and change diachronically from generation to generation.
5. Generic differences have to be traced back to medial constraints of the almanacs, newspapers, and journals in which these texts were published, rather than to poetics (Meyer 1987, 1998).

The primary task of a history of German *Novellen* is to support these hypotheses with empirical evidence. Testing these hypotheses requires that we separate the history of the poetological genre concept from the generic and aesthetic expectations that readers based on their reading of texts they perceived as 'Novellen' or 'Erzählungen.' As I am interested in aesthetic expectations that were communicated by genre concepts in historical cultures, the apparent conflict between the second and the third hypothesis requires me to take into account that poetological reflection, as well as reading practice, might have served as two different and conflicting sources for generalizing aesthetic expectations. As the poetological concept and reading experience do not match in the case of *Novellen*, I regard the relationship between poetics and reading practice as dialectical. By dialectical, I mean that the strong aesthetic expectations historical readers gained from their poetological knowledge of the concept of *Novelle* were undermined by their reading practice, so that these readers either had to ignore, decide, or mediate this conflict.

4. Inductive Procedure

CLGS comes into play on the level of inferring categorical expectations from the recurrent features of text groups. It is reasonable to ask why textual analysis should be proceeded by computational analysis and not by close reading alone. The text group that has generally been taken as the basis for constructing the aesthetic category of *Novellen* in literary studies is rather small indeed. It consists of ten to twenty texts that were not explicitly published as *Novellen* in the first place.[9] Inferring categorical

[9] This holds for most canonized texts such as Johann Wolfgang von Goethe's tale *Die neue Melusine*, or the novellas by Adalbert von Chamisso, E. T. A. Hoffmann, Heinrich von Kleist, Franz Grillparzer, Annette von Droste-Hülshoff, Gottfried Keller, and Franz Kafka. By contrast,

expectations based on extensive close reading of these texts would lead to the same concept of the *Novelle* that is common to literary studies. The fourth hypothesis mentioned above claims that the concepts of *Novelle* differ between groups and change from generation to generation. If this thesis holds, the inferences based on a small canon of high literature will not elucidate the historical understanding of genre. Thus, an operationalization has to be chosen that makes it possible to test this hypothesis. The first step is to define historical groups and generations. Theories about system change (Titzmann 1991) and media change (Meyer 1987) suggest the following periodization: 1790–1820, 1820–1850, and 1850–1880–1910 (Schröter 2019, 233–35). There is also strong evidence that from 1850 on the social and medial system was subject to complex differentiation (Mellmann and Reiling 2016). Thus, historical situations S_i, defined by temporal as well as social group parameters, have to be generated heuristically and based on prior knowledge. As the formation of different situations is meant to identify stable language use within each situation, but change between two situations, it might turn out that two provisionally distinguished situations should be either merged or further differentiated. In the next step, we must determine which genre labels were used within each situation S_i for the set of texts to be considered. Though ugly in regards to readability, terms such as 'Novelle_$\mathrm{S_i}$' (in contrast to 'Erzählung_$\mathrm{S_i}$') must be used to designate the extension of relevant text groups. We do not know the intension or meaning of these terms in advance, but only the extension ('Novelle_$\mathrm{S_i}$' designates all texts that were regarded as *Novellen* in the historical situation S_i). This technique comes close to what is called "inductive procedure" in genre theory, as it is opposed to the deductive procedure (Müller 2010). In order to rid the inductive method of its methodological deficiencies, I reformulated this method in terms of quantity theory in such a way that the procedure fulfills its essential requirements (Schröter 2019, 240–47).[10]

the equally canonical novellas by Adalbert Stifter, Theodor Storm, C. F. Meyer, Thomas Mann, Arthur Schnitzler, and Robert Musil were published as 'Novellen' in the first print. Works by these authors are contained in all collections; see Wiese (1963); Swales (1977); Freund (1998); Reclam (2011).

[10] The problem with the inductive procedure is that, in the case of inconsistent historical use of names for genres, it is not suitable for inferring generic characteristics from text groups. My set-theoretical reformulation of the inductive procedure makes it possible to reconstruct the generic expectations historically associated with the generic label by making use of classificatory predicates. However, it is not the concept of genre that this procedure aims to determine prior to historiographic work. Rather, the aim is to relate the set of texts that were communicated as 'Novelle' or 'Erzählung' in the relevant historical situation S_i to classificatory sets by means of intersections. Each of these intersections—for example, the intersection of the texts referred to as 'Novelle' and of fictional journal prose—can be described with regard to the relevant text characteristics of the defined set (for example, fictional journal prose). It is important not to summarily define such intersections as a systematic concept of novella, as is often done. Such a definition would result in a classificatory concept of the respective genre, which in turn would

The set-theoretical reformulation of the inductive procedure enables us to reconstruct the generic expectations in the cases where genre labels have determined the perception of connected text groups.[11] It is, of course, another empirical and social-psychological assumption that the genre labels of the original publication caused the perception of text groups and led readers to produce genre concepts through a process of abstraction. In contrast to the way literary studies and CLGS have until now often modeled literary genres, ascribing a genre to a text is neither an ontological nor an epistemic matter in the case of disordered genres. It should not be regarded as simply true or false with regard to textual features and a defined genre concept that a certain text is a *Novelle*. Instead, it is our task, first, to collect original genre labels in heuristically defined historical situations and, second, to work out the rules for the use of these labels within these situations.

5. A Machine Learning Task

The following section deals with the problem of how a concrete machine learning task can be modeled in order to determine the aesthetic understanding of genre in a certain period. As indicated at the end of the last section, the fundamental question is twofold: did the members of S_i base their historical genre concept on inferences that consist of the following steps: (1) texts were grouped by the genre labels 'Novelle,' 'Erzählung,' 'Märchen' (tale), and 'Roman' (novel); and (2) recurrent textual patterns and text types were inferred from this type of grouping? This twofold question is also suited for elaborating in more detail the third of the hypotheses mentioned above—that there was no difference between *Erzählungen* and *Novellen* on a textual basis. According to the psychological frame I am suggesting, this thesis means that people within the historical situation S_i were not able to perceive any differences on the basis of grouping texts according to their original labels.

The corpus includes 509 texts from the long nineteenth century. What is important and new to this corpus is that it consists up to 30 percent (120 texts) of journal prose fiction texts that are currently being digitized in the course of a postdoctoral project and have not yet been discussed in literary studies.[12] The corpus design followed

no longer be suitable for comprehending the historical use of the generic label and hence the semantics of a genre in historical literary communication.

[11] This strategy is also pursued by Underwood (2019, 37), who defines a genre as a "group of books recognized by some specific, historically situated groups of readers."

[12] The project (*Habilitation*) "Ästhetische und soziale Funktionen der Erzählungen und Novellen im 19. Jahrhundert" ("Toward a functional history of nineteenth-century German novellas") started in 2018 and has been funded by the research fund of the Philosophical Faculty of the

Table 1 Corpus grouped by periods and genre labels

Situation/period	S_1: 1790–1820	S_2: 1820–1850	S_3: 1850–1880	S_4: 1880–1910	S_5: 1880–1910	S_0: total
*Novelle*_S_i	4	86	83	17	7	197
Other-novellas_S_i	34	31	35	15	5	120
*Erzählung*_S_i	30	45	13	3	4	95
*Märchen*_S_i (tales)	28	9	0	1	0	38
*Romane*_S_i (novels)	13	6	9	10	3	41
Non-fictional-report_S_i	5	4	0	0	–	9
Dorfgeschchte_S_i	0	2	3	0	–	5
Kriminalerzählung	0	0	2	0	–	2
Total	**114**	**183**	**145**	**46**	**19**	**509**

two main criteria: the share of genre labels should be representative of the historical situation in the media market, and canonized texts should not be overrepresented. These maxims are close to being satisfied for the periods from 1790 to 1850 but not yet for the periods from 1850 to 1940. The following Table 1 shows the proportions of the corpus grouped by periods and genre labels (see Table 1).

In the following, the accuracy scores of a supervised machine learning algorithm trained on texts grouped by genre labels shall be interpreted as a proxy for the cognitive ability of historical readers and authors to perceive these groups as genres. Thus, two borders have to be defined: first, that of the best possible classification; and second, a baseline that indicates the indistinguishability of two or more genres. I tested bootstrapped accuracy scores for three classification algorithms (logistic regression, support vector machine (SVM), and k (=5) nearest neighbors (KNN)), for pairs of two genres, and for different feature sets. Table 2 shows the three best predicted genre pairs for all algorithms and feature sets. As feature sets, I chose the absolute counts and the normalized frequency of 2,000 most frequent words with and without stop words:[13] (a) 2000_mfw_abs_with_stop_words, (b) 1615_mfw_abs_without_stopwords, (c) 2000_mfw_norm_with stop_words, (d) 2000_mfw_norm_without_stop_words.[14]

University of Wuerzburg. Over the course of the project, more than 2000 texts will be digitized. These 509 texts were the intermediate result of the digitization process.

[13] As normalization according to what is known as the Manhattan style (called ‘l1’ in *Python*’s sklearn) yielded results below the baseline (see below), I used the default quadratic normalization (‘l2’ in sklearn).

[14] Two thousand most frequent words without stop words and without proper names resulted in 1,615 words. For the justification of the use of logistic regression see Underwood (2016) and Underwood (2019).

Table 2 Bootstrapped accuracy scores for different feature sets, learning algorithms, and text groups

Features	*Roman* (novel) versus *Märchen* (tale)			*Novelle* versus *Märchen* (tale)			*Novelle* versus *Roman* (novel)		
	Log Reg	SVM	KNN	Log Reg	SVM	KNN	Log Reg	SVM	KNN
a) abs	0.75	0.84	0.97	0.78	0.84	0.82	0.65	0.72	0.92
b) abs, no stop words	0.83	0.95	0.93	0.83	0.89	0.80	0.68	0.73	0.87
c) norm	0.72	0.74	0.70	0.76	0.79	0.74	0.56	0.48	0.54
d) norm, no stop words	0.49	0.47	0.52	0.48	0.45	0.54	0.45	0.46	0.52

We see accuracy scores above 0.9 for absolute counts of most frequent words in particular for SVM and KNN.[15] This effect is due to the average length of text groups. Tales are significantly shorter than *Novellen* and *Novellen* are shorter than novels.[16] If we compare the stronger accuracy scores for feature sets (a) and (b) with the weaker scores for (c) and (d), it appears that all algorithms are trained basically on text length in case of (a) and (b). Thus, accuracy scores above 0.9 are possible in principle. However, I am interested here in the difference between text groups based on compositional strategy instead of text length.

In the next step, the baseline has to be established. This baseline also operationalizes the third hypothesis, that the labels 'Novelle' and 'Erzählung' do not mark any substantial generic differences on the level of textual features (see section 2). This hypothesis serves as the null hypothesis. In a pairwise classification task, the accuracy scores are expected to be 0.5 if the null hypothesis is true. In order to test the algorithm, I randomized the attribution of the genre labels in the sample before training and in each iteration of the bootstrapping. As expected, the empirical mean accuracy score with randomized genre labels is for all feature sets, algorithms, and for all pairs of text groups 0.5.

Surprisingly, the feature set (d) (normalized most frequent words with stop words removed) is on the level of that baseline (see Table 2). Table 3 shows the distribution of 1,000 iterations of bootstrapped sampling, training, and validating pairs of text groups with feature set (c) (2,000 most frequent words, with quadratic normalization and with stop words) and with logistic regression algorithm, which is preferred by Underwood (2019) because of its interpretability.[17]

[15] Improvement for SVM and logistic regression, when stop words are removed (feature set b), is not significant.

[16] An ANOVA test (analysis of variance) showed that the difference between average length is significant also with regard to the extremely large variance of text length in each group.

[17] Each sample consists of a pair of two groups that are grouped by their original genre label. Both groups are of equal size in the sample. The size of the validation set was set to 0.2 of the respective sample.

Table 3 Distribution of accuracy scores for pairwise training and validation for S_0: 1790–1940 (bootstrapping with 1,000 iterations, logistic regression)

Pair	Sample size	Mean of bootstrapped accuracy scores	Standard deviation	0.95 Empirical confidence interval for the distribution of bootstrapped accuracy scores
Novellen_S_0 versus Erzählungen_S_0	190	0.64	0.08	0.47–0.81
Novellen_S_0 versus other novellas_S_0	242	0.56	0.07	0.43–0.69
Novellen_S_0 versus Romane_S_0 (novels)	88	0.55	0.13	0.28–0.78
Novellen_S_0 versus Märchen_S_0 (tales)	76	0.76	0.11	0.50–0.93
Novellen_S_1+S_2 (before 1850) versus Novellen_S_3-S_5 (after 1850)	180	0.67	0.07	0.50–0.80
Novellen_S_1+S_2 versus Erzählugen_S_1+S_2 (before 1850)	180	0.64	0.09	0.43–0.80

All accuracy scores for the pairs listed in Table 3 clearly miss the level of 0.9, which was achieved with features set (a) for *Märchen* ('tales') versus *Novellen* or novels. In fact, the results are much closer to the baseline, so that we have to ask whether the null hypothesis can be rejected. According to the empirical 0.95 confidence interval, the difference between *Novellen* and *Romane*, *Erzählungen* and other novellas without a genre label is clearly insignificant.[18] Surprisingly the difference between *Novellen* and tales is as significant as that between *Novellen* before 1850 and *Novellen* after 1850. This indicates that the semantic change of the concept of *Novelle* in the nineteenth century is stronger than the semantic difference between *Erzählung*, *Novelle*, and *Roman* with regard to textual features (and compositional strategy) beyond text length. Hence the null hypothesis cannot be rejected for the pairs of *Novellen*, *Erzählungen*, other novellas and novels.[19] There is, however, one more interesting detail: the difference between *Novellen* and *Erzählungen*, though not strongly significant, might in effect be larger than the difference between *Novellen* and *Romane* (novels) if the effect size is calculated as the difference between two means. This result corresponds to a common assumption

[18] This can be directly drawn from Table 4, as for the mentioned pairs, in more than 2.5 percent of all iterations the accuracy scores are below the baseline of 0.5.

[19] In contrast to genre categories, media types (such as almanacs, journals, or anthologies) prove more stable within limited periods. This might support the fifth thesis (see section 3 above), which was put forward by Meyer. However, this essay is about genre labels and not about the role of media.

that in the first half of the nineteenth century, the labels *Novelle* and *Roman* were used synonymously (Meyer 1998). However, due to small sample sizes for the periods after 1850 (S_3 and S_4, see Table 1), it is not yet possible to compare text groups relative to shorter periods except for the largest group of the *Novellen* for all periods and *Erzählungen* for the time before 1850 (see last two rows of Table 3).

Although greater effort could be put into validating and refining the technical procedures, which I cannot provide here, I maintain that we should be able to interpret the weak results represented here from a psychological and aesthetic perspective. From a psychological point of view, the results suggest that readers in the nineteenth century were able to learn what they should expect from tales compared to what they should expect from novels. Likewise, they were able to identify *Novellen* compared to tales. In contrast, readers might have had difficulty understanding the differences between *Novellen* and novels, even more so than between *Erzählungen* and *Novellen*. Thus, historical genre semantics that were induced from reading texts with certain labels (see section 3) do not yield blurred or clear-cut genres per se. Instead, genre semantics based on label use rely upon relating two or more text groups. In the nineteenth century, the *Novelle* might have appeared as a clear-cut and a blurred genre at the same time: historical readers saw it as clear-cut when they compared *Novellen* with tales, and they saw it as blurred when they compared *Novellen* with novels or *Erzählungen*.

In order to improve methods, optimization by means of starting from well-formed classes should not meet the requirements of the classificatory interest as presented in the first section of this essay. Rather, it should be oriented toward historical language use. Thus, I construed the group of novellas_merged, which is a sample of *n* texts from the set of *Novellen*, *Erzählungen*, and no-label-novellas. This sample is intended to map a historical practice that was common to nineteenth-century reading culture. Realist mid-length narrative prose was merged and regarded as 'Novellen' or 'Erzählungen' irrespective of the original genre labels. Although 'novella' is introduced here as classificatory terms of my description language,[20] it is intended to represent the historical practice of merging all fictional prose texts in journals, which was mostly labeled as 'Novelle' or 'Erzählung.' Based on this grouping, we can test whether the novellas_merged group helped historical readers to distinguish novellas not only from tales, but also from novels. Thus, in Table 4, the novellas_merged group is compared to novels and tales.

This result does not give a definitive answer. At first glance, it seems that the novellas_merged group is not significantly different from novels. In order to see things

[20] The term 'novella' serves a classificatory interest as elaborated in the first section of this paper and is defined as realist narrative mid-length prose fiction, published in journals and anthologies or as a separate book publication. The term 'no-label-novella' designates novellas that had no genre label in their first publication.

Table 4 Novellas_merged compared to novels and tales

Pair	Sample size	Mean of bootstrapped accuracy scores	Standard deviation	95% Empirical confidence interval for the distribution of bootstrapped accuracy scores
Novellas_merged_S_0 versus tales_S_0	76	0.73	0.12	0.50–0.93
Novellas_merged_S_0 versus novels_S_0	88	0.57	0.12	0.33–0.78

more clearly, the following scatter plot of a principle component analysis based on the absolute counts of most frequent words (feature set a) gives a visual sense of the obtained classification results for the original groups according to actual genre labels.

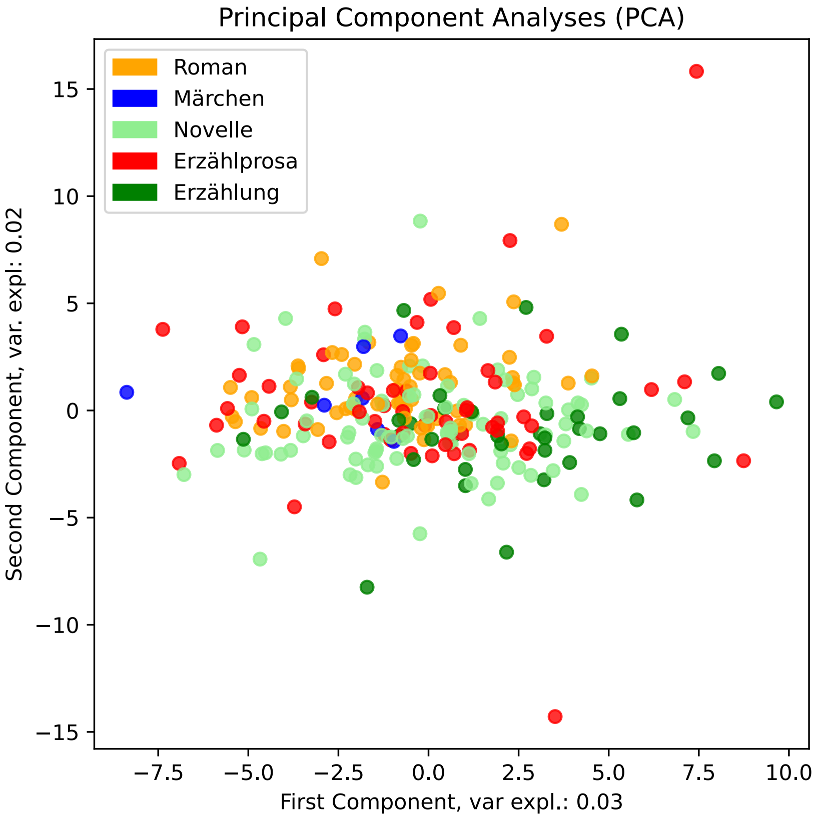


Fig. 1 Principal component analysis and genre labels (Schröter, CC BY).

Figure 1 provides all texts with their original genre labels. This visualization suggests that there is no genre structure at all in the data. If all mid-length prose fiction is regarded as one group (novellas merged) (see Figure 2), the plot looks more uncluttered.

Now we can understand why novellas_merged are not clearly different from novels. Many novellas spread into the group of novels. Nevertheless, we get the picture that, by and large, novellas as the text type of mid length prose fiction are slightly different from fairy tales (*Märchen*) and novels (*Romane*) and that the internal differentiation of novellas into *Novellen* and *Erzählungen*, which is so characteristic of German culture, does not correspond to different text types.

Finally, a considerable bias within the corpus can be seen in Figure 2. Fairy tales, which are represented only in a very small sample of 7 instances here, seem to be closer to each other compared to the remainder of the genre groups. This suggests that fairy

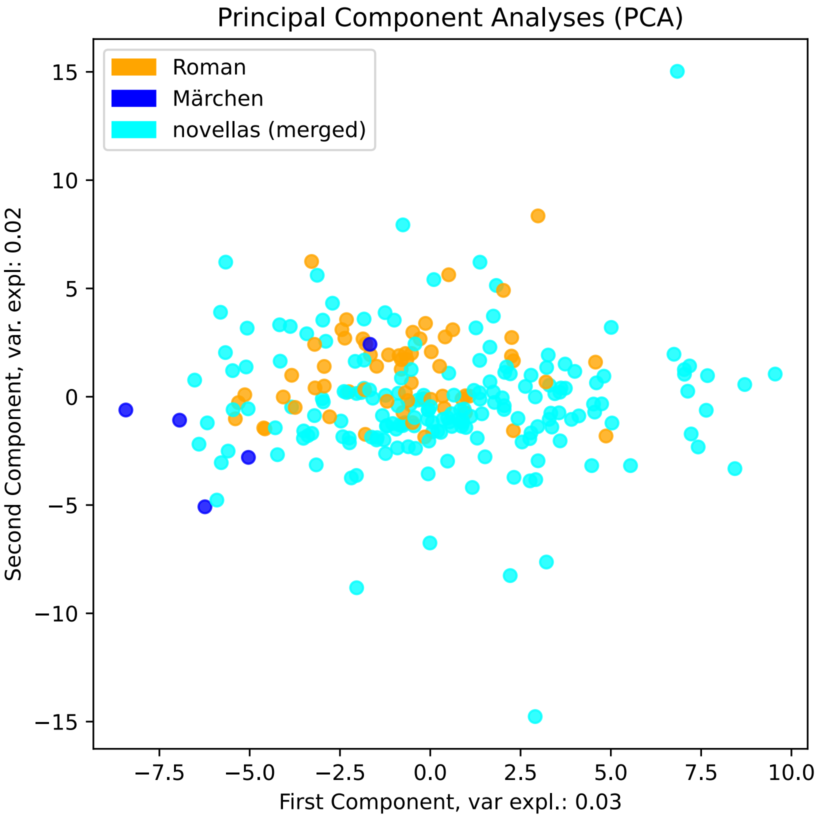


Fig. 2 Principal component analysis for merged genres (Schröter, CC BY).

tales might be slightly more homogeneous in style than novels. As the 38 fairy tales in the whole corpus have been written by only ten different authors (17 by the Brothers Grimm; by contrast the 41 novels are written by 37 different authors), this result should be traced back to the style of the author rather than that of the genre. This bias is already controlled here by sampling only one text written by each author. From the perspective of literary history, things are intricate because this kind of bias might be essential to nineteenth-century knowledge: most of the well-known fairy tales of this period were written or transmitted by the Brothers Grimm. Thus, the composition of fairy tales in the corpus might indeed be regarded as fairly representative of nineteenth-century culture. The well-known fairy tales are rather homogeneous in style also, but not only, because they were written by a small number of authors. However, the main insight here is that, form the perspective of CLGS, and featuring a bag of words model, PCA and classification tasks indicate unanimously that it is more reasonable to assume only one text type of the novella, which can be defined as mid-length prose fiction, rather than assuming different novella sub-types.

6. Conclusion: Toward a Psychological Frame and Why Aesthetic Interest Matters

At first glance the results shown above seem to suggest that a wide classificatory concept of novella encompassing all narrative mid-length prose fiction should be favored. This would weaken my claim that the aesthetic interest should be prioritized over the classificatory interest. On closer observation, it becomes clear that these results actually support my claim. Initially, it was necessary to use a definition of the term 'novella' in the description language because I had to identify the group of no-label-novellas. Several scholars have chosen 'Novelle' as a term to designate all mid-length prose fiction in journals (see Lukas 1998). The extension of this definition is captured by the novellas_merged group, which covers all texts labeled as 'Novellen,' 'Erzählungen,' and prose fiction in journals without label or with different labels. CLGS could be tempted to do the same, because this definition would yield much better clustering and slightly better classification results. As definitions are arbitrary, there is nothing wrong with this procedure from the perspective of genre theory. From the perspective of genre history however, Lukas's decision is fatal because it blocks the access to the historical use of the term 'Novelle' once and for all. It is possible for 'novella' to serve as a term, but 'Novelle' should be referred to as a historical genre label that can be used to construct samples of texts grouped by original genre labels. If the intention is to reconstruct historical genre understanding from historical language use, the group of texts that were labeled as *Novellen* in a certain historical situation S_i must not be defined by textual features but by the historical use of the genre label. I maintain that this inductive procedure should be applied to all disordered genres.

The fourth section of this essay elaborated a technique to answer the question of whether authors, readers, and other actors in a historical situation S_i might have been able to infer categorical genre expectations based on the procedure of grouping texts according to their original genre labels. I maintain that this question should be the first step to reconstruct historical genre understanding. However, this question actually pertains to the *psychological* issue of whether the members of S_i were able to perceive categorical differences between novels and *Novellen*. Walton already emphasized this psychological dimension (see 1970, 338). However, up to this point, my answer has consisted of bootstrapped accuracy scores for classification tasks and of a scatter plot based on principal component analysis. Therefore, in the next step I will have to clarify how statistical significance correlates with psychological significance as outlined in Figure 3.

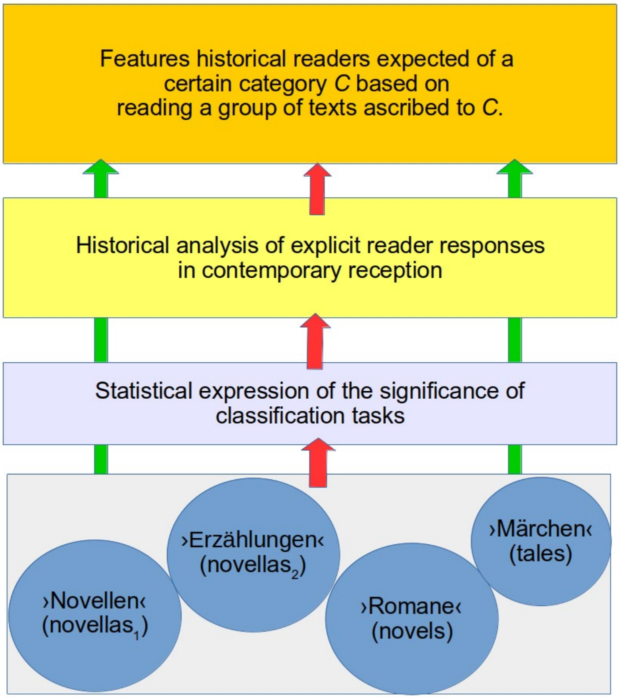

Features historical readers expected of a certain category *C* based on reading a group of texts ascribed to *C*.
Historical analysis of explicit reader responses in contemporary reception
Statistical expression of the significance of classification tasks
›Novellen‹ (novellas$_1$)
›Erzählungen‹ (novellas$_2$)
›Romane‹ (novels)
›Märchen‹ (tales)

Fig. 3 Outline of correlating statistical and psychological significance (Schröter, CC BY).

The vertical axis of Figure 3 reflects two different types of processes. The green arrows in the background indicate a latent historical process that is intended to be revealed between the lower und the upper box. The lower box shows the input: the texts that were read and grouped by genre labels. As a result, readers come to their genre concepts based (at least in part) on reading and grouping texts. From an aesthetic and historiographic perspective, genre research should determine the semantics of the category *C* (in the upper box). The levels between the upper and the lower box are latent and need to be revealed by historical research because these levels are suitable to explain the development and change of historical semantics. This essay has taken the first step by giving a statistical expression of the distinctiveness of text groups (second box from the bottom). The third box from the bottom refers to the psychological task of formulating statistical results as psychological hypotheses. I would like to propose that two procedures can be combined here: empirical studies of literary response might help to translate the statistical results of computational text analysis to psychological pattern recognition. There is occasional evidence that authors and readers in the nineteenth century, such as Gottfried Keller, doubted that there was any actual difference between *Novellen* and *Erzählungen* and between *Novellen* and novels. More systematic evidence from explicit reader responses could help to validate statistical results. This procedure is shown in the yellow box. If combined with methods of CLGS as elaborated in the fourth section of this paper and as illustrated in the second box from the bottom in Figure 2, classification tasks can be interpreted as a metric that measures the degree of looseness of historical genre concepts. This would lead to a new way of interpreting supervised learning, that comes close to Underwood's (2019) proposal. This method of interpretation would use supervised machine learning to make hermeneutical sense of conceptual looseness within and—if further refined—between historical reading cultures, rather than using it to determine any presumed textual differences between texts in order to reproduce distinctions that have already been established. I argue that research into aesthetic genres should focus on this kind of historical semantics even in cases where the use of genre concepts seems to be arbitrary.

Data repository

Julian Schröter. 2024. julianschroeter/19CproseCorpus: 19th Century Prose Corpus (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.10801777

Code repository

Julian Schröter. 2023. julianschroeter/PyNovellaHistory: Python Code for the Project on the history of the German 19th-century novella (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.7945348

ORCID®

Julian Schröter https://orcid.org/0000-0003-0168-2608

References

Baker, Christine L. 1999. *Landscapes in Theodor Storm's Novellen*. Leicester: University of Leicester.

Borcherdt, Hans Heinrich. 1926. *Geschichte des Romans und der Novelle in Deutschland*. Leipzig: Weber.

Carnap, Rudolf. 1950. *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Derrida, Jacques. 1980. "The Law of Genre." *Critical Inquiry* 7 (1): 55–81.

Evert Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. "Understanding and Explaining Delta Measures for Authorship Attribution." *Digital Scholarship in the Humanities* 32 (2): ii4–ii16.

Fishelov, David. 1991. "Genre Theory and Family Resemblance – Revisited." *Poetics* 20 (2): 123–138.

Freund, Winfried 1998. *Deutsche Novellen*. München: Fink.

Freund, Winfried. 2009. *Novelle*. Erweiterte und bibliographisch ergänzte Ausgabe. Stuttgart: Reclam.

Fricke, Harald. 1981. *Norm und Abweichung: Eine Philosophie der Literatur.* München: Beck.

Fricke, Harald. 2010a. "(A) Aspekte der literaturwissenschaftlichen Gattungsbestimmung. 1 Methodische Aspekte. 1.1 Definitionen und Begriffsformen." In *Handbuch Gattungstheorie*, edited by Rüdiger Zymner, 7–10. Stuttgart, Weimar: Metzler.

Fricke, Harald. 2010b. “(A) Aspekte der literaturwissenschaftlichen Gattungsbestimmung. 1 Methodische Aspekte. 1.5 Invarianz und Variabilität von Gattungen.” In *Handbuch Gattungstheorie*, edited by Rüdiger Zymner, 19–21. Stuttgart, Weimar: Metzler.

Füllmann, Rolf. 2010. *Einführung in die Novelle*. Darmstadt: WBG.

Hempfer, Klaus W. 2010a. “Generische Allgemeinheitsgrade.” In *Handbuch Gattungstheorie*, edited by Rüdiger Zymner, 15–19. Stuttgart, Weimar: Metzler.

Hempfer, Klaus W. 2010b. “Zum begrifflichen Status der Gattungsbegriffe: Von ‘Klassen’ zu ‘Familienähnlichkeiten’ und ‘Prototypen.’” *Zeitschrift für französische Sprache und Literatur* 120 (1): 14–32.

Hempfer, Klaus W. 2014. “Some Aspects of a Theory of Genre.” In *Linguistics and Literary Studies/ Linguistik und Literaturwissenschaft, Interfaces, Encounters, Transfers/ Begegnungen, Interferenzen und Kooperationen*, edited by Monika Fludernik and Jacob Daniel, 405–22. Berlin, Boston: De Gruyter.

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

Himmel, Hellmuth. 1963. *Geschichte der deutschen Novelle*. Bern, München: Francke.

Jockers, Matthew Lee. 2013. *Macroanalysis: Digtital Methods and Literary History*. Urbana: University of Illinois Press.

Kiefer, Sascha. 2010. *Die deutsche Novelle im 20. Jahrhundert*. Köln, Weimar, Wien: Böhlau.

Klein, Johannes. 1960. *Geschichte der deutschen Novelle von Goethe bis zur Gegenwart*. 4th ed. Wiesbaden: Steiner.

Kunz, Josef. 1970. *Die deutsche Novelle im 19. Jahrhundert.* Berlin: Schmidt.

Lukas, Wolfgang. 1998. “Novellistik.” In *Zwischen Restauration und Revolution 1815–1848. Hansers Sozialgeschichte*, vol. 5, edited by Gerd Sautermeister et. al., 251–80. München: Hanser.

Meier, Albert, and Simone Vrckovski. 2014. *Novelle: eine Einführung*. Berlin: Schmidt.

Mellmann, Katja, and Jesko Reiling, eds. 2016. *Vergessene Konstellationen literarischer Öffentlichkeit zwischen 1840 und 1885. Studien und Texte zur Sozialgeschichte der Literatur*. Berlin, Boston: De Gruyter.

Meyer, Reinhart. 1987. *Novelle und Journal, I: Titel und Normen: Untersuchungen zur Terminologie der Journalprosa, zu ihren Tendenzen, Verhältnissen und Bedingungen*. Stuttgart: Steiner.

Meyer, Reinhart. 1998. “Novelle und Journal.” In *Zwischen Restauration und Revolution 1815–1848. Hansers Sozialgeschichte*, vol. 5, edited by Gerd Sautermeister et. al., 234–50. München: Hanser.

Müller, Ralph. 2010. “Korpusbildung.” In *Handbuch Gattungstheorie*, edited by Rüdiger Zymner, 23–25. Stuttgart, Weimar: Metzler.

Pawłowski, Tadeusz. 1980. *Begriffsbildung und Definition*. Berlin, New York: De Gruyter.

Polheim, Karl Konrad. 1965. *Novellentheorie und Novellenforschung: Ein Forschungsbericht.* Stuttgart: Metzler.

Polheim, Karl Konrad. 1981. “Gattungsproblematik.” In *Handbuch der deutschen Erzählung*, edited by Karl Konrad Polheim, 9–16. Düsseldorf: Bagel.

Rath, Wolfgang. 2008. *Die Novelle: Konzept und Geschichte*. 2nd ed. Göttingen: Vandenhoeck & Ruprecht.

Reclam. 2011. *Erzählungen und Novellen des 19. Jahrhunderts* Vol. 1. Stuttgart: Reclam.

Rosch, Eleanorl. 1978. "Principles of Categorization." In *Cognition and Categorization*, edited by Eleanor Rosch and Barbara B. Lloyd, 27–48. Hillsdale, NJ: Lawrence Erlbaum.

Ryan, Marie L. 1981. "Introduction: On the Why, What, and How of Generic Taxonomy." *Poetics* 20 (2–3): 109–126.

Schöch, Christof. 2017. "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Schröter, Julian. 2019. "Gattungsgeschichte und ihr Gattungsbegriff am Beispiel der Novellen." *Journal of Literary Theory* 13 (2): 227–57.

Strube, Werner. 1986. "Sprachanalytisch-philosophische Typologie literaturwissenschaftlicher Begriffe." In *Zur Terminologie der Literaturwissenschaft*, edited by Christian Wagenknecht, 35–49. Stuttgart: Metzler.

Swales, Martin. 1977. *The German Novelle*. New Jersey, Princeton: Princeton University Press.

Tarski, Alfred. 1935. "Der Wahrheitsbegriff in den formalisierten Sprachen." *Studia Philosophica* 1: 261–405.

Titzmann, Michael. 1991. *Modelle des literarischen Strukturwandels. Studien und Texte zur Sozialgeschichte der Literatur*. Tübingen: Niemeyer.

Underwood, Ted. 2016. "The Life Cycles of Genres." *Journal of Cultural Analytics.* 2 (2) https://doi.org/10.22148/16.005.

Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change*. Chicago, London: The University of Chicago Press.

Walton, Kendall L. 1970. "Categories of Art." *Philosophical Review* 79 (3): 334–67.

Wiese, Benno von. 1963. *Die deutsche Novelle von Goethe bis Kafka*. Düsseldorf: Bagel.

Zymner, Rüdiger. 2003. *Gattungstheorie. Probleme und Positionen der Literaturwissenschaft*. Paderborn: mentis.

Exploring Fictional Styles along Universal Dimensions of Register Variation

Douglas Biber and Jesse Egbert

Abstract Multi-dimensional (MD) analyses have been carried out to identify the linguistic parameters of register variation in many different discourse domains and many different languages (see, e.g., Biber 1988; 1995; 2014). Each MD study has identified linguistic dimensions that are peculiar to a particular language/discourse domain. However, the more theoretically interesting finding is that linguistically similar dimensions emerge in nearly all MD studies. Two of these dimensions are especially robust, making them strong candidates for universal dimensions of register variation: 1) a fundamental opposition between clausal/“oral” discourse versus phrasal/“literate” discourse, and 2) the opposition between “narrative” versus “non-narrative” discourse. It turns out that these same two functional parameters are fundamentally important in the discourse domain of fictional literature. The present paper overviews results of MD studies of English across discourse domains, and shows how these two universal dimensions are also fundamentally important in fictional literature.

Keywords Multi-dimensional analyses, register, literature

1. Introduction

One major focus of previous corpus-based research has been to describe the ways in which linguistic features vary across registers (e.g., conversation, classroom teaching, newspaper editorials). Such research can be carried out for many different research purposes, including a detailed description of a single register, comparing the patterns of register variation, or even describing patterns of variation across the styles of individual speakers or authors.

Biber, Douglas and Egbert, Jesse: Exploring Fictional Styles along Universal Dimensions of Register Variation, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 197–215. DOI: https://doi.org/10.17885/heiup.1157.c19372

Multi-Dimensional (MD) analysis is a methodological approach that has been applied to all of these general research goals. MD analyses have been conducted for many different discourse domains and many different languages. Using bottom-up statistical analyses, these studies have investigated specific patterns of register variation in several different discourse domains of English, as well as the more general patterns of register variation in numerous languages. Each study identifies linguistic dimensions that are peculiar to that particular language/domain. However, the more theoretically interesting finding is that linguistically similar dimensions emerge in nearly all of these studies. Two of these dimensions are especially robust, making them strong candidates for universal dimensions of register variation: 1) a fundamental opposition between clausal/"oral" discourse versus phrasal/"literate" discourse, and 2) the opposition between narrative versus non-narrative discourse. It turns out that these same two dimensions are important for distinguishing among the author styles employed in fictional novels. In the sections below, we first introduce the methodology of MD analysis, then briefly survey previous MD research studies with an emphasis on these potentially universal patterns of register variation, and finally present a new MD analysis of fictional novels, showing how these same two parameters of variation are important in that discourse domain.[1]

2. Overview of MD Analysis

The *Multi-Dimensional (MD)* analytical approach was originally developed to investigate the linguistic patterns of variation among spoken and written registers (see, e.g., Biber 1986; 1988; 1995). Studies in this research tradition have used large corpora of naturally-occurring texts to represent the range of spoken and written registers in a language. These registers are compared with respect to "dimensions" of variation (identified through a statistical factor analysis), comprising constellations of linguistic features that typically co-occur in texts. Each dimension is distinctive in three respects:

— It is defined by a distinct set of co-occurring linguistic features;
— It is associated with particular communicative functions;
— There are different patterns of register variation associated with each dimension.

The MD approach uses statistical factor analysis to reduce a large number of linguistic variables to a few basic parameters of linguistic variation: the "dimensions." In MD

[1] Sections 2 and 3 are based on earlier summaries of the research goals, methods and findings of MD analysis, especially Biber (2014) and Biber (2019).

analyses, the distribution of individual linguistic features is analyzed in a corpus of texts. Factor analysis is then used to identify the systematic co-occurrence patterns among those linguistic features—the "dimensions"—and then texts and registers are compared along each dimension. Each dimension comprises a group of linguistic features that usually co-occur in texts (e.g., nouns, attributive adjectives, prepositional phrases). The dimensions are then interpreted to assess their underlying functional associations.

The first book-length MD analysis (Biber 1988) investigated the relations among general spoken and written registers in English, based on analysis of the LOB (Lancaster-Oslo-Bergen) Corpus (15 written registers) and the London–Lund Corpus (6 spoken registers). 67 different linguistic features were analyzed computationally in each text of the corpus. Then, the co-occurrence patterns among those linguistic features were analyzed using factor analysis, identifying the underlying parameters of variation: the factors or "dimensions."

After the statistical analysis is completed, dimensions are interpreted functionally, based on the assumption that linguistic co-occurrence reflects underlying communicative functions. That is, linguistic features occur together in texts because they serve related communicative functions. Table 1 summarizes the first two dimensions from the 1988 factor analysis, including a list of the most important linguistic features comprising each dimension as well as the interpretive functional labels.

Table 1 Summary of the major linguistic features co-occurring on Dimensions 1 and 2 from the 1988 MD analysis of register variation

	Dimension 1: **Involved versus informational production**	**Dimension 2:** **Narrative versus non-narrative discourse**
Positive features	mental (private) verbs, *that* complementizer deletion, contractions, present tense verbs, *WH*-questions, 1st and 2nd person pronouns, pronoun *it*, indefinite pronouns, *do* as pro-verb, demonstrative pronouns, emphatics, hedges, amplifiers, discourse particles, causative subordination, sentence relatives, *WH*-clauses	past tense verbs, 3rd person pronouns, perfect aspect verbs, communication verbs
Negative features	nouns, long words, prepositions, type/token ratio, attributive adjectives	present tense verbs, attributive adjectives

Each dimension can have positive and negative features. Rather than reflecting importance, positive and negative signs identify two groupings of features that occur in a complementary pattern as part of the same dimension. That is, when the positive features occur together frequently in a text, the negative features are markedly less frequent in that text, and vice versa.

For Dimension 1, the interpretation of the negative features is relatively straightforward. Nouns, word length, prepositional phrases, type/token ratio, and attributive adjectives all reflect an informational focus, a careful integration of information in a text, and precise lexical choice. The set of positive features for Dimension 1 is more complex, although all of these features have been associated with interpersonal interaction, a focus on personal stance, and real-time production circumstances. For example, first and second person pronouns, *WH*-questions, emphatics, amplifiers, and sentence relatives can all be interpreted as reflecting interpersonal interaction and the involved expression of personal stance (feelings and attitudes). Other positive features are associated with the constraints of real-time production, resulting in a reduced surface form, a generalized or uncertain presentation of information, and a generally "fragmented" production of text; these include *that*-deletions, contractions, pro-verb *do*, the pronominal forms, and final (stranded) prepositions.

Overall, Dimension 1 represents a parameter marking interactional, stance-focused, and generalized content (the positive features on Table 1) versus high informational density and precise word choice (the negative features). Two separate communicative considerations seem to be represented here: the primary purpose of the writer/speaker (involved versus informational), and the production circumstances (those restricted by real-time constraints versus those enabling careful editing possibilities). Reflecting both of these parameters, the interpretive label "Involved versus informational production" was proposed for the dimension underlying this factor.

A second major step in interpreting a dimension is to consider the similarities and differences among registers with respect to the set of co-occurring linguistic features. To achieve this, *dimension scores* are computed for each text, by summing the individual scores of the features that co-occur on a dimension (see Biber 1988: 93–97). Once a dimension score is computed for each text, the mean dimension score for each register can be compared across registers. For example, Figure 1 plots the mean dimension scores of registers along Dimension 1 from the 1988 MD analysis. The registers with large positive values (such as face-to-face conversations) have high frequencies of positive Dimension 1 features (e.g., present tense verbs, private verbs, etc.) combined with low frequencies of negative Dimension 1 features (e.g., nouns, prepositional phrases, etc.). Registers with large negative Dimension 1 values (e.g., academic prose, official documents) have the opposite linguistic characteristics.

The relations among registers shown in Figure 1 confirm the interpretation of Dimension 1 as distinguishing among texts along an oral/literate continuum. At the positive extreme, conversations are highly interactive and involved, with the language produced under real-time circumstances. And at the negative extreme, registers such as academic prose are non-interactive but highly informational in purpose, produced under circumstances that permit extensive revision and editing.

```
INVOLVED
     |
     | TELEPHONE CONVERSATIONS
     |
  35 + FACE-TO-FACE CONVERSATIONS
     |
     |
     |
  30 +
     |
     |
     |
  25 +
     |
     |
     |
  20 + Personal letters
     | PUBLIC CONVERSATIONS, SPONTANEOUS SPEECHES
     | INTERVIEWS
     |
  15 +
     |
     |
     |
  10 +
     |
     |
     |
   5 +
     | Romance fiction
     | PREPARED SPEECHES
     |
   0 + Mystery and adventure fiction
     | General fiction
     | Professional letters
     | BROADCASTS
  –5 +
     | Science fiction
     | Religion
     | Humor
 –10 + Popular lore, editorials, hobbies
     |
     | Biographies
     | Press reviews
 –15 + Academic prose, Press reportage
     |
     | Official documents
     |
INFORMATIONAL
```

Fig. 1 Mean scores of registers along Dimension 1: Involved versus informational production. Written registers are in italics; spoken registers are in CAPS. ($F = 111.9$, $p < .0001$, $r^2 = 84.3$ percent), (adapted from Figure 7.1 in Biber 1988).

The overall comparison of spoken and written registers in the 1988 MD analysis requires consideration of all dimensions of variation, which each define a different set of relations among spoken and written registers. For example, Dimension 2 (see Table 1) is interpreted as "Narrative versus non-narrative concerns." The positive features—past tense verbs, third-person pronouns, perfect aspect verbs, communication verbs, and present participial clauses—are associated with past-time narration. In contrast, the negative features—present tense verbs and attributive adjectives—have non-narrative communicative functions. The distribution of registers along Dimension 2 supports the interpretation as narrative versus non-narrative concerns. All types of fiction have high positive scores on this dimension, reflecting their emphasis on narrating events. In contrast, registers which are typically more concerned with events currently in progress (e.g., broadcasts) or with building arguments rather than narrating (e.g., academic prose) have negative scores on this dimension.

The other dimensions in the analysis can be interpreted in a similar way. Overall, the 1988 MD analysis showed that English registers vary along several underlying dimensions associated with different functional considerations, including: interactiveness, involvement and personal stance, production circumstances, informational density, informational elaboration, narrative purposes, situated reference, persuasiveness or argumentation, and impersonal presentation of information.

Subsequent MD studies have shown that some of these dimensions turn out to be peculiar to English or to particular discourse domains (see Biber 1995; 2014). However, two linguistic parameters of variation have emerged consistently as dimensions across MD studies: a basic oral/literate parameter of variation, and a narrative/non-narrative dimension. The linguistic compositions and functional associations of these dimensions are remarkably stable across discourse domains (and languages, such as Spanish, Korean, Somali, and Czech; see Biber 2014 for an earlier survey), although the particular patterns of register variation differ according to the specific language/culture/discourse domain. In the following section, we briefly document the characteristics of these dimensions in previous MD studies of English, and then turn to an MD analysis of fictional styles in Section 4.

3. Universal Dimensions in Previous MD Studies of English

Numerous studies have undertaken MD analyses of particular discourse domains in English (as well as studies of different languages; see Biber 2014). Given that each of these studies is based on a different corpus of texts (representing a different discourse domain or different language), and based on increasingly comprehensive sets of linguistic features (as computational techniques for linguistic analysis have improved),

Table 2 The oral/literate dimension in selected MD studies of English

Discourse domain	Linguistic features defining Dimension 1	Register pattern along Dimension 1
University spoken and written registers; Biber (2006)	contractions, pronouns, present tense verbs, progressive aspect, time / place / stance adverbials, *that*-clauses, *WH*-clauses, adverbial clauses VERSUS nouns, nominalizations, attributive adjectives, prepositional phrases	service encounters, office hours, study groups, classroom teaching VERSUS textbooks, course packs, institutional writing
Conversational text types; Biber (2008a)	contractions, 1st and 2nd person pronouns, activity verbs VERSUS long words, abstract nouns, nominalizations, attributive adjectives, prepositional phrases	casual conversations VERSUS work-place conversations
Academic research articles across disciplines; Gray (2013)	pronouns, causative verbs, modals, stance and time adverbials, conditional adverbial clauses, *that* complement clauses, *wh*-clauses, *to*-clauses VERSUS nouns, past tense verbs, prepositions, type-token ratio, word length; passives	theoretical philosophy VERSUS quantitative biology, quantitative physics

it is reasonable to expect that they would each identify a unique set of dimensions. However, despite these differences in design and research focus, there are striking similarities in the dimensions that are uncovered across studies.

Most importantly, in nearly all previous MD studies, there is a dimension associated with an oral/literate opposition (cf. Biber 2014). Linguistically, this opposition is realized as two fundamentally different ways of constructing discourse: clausal versus phrasal. That is, across studies the "oral" pole of this dimension consists of verb classes (e.g., mental verbs, communication verbs), grammatical characteristics of verb phrases (e.g., present tense, progressive aspect), and modifiers of verbs and clauses (e.g., adverbs and stance adverbials). Interestingly, these "oral" features also include dependent clauses that function as clausal constituents, including adverbial clauses and finite complement clauses. In contrast, the "literate" pole usually consists of phrasal devices that function as elements of noun phrases, especially nouns, nominalizations, attributive adjectives, and prepositional phrases. Functionally, this dimension is interpreted as distinguishing between a personal/involved focus (personal stance, interactivity, and/or real-time production features) versus informational focus. And in nearly every case, this parameter is the first dimension identified by the statistical factor analysis (i.e., it is the most important factor, accounting for the greatest amount of shared variance).

Table 2 summarizes the composition of this oral/literate dimension in selected studies of English discourse domains. It is perhaps not surprising that Dimension 1

Table 3 Narrative dimensions in MD studies of particular discourse domains in English

Discourse domain	Linguistic features defining the narrative dimension	Register pattern along the dimension
University spoken and written registers; Biber (2006)	3rd person pronouns, human nouns, communication and mental verbs, past tense VERSUS concrete and quantity nouns	office hours, study groups VERSUS textbooks, course packs, institutional writing
Conversational text types; Biber (2008a)	past tense, 3rd person pronouns, communication verb + *that*-clause VERSUS present tense	narrative conversations VERSUS other conversations
Academic research articles across disciplines; Gray (2013)	past tense verbs, perfect aspect, communication verbs, 3rd person pronouns, time adjectives, etc. VERSUS technical nouns, passive voice verbs	history / political science / applied linguistics VERSUS theoretical / quantitative physics

in the original 1988 MD analysis was strongly associated with the oral/literate opposition, given that the corpus in that study ranged from spoken conversational texts to written expository texts. For the same reason, it is somewhat predictable that a similar dimension would have emerged from the study of spoken and written registers in world English varieties (Xiao 2009), and in the study of eighteenth-century general written and speech-based registers (Biber 2001).

However, it is more surprising that restricted comparisons of spoken and written registers would uncover a first dimension with a similar set of co-occurring linguistic features, associated with a similar opposition between oral and informational-literate registers, such as the studies of university spoken and written registers (Biber 2006), elementary school registers (Reppen 2001), and English as a second language (ESL) spoken and written exam responses (Biber, Gray, and Staples 2016).

The most surprising finding here is the existence of a similar first dimension in MD studies of registers from a single mode. Those include studies focused exclusively on spoken registers (e.g., call center interactions and conversations, Friginal 2009) as well as those focused exclusively on written registers (e.g., legal registers and research articles, Goźdź-Roszkowski 2011). In all of these cases, the linguistic composition of Dimension 1 is surprisingly similar, generally opposing verbs, dependent clauses, pronouns, and interpersonal features versus nouns and phrasal noun modifiers.

The second linguistic parameter that has emerged in all MD studies is a dimension associated with narration. Linguistically, this dimension is consistently defined by features like past tense verbs, 3rd person pronouns, human nouns, temporal adverbs, and communication verbs. In terms of register differences, this dimension distinguishes narrative, time-organized descriptions of past-time events versus all other registers. Table 3 summarizes the narrative dimensions across a few MD studies of English.

4. Universal Dimensions in Fictional Novels

Two previous MD studies have focused specifically on the discourse domain of fictional novels: Biber's (2008b) study of novels in the nineteenth and twentieth centuries, and Egbert's (2012) study of nineteenth-century novels. The earlier study was based on 185 nineteenth- and twentieth-century novels collected from the Longman Corpus Network and from Project Gutenberg (comprising approximately 8.5 million words). Egbert's study then extended the earlier analysis by adding 100 fiction texts written by ten of the most famous nineteenth-century fiction authors (e.g., Louisa May Alcott, Charles Dickens, Henry James, Herman Melville, Mark Twain). Ten complete texts were collected for each author, making a total of approximately 10 million words.

Separate MD analyses were conducted in the two studies, with three factors extracted in the 2008 study and four factors extracted in the 2012 study. Table 4 shows that two of those dimensions are highly similar between the two analyses: Dimension 1, representing a basic oral/literate opposition, and Dimension 3, representing a 'narrative'/'non-narrative' opposition. Thus, the first dimension in both analyses shares many of the linguistic characteristics of the 'oral' versus 'literate' dimensions uncovered in other MD analyses, including verbs, adverbials, pronouns, and finite dependent clauses co-occurring as 'oral' linguistic features as opposed to nouns, attributive adjectives, and prepositional phrases co-occurring as 'literate' linguistic features. And the third dimension in both analyses shares many of the linguistic characteristics of the 'narrative' versus 'non-narrative' dimensions uncovered in other MD analyses, including past tense verbs and 3rd person pronouns as co-occurring 'narrative' features as opposed to present tense verbs.

At the same time, though, there are differences between the two analyses. For example, Dimension 1 in the Biber (2008b) study includes present tense, communication verbs, and 1st and 2nd person pronouns among the "oral" group of co-occurring features, while those features are grouped on to Dimension 3 in the Egbert (2012) study. These differences reflect the fact that 'oral' discourse in fictional novels (positive Dimension 1 characteristics) is also often dialogue and therefore 'non-narrative' discourse (negative Dimension 3). As a result, some interactive linguistic features tend to co-occur with present tense verbs marking non-narrative discourse. (In fact, 2nd person pronouns co-occur with present tense verbs on Dimension 3 in both MD analyses.)

Dimension 1 in the 2008 analysis is interpreted as 'Interactional (dialogue) versus informational (prose) focus.' This interpretation reflects the fact that the positive set

Table 4 The oral/literate dimensions and narrative-non-narrative dimensions in MD studies of fictional novels in English

Biber (2008b)	**Dimension 1: Interactional/involved versus informational focus**
	Features with positive loadings: – verbs: present tense – common verbs: mental, communication, pro-verb *do*, copula *be* – pronouns: 1st person, 2nd person – modals: possibility, necessity, prediction – adverbials: certainty – *that*-clauses: controlled by likelihood verbs, controlled by certainty verbs, controlled by communication – verbs, controlled by other mental/stance verbs – *that*-omission – *to*-clauses: controlled by desire verbs
	Features with negative loadings: – nouns: total nouns, place nouns, concrete nouns – prepositional phrases – adjectives: attributive – word length, type/token ratio – adverbials: place
	Dimension 3: Past versus present orientation (time and person)
	Features with positive loadings: – past tense verbs – perfect aspect verbs – 3rd person pronouns
	Features with negative loadings: – 2nd person pronouns – present tense verbs – contractions – nouns

of features are very similar to the set of interactive and involved co-occurring features found in face-to-face conversation. As a result, these features are common in novels that rely heavily on dialogue among characters. In contrast, the negative features are typical of informational written registers, and thus in novels, these features are characteristic of novels that rely heavily on descriptive or narrative prose. The functional interpretation of Dimension 1 in the 2012 study focuses on a slightly different opposition: thought presentation versus (informational) description.

Despite these differences, the similarities between the two analyses are strong in that both have strong dimensions associated with the 'oral'/'literate' opposition as well as the 'narrative'/'non-narrative' opposition. That is, Dimension 1 in both studies follows the pattern of the 'oral'/'literate' dimension in other studies, opposing a

Table 4 (*continued*)

Egbert (2012)	Dimension 1: Thought presentation versus description
	Features with positive loadings: – verbs: mental verbs, existence verbs, perfect aspect, possibility modals – pronouns: indefinite, *it* – adverbials: stance adverbials, general adverbs – dependent clauses: stance verb + *that*-clause; desire verb + *to*-clause; *WH*-clauses, *that*-clauses with complementizer deletion Features with negative loadings: – nouns – attributive adjectives – prepositional phrases
	Dimension 3: Narration versus dialogue [polarity reversed]
	Features with positive loadings: – past tense verbs – simple occurrence verbs – 3rd person pronouns Features with negative loadings: – present tense verbs – *have* as main verb – communication verbs – modal verbs – 1st and 2nd person pronouns – *WH*-questions

set of verbs, adverbs, and dependent clauses versus nouns, attributive adjectives, and prepositional phrases. And Dimension 3 in both studies follows the patterns of 'narrative'/'non-narrative' dimensions in previous studies, opposing past tense verbs and 3rd person pronouns versus present tense verbs.

It turns out that these dimensions are highly useful for distinguishing among the styles of fiction authors. Egbert (2012) focuses specifically on nineteenth-century authors. For example, with respect to Dimension 1, authors like James, Alcott, and Twain tend to rely on a clausal "oral" style, while authors like Melville and Kipling tend to rely on a more phrasal "literate" style. There are also important differences with respect to Dimension 3. For example, Hawthorne and Melville prefer a more narrative prose style, while Alcott and Twain rely much more heavily on present-time, interactive dialogue.

Biber (2008b) similarly identifies important differences among the prose styles of fiction authors in both the nineteenth and twentieth centuries. For example, with respect to Dimension 1, the children's novel *The House on Pooh Corner* by A. A Milne is extremely 'oral,' with frequent features reflecting personal involvement and interactivity. But these characteristics are not at all restricted to children's literature. For example, Forster's *A Room with a View* is nearly as marked for positive Dimension 1 features as *Pooh*.

1. Excerpt from *A Room with a View*
   E.M. Forster
   [present tense verbs, modals, and 1st and 2nd person pronouns are shown in italics]

   "*I want* so to see the Arno. The rooms the Signora promised *us* in her letter *would have looked* over the Arno. The Signora had no business to do it at all. Oh, it *is* a shame!"
   "Any nook *does* for *me*," Miss Bartlett continued; "but it *does seem* hard that *you shouldn't have* a view."
   Lucy felt that she had been selfish. "Charlotte, *you mustn't spoil me*: of course, *you must look* over the Arno, too. I meant that. The first vacant room in the front—"
   —"*You must have* it," said Miss Bartlett, part of whose travelling expenses were paid by Lucy's mother—a piece of generosity to which she made many a tactful allusion.
   "No, no. *You must have* it."
   "*I insist* on it. Your mother *would never forgive me*, Lucy."

At the other (informational) extreme of Dimension 1, we also find both children's literature (e.g., *The Tale of Peter Rabbit*, Beatrix Potter) and adult fiction (e.g., *Ulysses*, James Joyce). Surprisingly, the most "informational" novel in our corpus is a children's novel: Henry Williamson's *Tarka the Otter*, illustrated in Text Excerpt 2.

2. Excerpt from *Tarka the Otter*
   Henry Williamson
   [Nouns, attributive adjectives, and prepositional phrases are shown in italics]

   She ran *over the bullock's drinking-place* and passed *through willows to the meadow*, seeking *old dry grasses and mosses under the hawthorns growing by the mill-leat*, and gathering them *in her mouth with wool pulled from the over-arching blackberry brambles* whose *prickles* had caught *in the fleeces of sheep*. She re-

> turned *to the river bank* and swam *with her webbed hind-feet to the oak tree*, climbed *to the barky lip of the holt*, and crawled within. *Two yards inside* she strewed her *burden on the wood-dust*, and departed *by water for the dry, sand-coloured reeds of the old summer's growth* which she bit off, frequently pausing to listen. *After several journeys* she sought *trout* by cruising *under water along the bank*, and *roach* which she found by stirring up *the sand and stones of the shallow* wherein they lurked.

Dimension 3 in the 2008 study also identifies important linguistic differences among novels. For example, children's novels like *Peter Rabbit* and *Tarka the Otter* (see excerpt 2) have a strong past narrative orientation (with large positive scores on Dimension 3). Adult novels like Virginia Woolf's *To the Lighthouse* also have a similar reliance on 3rd person past-time discourse, as illustrated in Text Sample 3:

3. Excerpt from *To the Lighthouse*
   Virginia Woolf
   [3rd person pronouns, past tense, and perfect aspect verbs shown in italics]

> Nothing *happened*. Nothing! Nothing! as *she leant* her head against Mrs Ramsay's knee. And yet, *she knew* knowledge and wisdom *were stored* in Mrs Ramsay's heart. How then, *she had asked* herself, *did one* know one thing or another thing about people, sealed as *they were*? Only like a bee, drawn by some sweetness or sharpness in the air intangible to touch or taste, *one haunted* the dome-shaped hive, *ranged* the wastes of the air over the countries of the world alone, and then *haunted* the hives with their murmurs and their stirrings; the hives which *were* people. Mrs Ramsay *rose*. Lily *rose*. Mrs Ramsay *went*. For days there *hung* about *her*, as after a dream some subtle change is felt in the person one *has dreamt* of, more vividly than anything *she said*, the sound of murmuring and, as *she sat* in the wicker arm-chair in the drawing-room window *she wore*, to Lily's eyes, an august shape; the shape of a dome.

At the other extreme, many novels adopt a present-time focus, which tends to co-occur with 2nd person pronouns and contractions. It would be easy to suppose that this fictional style occurs in novels that include extensive dialogues among characters, given that actual face-to-face conversations also employ these same co-occurring linguistic features. However, it turns out that there are additional factors associated with this discourse style in novels.

In some novels, fictional dialogue does employ a present-time focus, similar to the norm in actual conversation. Thus, consider Text Sample 1 (above) and the following interaction from Joseph Heller's *Catch-22*:

4. Dialogue from *Catch-22*
   Joseph Heller
   [present tense verbs shown in italics]

   "*Are* you crazy?" […] "I *suppose* you just *don't care* if you *kill* yourself, *do* you?"
   "It*'s* my self"
   "I *suppose* you just *don't care* if you *lose* your leg, *do* you?"
   "It*'s* my leg"
   "It certainly *is* not your leg!" […] "That leg *belongs* to the U.S. government."

Surprisingly, though, it is at least as common for fictional dialogue to occur with frequent past tense and perfect aspect verbs, being quite different from typical face-to-face conversation in this regard. Fiction authors rely on these past-time features in dialogue because they use dialogue to move the narrative story forward, and thus characters often report past events in their interactions (see the text excerpts in [5]).

5. Short dialogues from three novels, illustrating the dense use of past tense and perfect aspect in fictional interpersonal interactions.
   [past tense and perfect aspect verbs shown in italics]

   The Insidious Dr. Fu-Manchu
   Sax Rohmer

   "Ever *seen* one like it?" he *asked*.
   "Not exactly," I *confessed*. "It appears to *have been* deeply *cauterized*."
   "Right! Very deeply!" he *rapped*. "A barb steeped in the venom of a hamadryad
   went in there!"
   […]
   "There's only one treatment," he *continued*, rolling his sleeve down again,
   "and that's with a sharp knife, a match, and a broken cartridge.
   I *lay* on my back, raving, for three days afterwards, in a forest that *stank* with
   malaria, but I should *have been* lying there now if I *had hesitated*.
   Here's the point. It *was* not an accident!"

   Masters of Space
   E.E. Smith and E. Everett Evans

   "Mr. Ashby, *did* you have your interspace rigs set?"
   "No, sir. I *didn't think* of it, sir."

> "Doctor Cummings, why *weren't* yours out?"
> "I *didn't think* of such a thing, either—any more than you *did*," Sandra *said*.

> *The Highest Treason*
> Randall Garrett

> "The Board of Strategy *asked* me to tell you," Tallis *continued*. "After all, my recommendation *was* partially responsible for the decision." [...] "It *was* a hard decision, Sepastian—you must realize that. We *have been* at war with your race for ten years now. We *have taken* thousands of Earthmen as prisoners, and many of them *have agreed* to co-operate with us [...]"

In contrast, it turns out that many novels—especially modern novels—employ a present-time style for narrative and descriptive prose, and as a result these novels have large negative scores for Dimension 2. Excerpt [6] illustrates this style:

6. *The Middleman*
Olen Steinhauer
[Present tense verbs in italics]

> All day I *sit* by the lime green swimming pool, sun-screened so I won't turn black, going through my routine of isometrics while Ransome's indios *hack* away the virgin forests. Their hate is intoxicating. They *hate* gringos—from which my darkness *exempts* me—even more than Gutierrez. They *hate* in order to keep up their intensity.
>
> I *hear* a litany of presidents' names, Hollywood names, Detroit names—Carter, *chop*, Reagan, *slash*, Buick, *thump*—*bounce* off the vines as machetes *clear* the jungle greenness.
>
> We spoke a form of Spanish in my old Baghdad home. I always *understand* more than I *let on*.

For these same reasons, Dimensions 1 and 3 are not strongly related, having a Pearson correlation of only r = -0.19 for the 185 novels analyzed in the 2008 study. Figure 2 plots 15 of these novels in the two-dimensional space, illustrating this lack of a strong relationship. Although novels occupy much of the space, there is a noticeable absence of novels that have large scores for both dimensions. For example, *To the Lighthouse* has the largest positive score for Dimension 3 ("Past orientation") but a score near 0.0 for Dimension 1; *The Middleman* similarly has the largest negative score for Dimension 3 ("Present orientation") but a score near 0.0 for Dimension 1. And we find a very similar pattern with respect to Dimension 1: *Room with a View* and *The House on Pooh Corner*

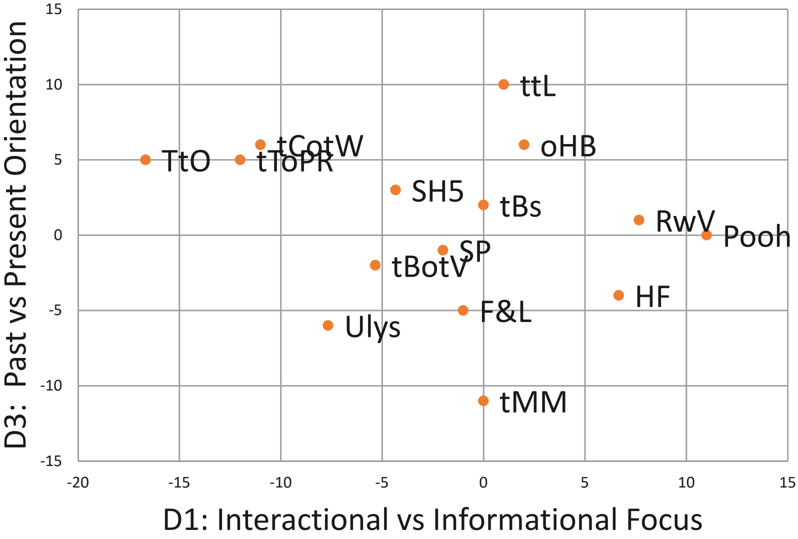


Fig. 2 Dimension 1 versus Dimension 3 scores for selected novels. Key: Pooh = *House on Pooh Corner*; RwV = *A Room with a View*; HF = *Huckleberry Finn*; oHB = *Of Human Bondage*; ttL = *To The Lighthouse*; tBs = *The Borrowers*; tMM = *The Middleman*; F&L = *Fear and Loathing In Las Vegas*; SP = *Mr Sammler's Planet*; SH5 = *Slaughterhouse-Five*; tBotV = *The Bonfire Of The Vanities*; Ulys = *Ulysses*; tCotW = *The Call of the Wild*; tToPR = *The Tale Of Peter Rabbit*; TtO = *Tarka the Otter*. (Biber/Egbert, CC BY)

both have large positive scores for Dimension 1 ("Interactional/involved focus") but scores near 0.0 for Dimension 3. The only exception to this generalization is *Tarka the Otter*, which has a very large negative score for Dimension 1 ("Informational focus") coupled with a moderately large positive score for Dimension 3 ("Past orientation"). In general, though, these two parameters are largely unrelated, indicating that authors can choose between a highly interactive/involved style versus a highly informational style independently of their choice between a highly narrative style versus highly present-time oriented style.

5. Conclusion

Both of the previous MD analyses of fictional novels identify an additional dimension that is peculiar to the discourse domain of fictional novels. In the 2008 study, this dimension was interpreted as "Concrete actions/events versus abstract description," opposing phrasal verbs, activity verbs, progressive aspect, concrete nouns, and place adverbials versus nominalizations, mental nouns, abstract nouns, long words, and attributive adjectives. And similarly, the 2012 study identified a third dimension with almost the same interpretation: "Abstract exposition versus concrete action." Authors like Sammler (in the 2008 study) and Kipling (in the 2012 study) relied heavily on the linguistic features associated with concrete action, while authors like Eliot relied heavily on the features of abstract description/exposition. Surprisingly, Twain's *Adventures of Huckleberry Finn* was also marked for the dense use of "abstract description" features.

This third dimension does not have a direct counterpart in MD studies of other discourse domains. Rather, it reflects a functional distinction that is important for distinguishing among fictional novels: those stories that are action-oriented versus novels with a much stronger focus on the description of people and places, which often also includes commentary on motives, actions, or society in general. In this regard, these MD studies of fictional novels are similar to all previous MD studies, in that they have all identified dimensions of variation that are specialized to a discourse domain or language. These specialized dimensions reflect the particular communicative priorities of each language/culture or domain of use.

From both theoretical and methodological perspectives, it is not surprising that each MD analysis would uncover specialized dimensions that are peculiar to a given language and/or discourse domain. After all, each of these studies differs with respect to the set of registers represented in the corpus for analysis, and the set of linguistic features included in the analysis. Given those differences, it is reasonable to expect that the parameters of variation that emerge from each analysis will be fundamentally different. And to some extent, this expectation is met, with specialized dimensions emerging in nearly all MD analyses.

However, given this background, the existence of universal dimensions of variation that emerge in nearly all MD studies is quite unexpected. Two of these dimensions are especially important, regardless of the discourse domain: a dimension associated with 'oral' versus 'literate' discourse, and a dimension associated with narrative discourse.

The robustness of narrative dimensions across languages and discourse domains indicates that this rhetorical mode is basic to human communication, whether in speech or in writing. Rhetoricians and discourse analysts have long argued for the central role of narration in communication. MD studies confirm that claim, showing the importance of this rhetorical mode in virtually all discourse domains (spoken and written; interpersonal and informational; etc.). And, as we show in Section 4 above, this functional parameter is also important for distinguishing among fictional novels, both in the nineteenth century and in the twentieth century.

But the most surprising pattern discovered through MD analysis is the oral/literate opposition, which emerges as the very first dimension in nearly all MD studies. In studies based on general corpora of spoken and written registers, this dimension clearly distinguishes between speech and writing. However, other studies show that this is not a simple opposition between the spoken and written modes. In fact, this dimension emerges consistently in studies restricted to only spoken registers, as well as studies restricted to written registers.

In terms of communicative purpose, the 'oral' registers characterized by this dimension focus on personal concerns, interpersonal interactions, and the expression of stance. In contrast, 'literate' registers focus on the presentation of propositional information, with little overt acknowledgment of the audience or the personal feelings of the speaker/writer. Linguistically, this first dimension opposes two discourse styles: an 'oral' style that relies on pronouns, verbs, adverbs, versus a 'literate' style that relies on nouns and nominal modifiers. The oral style relies on clauses to construct discourse—including a dense use of dependent clauses. In contrast, the complexity of the literate style is phrasal.

It turns out that this same opposition is fundamentally important for distinguishing among the styles of fictional novels. Authors like Melville, and even some children's novels like *Tarka the Otter*, are notable for their dense reliance on an informational style employing phrasal grammatical features. At the other extreme, novels like *House on Pooh Corner* and *Room with a View* are marked by their highly 'oral' style relying on verbs, pronouns, adverbs and dependent clauses. As such, we have shown here how variation in the discourse domain of fictional novels is patterned in similar ways to all other discourse domains in that it reflects the two general functional parameters of 'oral'/'literate' discourse and 'narrative'/'non-narrative' discourse, while at the same time being organized with respect to additional functional-linguistic dimensions.

References

Biber, Douglas. 1986. "Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings." *Language* 62 (2): 384–414.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Perspective*. Cambridge: Cambridge University Press.

Biber, Douglas. 2001. "Dimensions of Variation among Eighteenth-Century Speech-Based and Written Registers." In *Multi-Dimensional Studies of Register Variation in English*, edited by Susan Conrad and Douglas Biber, 200–14. London: Longman.

Biber, Douglas. 2006. *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Biber, Douglas. 2008a. "Corpus-Based Analyses of Discourse: Dimensions of Variation in Conversation." In *Advances in Discourse Studies*, edited by Vijay Bhatia, John Flowerdew, and Rodney H. Jones, 100–14. London: Routledge.

Biber, Douglas. 2008b. "Using Corpus-Based Analysis to Study Fictional Style: A Multi-Dimensional Analysis of Variation among and within Novels." Invited plenary lecture, International Society for the Empirical Study of Literature, University of Memphis.

Biber, Douglas. 2014. "Using Multi-Dimensional Analysis to Explore Cross-Linguistic Universals of Register Variation." *Languages in Contrast* 14 (1): 7–34.

Biber, Douglas. 2019. "Multi-Dimensional Analysis: A Historical Synopsis." In *Multi-Dimensional Analysis: Research Methods and Current Issues*, edited by Tony Berber-Sardinha and Marcia Veirano Pinto, 11–26. London: Bloomsbury.

Biber, Douglas, Bethany Gray, and Shelley Staples. 2016. "Predicting Patterns of Grammatical Complexity across Language Exam Task Types and Proficiency Levels." *Applied Linguistics* 37 (5): 639–68.

Egbert, Jesse. 2012. "Style in Nineteenth Century Fiction: A Multi-Dimensional Analysis." *Scientific Study of Literature* 2 (2): 167–98.

Friginal, Eric. 2009. *The Language of Outsourced Call Centers*. Amsterdam: John Benjamins.

Goźdź-Roszkowski, Stanislaw. 2011. *Patterns of Linguistic Variation in American Legal English: A Corpus-Based Study.* Frankfurt am Main: Peter Lang.

Gray, Bethany. 2013. "More than Discipline: Uncovering Multi-Dimensional Patterns of Variation in Academic Research Articles." *Corpora* 8 (2): 153–81.

Reppen. Randi. 2001. "Register Variation in Student and Adult Speech and Writing." In *Multi-Dimensional Studies of Register Variation in English*, edited by Susan Conrad and Douglas Biber, 187–99. London: Longman.

Xiao, Richard. 2009. "Multidimensional Analysis and the Study of World Englishes." *World English* 28 (4): 421–50.

Dutch Strong and Weak Pronouns as a Stylistic Marker of Literariness

Andreas van Cranenburgh

Abstract Certain languages exhibit distinctions between strong and weak forms of pronouns. Linguists have attempted to explain the preferences for the different forms of pronouns in terms of pragmatic factors, specifically discourse salience and contrast. These factors only partially account for the variation observed. In this article we propose to add another factor, style. We investigate the case of Dutch with a corpus of literary novels. We present quantitative results in the form of corpus frequencies and correlations with literary prestige, as well as qualitative judgments from a manual analysis, and finally a statistical analysis of coreference annotations. This complements the linguistic studies, which have focused on testing explanations in specific contexts in controlled experiments, without testing the relevance of those explanations in naturalistic data. Our results suggest that style is a prominent factor in the strong/weak pronoun distinction, since the linguistic explanations have limited predictive power, while our corpus study shows that a high proportion of strong pronouns is associated with literary prestige and Dutch authorship.

Keywords strong and weak pronouns, literariness, Dutch

1. Introduction

What makes a literary novel *literary*? This is a question without an empirically satisfying answer. Various explanations have been suggested. Adherents of Bourdieu claim that the cultural capital of critics and publishers determines the perceived prestige of novels. Proponents of Kantian aesthetics contend that the demarcation of art rests on (inter)subjective, normative value-judgments (i.e., we expect others to agree about the greatness of art, as opposed to matters of taste which are purely subjective). A third group, the formalists, make an even bolder claim, namely that *literariness* is an intrinsic, objective property of texts; proposed mechanisms are defamiliarization and estrangement. Literary language contrasts itself with everyday language by standing out.

van Cranenburgh, Andreas: Dutch Strong and Weak Pronouns as a Stylistic Marker of Literariness, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 217–234. DOI: https://doi.org/10.17885/heiup.1157.c19373

While this paper makes no commitment to any of these explanations, our work most closely aligns with the last explanation, since we will compare objective textual features of texts and correlate them with perceptions of literary prestige. This work is part of a larger research project, The Riddle of Literary Quality,[1] which set out to investigate textual features that may be correlated with literary prestige. To this end, a large reader survey was held (Koolen et al., 2020). Readers from the general public rated 401 recent, best-selling Dutch-language novels (both original and translated) on a Likert scale of 1–7 (not at all literary to very literary). This allows us to estimate the relation between perceptions of literariness and stylistic markers in the texts. Previous work has already shown that literary prestige can be predicted from textual features to a large extent (van Cranenburgh and Bod 2017; van Cranenburgh et al. 2019). The present paper is not about improving on these predictive models, but zooms in on one specific linguistic aspect (the strong/weak pronoun distinction) which turns out to have a surprising correlation with literary prestige (van Cranenburgh et al. 2019), with the aim of better understanding this specific stylistic aspect; i.e., we focus on explanation, not prediction (Breiman 2001).

The Dutch language (along with other languages) has full and reduced versions of some of its personal pronouns (see Table 1). Full pronouns such as *jij* ('you') are also called emphatic or strong, while in Dutch the reduced pronouns such as *je* ('you') are weak pronouns; other types of reduced pronouns such as clitics in Romance languages are grammatically more restricted. On the one hand the distinction follows linguistic rules and cues related to contrast and salience of discourse referents (Bresnan 1998; Kaiser 2011). On the other hand the distinction can also be a stylistic choice, when both options are available. Weak pronouns are more informal and are required in fixed expressions such as *dank je* ('thank you'), whereas strong pronouns can be used for emphasis or to refer to a less salient referent; strong pronouns are required when expressing contrast or in comparisons such as *hij en zij* ('he and she').

This paper addresses the following research questions:

1. Can we explain the large proportion of strong pronouns in some highly literary novels?
2. To what extent is the pronoun form due to a stylistic choice rather than a grammatical preference or requirement?

The rest of this paper is structured as follows. Section 2 goes into the linguistic background of the strong/weak pronoun distinction. Our main results consist of a corpus study (section 3), a study of contrast and preference (section 4), and a statistical model based on coreference annotations (section 5). We end with a discussion of theoretical implications (section 6).

[1] https://literaryquality.huygens.knaw.nl.

2. Linguistic Background on Dutch Strong/Weak Pronouns

We first introduce the pronoun system of Dutch and enumerate contexts in which either form is required or preferred. We continue by discussing linguistic theories put forward to explain the choice of pronouns.

2.1 Strong/Weak Pronouns as Described by Reference Grammars

See Table 1 for an overview of Dutch personal pronouns. Some Dutch pronouns have strong and weak forms. These pronouns carry the same meaning, but either the strong or weak variant may be obligatory or preferred in certain contexts. At other times, it is a matter of free choice, i.e., a matter of style.

Table 1 Personal pronouns in Dutch. Pronouns with a common strong and weak counterpart are shown in italics; a comma indicates a subject/object distinction; the forms in parentheses are not common in written language

	Strong	**Weak**
1st sg	ik, *mij*	-, *me*
2nd sg	*jij*, *jou*	*je*
3rd sg fem	*zij*, haar	*ze*, (d'r)
3rd sg masc	hij, hem	(ie, 'm)
3rd sg neut	het	('t)
1st pl	*wij*, ons	*we*, -
2nd pl	jullie	-
3rd pl	*zij*, *hen/hun*	*ze*

The Dutch grammar Haeseryn et al. (1997) describes a range of properties of strong and weak pronouns. The most important feature is that phonologically, strong pronouns are often stressed (e.g., when used for emphasis or to contrast a referent with another referent), but weak pronouns are always unstressed. Strong pronouns tend be restricted for persons or concepts treated as persons, while weak pronouns readily refer to both persons and objects. The grammatical contexts where strong pronouns are obligatory are as follows (* marks an ungrammatical phrase; examples adapted from Haeseryn et al. 1997, 252–55):

(1) Comparisons:
ik ben rijker dan *jij*, *dan *je*
'I am richer than *you* (strong), *than *you* (weak)'

(2) Conjunctions of pronouns:
hij en *zij*, *hij en *ze*
'he and *she* (strong), *he and *she* (weak)'

(3) Certain oblique arguments (i.e., neither subject nor object):
voor *hen* die …, *voor *ze* die …
'for *those* (strong) who …', *'for *those* (weak) who …'

Conversely, the following contexts require a weak pronoun:

(4) Idioms:
dank *je*, *dank *jou*
'thank *you* (weak)', *'thank *you* (strong)'

(5) Generic you:
je weet maar nooit! **jij* weet maar nooit!
'*you* (weak) never can tell!' *'*you* (strong) never can tell'

Generally, strong pronouns are preferred in written language, while weak pronouns are preferred in spoken language, as they are considered more informal. There is a tendency to write strong pronouns which would be weak pronouns in spoken language, and conversely, to pronounce strong pronouns when weak pronouns are read. Strong pronouns sound unnatural when repeated in the same sentence or context. Use of repeated strong pronouns is associated with non-native speakers, since they may be unaware of this unwritten rule. While academic grammars such as Haeseryn et al. (1997) and Donaldson (2008) discuss strong and weak pronouns, the subtleties of their usage are not discussed in most textbooks used by second language learners.

2.2 Linguistic Explanations for the Distribution of Strong/Weak Pronouns

Before going into the linguistic research on the strong/weak pronoun distinction in Dutch, it is helpful to look at research on the production of referring expressions in general. Arnold and Zerkle (2019) investigate why speakers might produce pronouns rather than descriptive noun phrases. Pronouns are strictly less informative than noun phrases, so what other reasons explain their use? The two main explanations they consider are pragmatic and rational factors. The pragmatic model argues that the choice to produce a pronoun can be explained by the speaker's cognitive status of the referent.

Concretely, there is an accessibility hierarchy spanning more or less salient referents. Pronouns are preferred for more salient, recent and frequent referents. The rational model argues that speakers optimize the balance between informativeness and efficiency, with shorter expressions being preferred if they do not cause confusion. Arnold and Zerkle (2019) conclude that the accessibility model only explains part of the variation observed, while efficiency cannot be the primary explanation either. Note also that these theories presuppose that the choice can be explained by rules or efficiency, which is not a given.

Research on strong and weak pronouns also considers the salience hierarchy. Kaiser (2011) summarizes the range of options for referring expressions as follows:

null > reduced pronoun > full pronoun > demonstrative > noun phrases … etc.
most salient referent less salient referent

Kaiser (2011) looks at Dutch specifically. The Dutch language does not have null pronouns except in very limited cases such as imperatives, but it does have reduced (specifically, weak) pronouns, and this distinction is made in both speaking and writing. Weak pronouns are therefore the most salient option available.

In addition to the salience explanation, Kaiser (2011) considers the explanation that strong pronouns are used to express contrast, i.e., the situation where there are multiple competing discourse referents, or where there is a switch to a new topic. She presents data from sentence completion as well as eye tracking experiments. Participants are manipulated using several conditions to test the salience and contrast explanations. The results show that salience does not explain the strong/weak distinction, while it does predict the choice between pronouns and demonstratives. The presence of contrast does result in a marked preference for strong pronouns. However, this does not imply the reverse: that the use of a strong pronoun is likely due to contrast between salient alternatives. This is due to the experimental setup of Kaiser (2011), which has the goal of probing the possible role of referential properties in the strong/weak pronoun distinction; a fortiori, non-referential properties are not considered. Another limitation of the results is that only the strong and weak pronouns *zij/ze* are considered (since first and second person pronouns are not as referentially ambiguous), and only where they occur in subject position (to avoid parallelism effects).

In her general discussion, Kaiser (2011) concludes that the results fit into a form-specific multiple-constraints approach (i.e., there is not a single constraint which can explain the distinction), since uses of strong pronouns that do not express contrast are readily attested. She proposes a Gricean approach in which the use of a strong pronoun where a weak pronoun is also licensed provides an implicature that the strong pronoun was preferred for a reason, such as contrast. The implicature is then further defined to be context-dependent and possibly underspecified (there may not be a reason). While this account can accommodate any new observation, it does not seem to make any specific, testable predictions.

We can conclude that the linguistic theories underdetermine the data. While speakers are influenced by pragmatics and efficiency, these are not sufficient explanations. We contend that what is missing from these experimental results is a consideration for naturalistic data. The reported experiments create artificial conditions with the goal of testing preconceived hypotheses. This serves to demonstrate that these factors play a role but cannot establish that they are sufficient. We suspect that an overlooked factor is the possibility that strong/weak pronouns also exhibit a stylistic dimension. Especially in the cases where the choice between a strong and weak pronoun is not required or preferred for grammatical reasons, the aforementioned associations of informality and differences in tone may play a role in the selection of a strong rather than a weak pronoun.

3. Study 1: Corpus Frequencies and Correlations

We consider the frequencies of pronoun forms and their correlation with literary ratings. We first look at the frequency of pronouns in general, and then focus on the proportion of strong pronouns in particular.

3.1 Materials and Methods

The corpus consists of 401 contemporary Dutch-language novels by 217 different authors; both originally Dutch and translated novels are included in similar proportions. The novels are best-selling and include different genres, such as thrillers, romantic novels, and literary fiction. In a large survey, readers from the general public rated the literariness of the 401 novels on a seven-point Likert scale (not at all literary to very literary). Survey participants first indicated which novels on the list they had read, and then rated those novels based on the title and author. We use the mean rating per novel as a representative score. While the resulting ratings are ordinal, a sufficient amount of ratings (50–1,000 per book) were collected, and the variance was limited, showing that there was substantial consensus on the literary ratings (for more details, cf. van Cranenburgh et al. 2019).

The texts of the novels were cleaned and automatically parsed with the Dutch *Alpino* parser.[2] The size of the corpus is five million sentences comprising 52 million

[2] https://www.let.rug.nl/vannoord/alp/Alpino/

tokens. For our corpus linguistic study, we focus on cases where there may be a choice between weak and strong pronouns, so we only consider pronouns with both versions. We exclude pronouns where the weak version is not common in written language (the neuter pronoun *'t*, female pronoun *d'r*, male pronoun *'m*); we also exclude forms that are exclusively possessive (e.g., *mijn*, *jouw)* and reflexive pronouns (e.g., *mezelf*, *zich)*. This leaves us with the following regular expressions to identify the pronouns of interest (matched at word boundaries, case insensitively):

Strong: `(mij|jij|jou|zij|wij|hen|hun)`
Weak: `(me|je|ze|we)`

After collecting the frequencies of strong and weak pronouns, we determine the correlation with the mean literary rating for each novel.

3.2 Results

We first look at the overall frequency of both pronoun types; for example, more literary novels might focus more on ideas than people. See Figure 1 for the results. There is indeed a negative correlation of pronoun frequency and literariness.

Figure 2 shows the percentage of strong pronouns with respect to both types. In other words, we control for the total number of pronouns, which may differ per novel. Weak pronouns are much more frequent than strong pronouns. On average 82 percent of pronouns with strong and weak forms are weak across the 401 novels. Similarly, in a 700 million word reference corpus (Lassy Large) with edited text from various domains, 76.2 percent of such pronouns are weak.

While we have already found a strong correlation for pronoun frequency, Figure 2 shows that the strong/weak distinction yields a stronger correlation, suggesting that the distinction has a stylistic dimension. A possible explanation would be that weak pronouns are a proxy for informality. However, the result may also be due to more complicated discourse structures in literature which employ strong pronouns for contrast, or a higher frequency of grammatical constructs that require strong pronouns.

Additionally, we see a striking set of nine outliers in Figure 2 with a substantially higher proportion of strong pronouns (>30 percent). The outliers are listed in Table 2. Except for Mitchell, they are novels written by Dutch authors; except for Van Kooten, they are all literary fiction according to the publisher labels and are rated as highly literary by the survey respondents.

Figure 3 shows the distribution of the different pronoun forms in the outliers. The heatmap contrasts the relative frequency (expressed as percentage) of each form with the average relative frequency across the whole corpus of 401 novels with a threshold

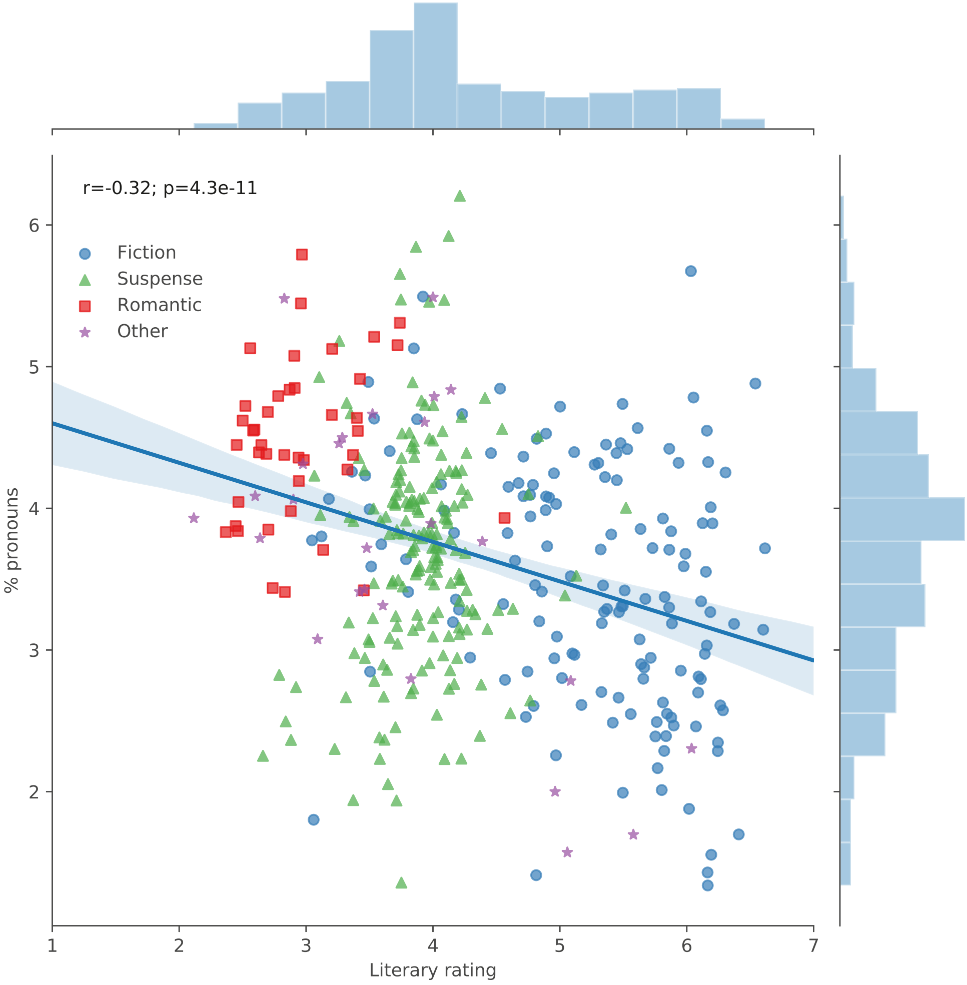


Fig. 1 Percentage of pronouns with respect to all words, correlated against literary ratings (van Cranenburgh, CC BY).

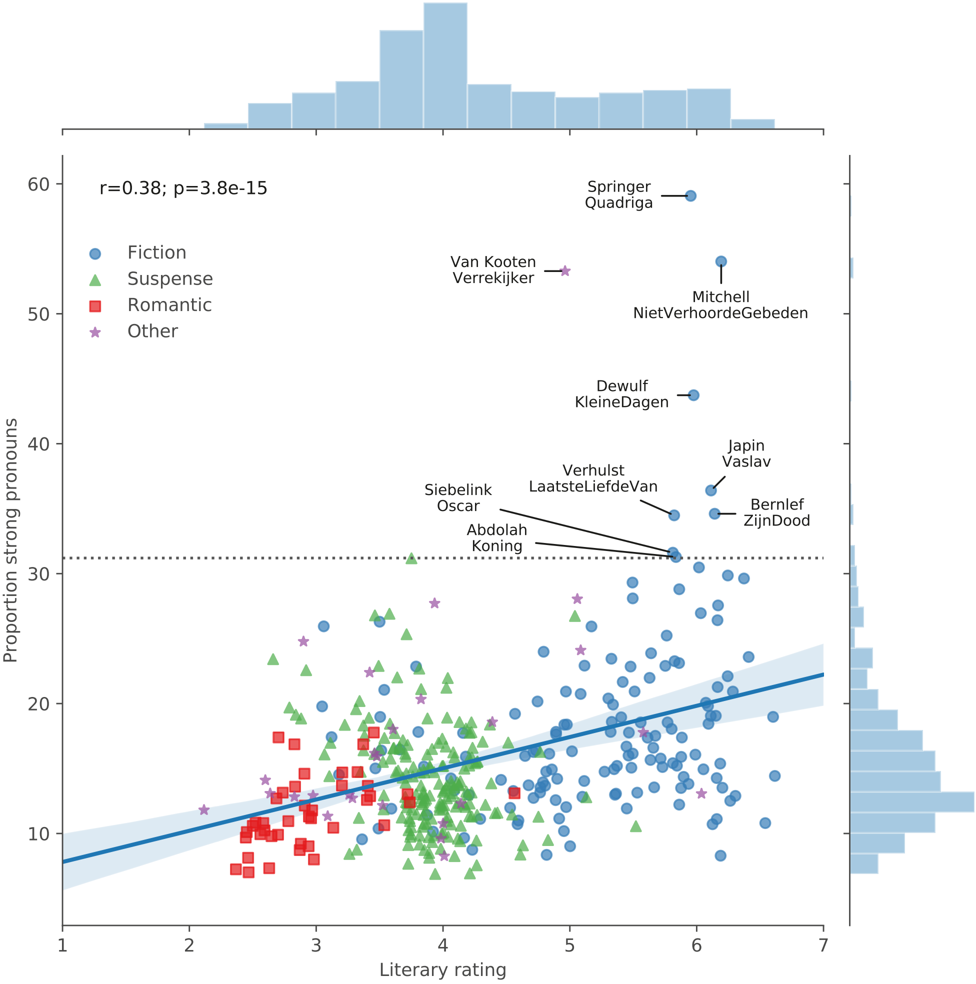


Fig. 2 Percentage of strong pronouns with respect to both pronoun types, correlated against literary ratings (van Cranenburgh, CC BY).

Table 2 Novels that are outliers with respect to the proportion of strong pronouns

	Strong %	Weak %	Both %	Strong prop.
Springer, Quadriga	1.69	1.17	2.85	59.1
Mitchell, The thousand autumns […]	0.84	0.71	1.55	54.0
Van Kooten, Verrekijker	1.07	0.93	2.00	53.3
Dewulf, Kleine dagen	1.57	2.02	3.59	43.7
Japin, Vaslav	1.22	2.13	3.34	36.4
Bernlef, De een zijn dood	1.03	1.94	2.97	34.6
Verhulst, De laatste liefde van […]	0.79	1.50	2.29	34.5
Siebelink, Oscar	0.83	1.80	2.63	31.6
Abdolah, Koning	0.75	1.64	2.39	31.3

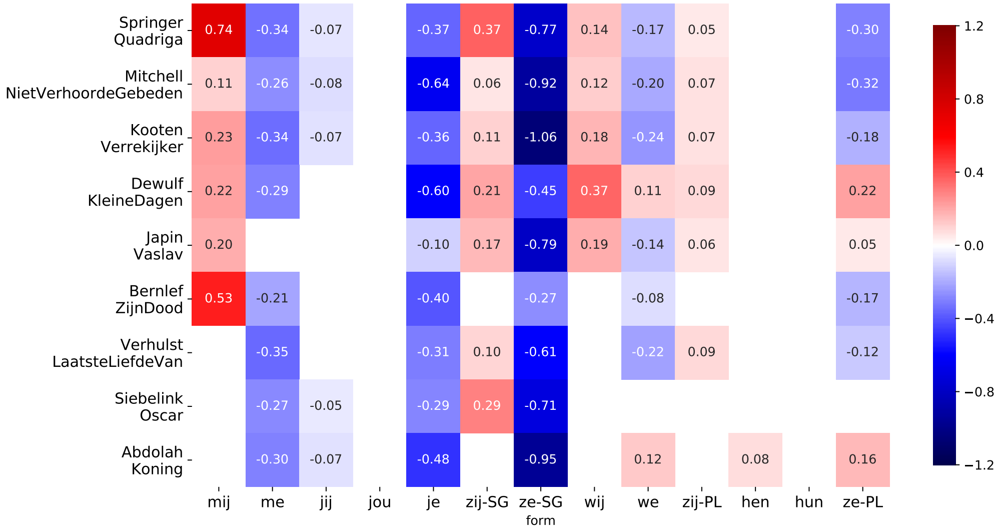


Fig. 3 Heatmap showing the divergence in frequency of the different pronoun forms across the outliers compared to the whole corpus (van Cranenburgh, CC BY).

of 0.05 absolute difference in relative frequency. We used part-of-speech (POS) tags to restrict the counts to occurrences of personal and personal/reflexive pronouns (specifically, VNW(pers,…) and VNW(pr,…); this excludes possessive pronouns and the archaic verb *zij* which were not excluded with the query used for Figure 2). Since the pronoun *zij/ze* can be both feminine/singular and plural, these are listed separately. We find that certain forms do not differ appreciably (*jou*, *hun*) in any of the novels, while this holds for the forms *mij* and *wij* in three and four novels, respectively. A few strong pronouns are shown in dark red, indicating that they are much more frequent than average; many more weak pronouns are shown in dark blue, indicating that they are much less frequent than average in these outlier novels.

Now that a corpus study has revealed these outlier novels, we will attempt to explain their outlier status using the linguistic contexts in which the pronouns are used.

4. Study 2: Coding Contrast and Preference

We now take a closer look at the outliers. While a comprehensive study of the observed variation would need to contrast the outliers with non-outliers, we focus here on investigating why the outliers display such an exceptionally high proportion of strong pronouns. We will consider whether any of the proposed reasons for the use of strong and weak pronouns apply: contexts in which one or the other is grammatically obligatory and contexts which may lead one or the other to be preferred.

4.1 Materials and Methods

We annotated the relevant strong/weak pronouns in the first 100 sentences of each of the nine outlier novels, resulting in 356 annotated pronouns. We used the following coding scheme:

- Pronoun form (strong versus weak)
- In this sentence, is the given pronoun:
  - Free choice (both strong and weak pronoun seem equally acceptable)
  - Preferred (the other pronoun would be dispreferred)
  - Obligatory (the other pronoun would be ungrammatical)
- Is the pronoun used for contrast (yes/no)?
- Pronoun type/POS: personal, possessive, generic, impersonal, verb (the form *zij* is also an archaic form of the verb to be)

The annotations were done by a single annotator (the present author), so no inter-annotator agreement score can be estimated. However, the binary distinction between grammatically obligatory or not is clear cut. The distinction between free choice and preferred is admittedly more subjective and judgments from multiple annotators would improve reliability; moreover, it might be better conceived as a spectrum of acceptability. However, the distinction is important for our research question, since the proposed explanations by Kaiser (2011) concern uses of pronouns where both forms are possible, but one is preferred. For the decision whether contrast was present, no context beyond the sentence was considered. Judging from the examples cited by Kaiser and Trueswell (2004), our notion of contrast is stricter; consider Kaiser and Trueswell (2004, 146):

(6) [context: Gilles and Ange are in the kitchen, and Ange notices Gilles looking outside intently. She tries to look [as well], knowing that outside are a garden, a river, the whole world.]
[…] maar hoe Ange zich ook inspant om van dat alles een glimp te ontwaren, *zij* ziet in de donkere ruit slechts de weerspiegeling van haar eigen keuken […] (Dorrestein, *Het hemelse gerecht*, p. 15)
'[…] but no matter how Ange exerts herself trying to catch a glimpse of all that, *she* sees in the dark pane nothing but her own kitchen […]'

If this pronoun is to be judged contrastive, this comes implicitly from the context and is a matter of interpretation; the pronoun has no emphasis or focus, and a weak pronoun would arguably fit equally well. Compare this example to the overt examples of contrast in the reference grammar *Algemene Nederlandse Spraakkunst* which require no context (Haeseryn et al. 1997, 252):

(7) Hij bedoelt *jou* niet, maar Mark.
'He does not mean *you*, but Mark.'

(8) Ik vind jouw verhaal veel geloofwaardiger dan dat van *hem*.
'I find your story much more credible than *his* one.'

Still, in (8) the contrast can also be on *jouw* (your) instead of *hem (*his). It seems difficult to operationalize the notion of contrast rigorously, and to avoid confirmation bias during annotation (for both presence and absence of contrast).

4.2 Results

Table 3 lists the distribution of pronoun types. For our research question, we focus on the personal pronouns, since the other types do not allow for both forms.[3] The personal pronouns also form the majority; we can therefore rule out that the other types are responsible for the outliers.

Table 3 Distribution of types

Personal	303
Possessive	26
Generic	20
Impersonal	6
Verb	1
Total	**356**

We will continue the analysis with only the 303 personal pronoun tokens. Figure 4 shows the breakdown of the other coded variables. We can conclude that the use of emphasis/contrast is rare. It can therefore be ruled out as an explanation for the outlying novels.

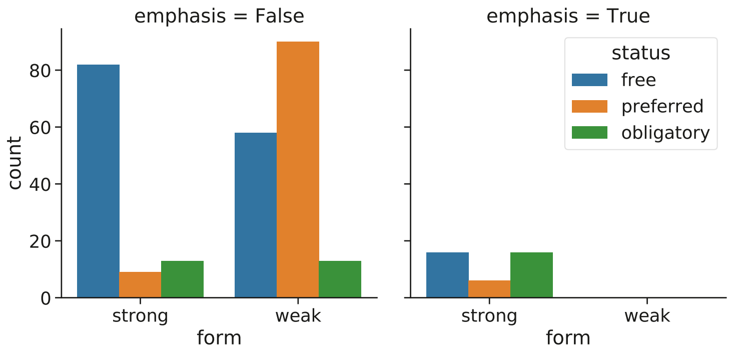


Fig. 4 Breakdown of manually analyzed pronouns (N = 303) (van Cranenburgh, CC BY).

When we focus on the pronouns without emphasis, which form the majority, we see that weak pronouns are often preferred, while strong pronouns are rarely preferred. Overall, a large proportion of both strong and weak pronouns are free choice, without preference for either form. This means that the grammatical explanations of salience and contrast cannot explain our observed outliers. Moreover, these results support the hypothesis that a large part of the strong/weak distinction is a stylistic matter. The following are typical examples of each category:

[3] The possessive *je* has the strong form *jouw*, but we choose to focus on personal pronouns.

(9) Weak pronouns:
a. Free: *We* speuren erfgenamen op
'*We* track down heirs'
b. Preferred: Dat weet *je* toch?
'*You* know that right?'
c. Obligatory: Mooie gouvernante is *me* dat.
'Nice governess that is.'

(10) Strong pronouns:
a. Free: Hoort u *mij?*
'Do you hear *me*?'
b. Preferred: Maar dan kennen ze *mij* niet.
'But then they haven't met *me*.'
c. Obligatory: Je ziet dat het niet van *mij* is!
'You can tell it's not *mine*!'

We also encountered some particularly interesting examples. The following are arguably unnatural usage of a strong reflexive pronoun, which may have been chosen for deliberate stylistic effect:

(11) a. Ik keek om *mij* heen
'I looked around *me*'
b. aangezien [...] heb ik altijd mijn eigen Duralexglas bij *mij*
'since [...] I always have my own Duralex glass with *me*'

In the following sentence, we have a clear example of a strong pronoun expressing contrast, as the strong pronoun picks out a different referent than the weak pronoun which occurs in the same sentence.

(12) Ik heb nooit kunnen vaststellen dat *ze* mij in de gaten hielden, al deden
ze dat natuurlijk wel, en *zij* in de eerste plaats.
'I have never been able to confirm that *they* were watching me, although
of course *they* did , and *she* most of all.'

However, it should be noted that this was the only clear example of contrast in the 303 pronouns we coded. This finding strongly contrasts with the preliminary corpus study of Kaiser and Trueswell (2004), who report that most uses of the strong pronoun *zij* are prompted by contrast.

5. Study 3: Coreference Analysis

A limitation of the approach in the previous section is the amount of subjectivity involved in the annotation. We now consider a more clearly defined task: coreference annotation. Given a pronoun, it is a well-defined question what other expressions in the text refer to the same entity. Coupled with the parse trees of sentences, we can directly test some of the proposed explanations for the distribution of strong and weak pronouns.

5.1 Materials and Methods

Fragments of a selection of 33 novels were manually annotated for coreference using the annotation scheme described in van Cranenburgh (2019). The set includes both the outliers as well as different kinds of novels without a high proportion of strong pronouns. The length of the fragments ranges from 1,000 to 20,000 tokens, rounded to the nearest sentence boundary. In total the subcorpus annotated for coreference contains 172,544 tokens. From the coreference annotations we extract the following predictors for each strong/weak pronoun:

1. pronoun function (core or non-core): non-core (i.e., not subject or object) arguments tend to be strong pronouns.
2. antecedent function (subject or other): the grammatical function of the antecedent (i.e., the closest preceding mention); to avoid sparsity, we only use an indicator for whether the antecedent is a subject or not. Subjects are more prominent and therefore more likely to be referred to by a strong pronoun.
3. distance: the distance to the antecedent in number of sentences. The distance is log transformed since it has a skewed distribution (most antecedents are close). A recently mentioned referent tends to be referred to by a weak pronoun.
4. chain density: the number of mentions in the same coreference chain as the pronoun in a window of 10 preceding sentences. A frequently used referent tends to be referred to by a weak pronoun.

These independent variables are compared to the dependent variable, whether a pronoun is strong or weak. We have also considered the number of competing mentions between the antecedent and the pronoun, but this was strongly correlated ($r = 0.83$) with distance and is therefore left out.

5.2 Results

A logistic regression with N = 3,549 strong/weak pronouns (Table 4) shows that except for distance, these variables are significant predictors.

Table 4 Logistic regression predicting strong pronouns from several proposed predictors

Dependent variable		**strong vs weak pronoun**	
No. observations		3,549	
Pseudo R-squ.		0.04412	
LLR p-value		1.650e-35	
	coef	**std err**	**p**
(intercept)	0.4882	0.157	0.002
pronfunc = core	−1.5708	0.128	0.000
antfunc = subj	−0.2851	0.082	0.001
log(distance)	0.0830	0.045	0.063
chain density	−0.0299	0.009	0.001

The continuous predictors are visualized in Figure 5. All signs of the coefficients match the hypothesized explanations: a negative coefficient indicates the variable makes a strong pronoun less likely, and vice versa. The logistic regression as a whole has a significant log-likelihood ratio as well, but the pseudo-R² is low. Since pseudo-R² is hard to interpret, we also calculate the area under the ROC-curve (a.k.a. concordance index); we compute this without cross-validation (within sample) and find C = 0.611. According to Hosmer and Lemeshow (2000, 162), C = 0.5 means no discrimination,

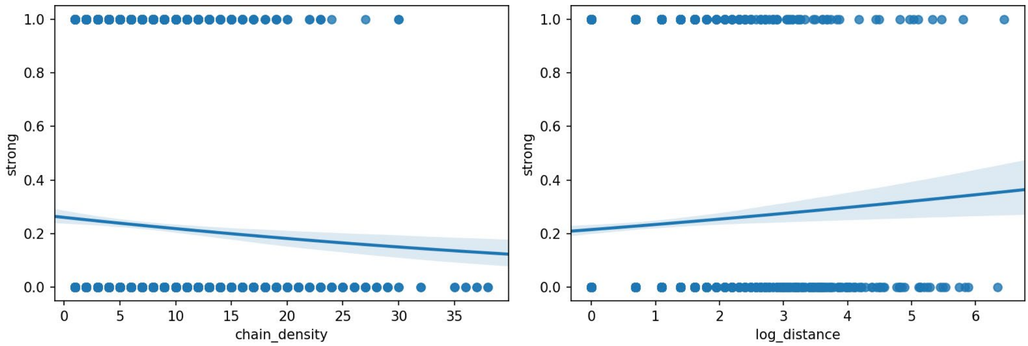


Fig. 5 Logistic regression plot of probability of a strong pronoun against chain density and distance to antecedent (van Cranenburgh, CC BY)

while $0.7 \leq C < 0.8$ means acceptable discrimination. We conclude that the model fit using these predictors is weak.

6. Discussion and Conclusion

We return to our research questions: we discovered a striking association between the use of strong/weak pronouns and literariness. There is a negative correlation with the number of pronouns and literariness, an expected result since both pronouns and less literary novels are associated with informal and spoken language. On the other hand, there is a positive correlation with the proportion of strong pronouns and literariness. A set of nine literary novels have a much larger than average proportion of strong pronouns. A manual analysis shows that these novels are not outliers due to grammatically obligatory strong pronouns, and a preference for strong pronouns is rare. Ergo, their authors freely chose to use a large number of strong pronouns, without being prompted by any of the proposed discourse-related factors. This was again confirmed by the statistical analysis of coreference annotations. The linguistic explanations for the use of strong and weak pronouns are shown to be significant variables but the amount of variance explained is limited; moreover, they cannot explain the outliers. What remains as a likely explanation is a stylistic dimension, given that choice is involved (whether the choice is deliberate is a second question). We submit that style is an important aspect of the use of strong and weak pronouns and referring expressions in general. Future research should investigate the stylistic effects of strong and weak pronouns in more detail by collecting fine-grained judgments from multiple annotators. Specifically, we should establish more precisely the degree of freedom in particular contexts, and the perceived unnaturalness of a variety of observed examples from both outliers and ordinary novels. The latter may turn out to be a clear instance of defamiliarization in literary language.

Code and data repository

https://github.com/andreasvc/strongweaklit

Acknowledgments

The author wishes to acknowledge helpful comments on this paper from Remco Knooihuizen, Jack Hoeksema, Kim Jautze, and an anonymous reviewer.

ORCID®

Andreas van Cranenburgh https://orcid.org/0000-0002-4545-1548

References

Arnold, Jennifer E., and Sandra A. Zerkle. 2019. "Why Do People Produce Pronouns? Pragmatic Selection vs. Rational Models." *Language, Cognition and Neuroscience*, 34 (9): 1152–75. https://doi.org/10.1080/23273798.2019.1636103.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726.

Bresnan, Joan. 1998. "Markedness and Morphosyntactic Variation in Pronominal Systems." In *Workshop Is Syntax Different*. http://web.stanford.edu/~bresnan/wow98-8.ps.

Donaldson, Bruce. 2008. *Dutch: A Comprehensive Grammar*. London: Routledge.

Haeseryn, Walter, Kirsten Romijn, Guido Geerts, Jacobus De Rooij, and Maarten Cornelis van den Toorn, eds. 1997. *Algemene Nederlandse Spraakkunst* [General Dutch Grammar]. Groningen: Martinus Nijhoff. https://e-ans.ivdnt.org/.

Hosmer, David W., and Stanley Lemeshow. 2000. *Applied Logistic Regression*. New York: Wiley.

Kaiser, Elsi. 2011. "Salience and Contrast Effects in Reference Resolution: The Interpretation of Dutch Pronouns and Demonstratives." *Language and Cognitive Processes* 26 (10): 1587–1624. https://doi.org/10.1080/01690965.2010.522915.

Kaiser, Elsi, and John Trueswell. 2004. "The Referential Properties of Dutch Pronouns and Demonstratives: Is Salience Enough." *Proceedings of Sinn und Bedeutung* 8 (August), 137–50: Konstanz: University of Konstanz. https://doi.org/10.18148/sub/2004.v8i0.754.

Koolen, Corina, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout. 2020. "Literary Quality in the Eye of the Dutch Reader: The National Reader Survey." *Poetics* 79:101439. https://doi.org/10.1016/j.poetic.2020.101439.

van Cranenburgh, Andreas. 2019. "A Dutch Coreference Resolution System with an Evaluation on Literary Fiction." *Computational Linguistics in the Netherlands Journal* 9: 27–54. https://clinjournal.org/clinj/article/view/91.

van Cranenburgh, Andreas, and Rens Bod. 2017. "A Data-Oriented Model of Literary Language." In *Proceedings of EACL*, 1228–38, Association for Computational Linguistics. http://aclweb.org/anthology/E17-1115.

van Cranenburgh, Andreas, Karina van Dalen-Oskam, and Joris van Zundert. 2019. "Vector Space Explorations of Literary Language." *Language Resources and Evaluation* 53: 625–50. https://doi.org/10.1007/s10579-018-09442-4.

Investigating the Relation between Syntactic Complexity and Subgenre Distinction

A Case Study on Two Contemporary French Authors

Robert Hesselbach

Abstract The purpose of this article is to explore the ways in which the analysis of the syntactic complexity of a text's sentences can help distinguish between works belonging to different literary subgenres written by the same author. Based on the considerations of an earlier study (Hesselbach 2019), syntactic complexity is understood here as an array of qualitative as well as quantitative features. Applying this method to a corpus of two contemporary French authors and their novels (1979–2002), comprising both crime fiction (*roman policier*) and 'high literature' (*littérature blanche*), the results of this study show that syntactic complexity has very little influence on genre distinction, at least for the two subgenres examined. In fact, a very stable distribution of results can be observed in both qualitative and quantitative terms. Furthermore, the evidence suggests that the degree of syntactic complexity is more likely to appear as an author-related characteristic.

Keywords syntactic complexity, (sub)genre distinction, French novel, Yasmina Khadra, Jean Echenoz

1. Introduction

The present study was conducted within the framework of the research group *CLiGS (Computational Literary Genre Stylistics)* at the University of Würzburg/Germany, one of whose objectives was to develop computer-based methods to identify

Hesselbach, Robert: Investigating the Relation between Syntactic Complexity and Subgenre Distinction.
A Case Study on Two Contemporary French Authors, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics
in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 235–259.
DOI: https://doi.org/10.17885/heiup.1157.c19374

genre-specific characteristics working with Romance, and more precisely French and Spanish, literature. Genre distinction from a digital point of view has been the subject of a number of recent research publications, as is evident from papers on stylometric analysis of ancient Greek literary texts (Gianitsos et al. 2019) or German novels using word frequencies and character tetragrams (Hettinger et al. 2016), human versus machine genre classification of Spanish novels (Calvo Tello 2021), sentiment analysis for Spanish American novels (Henny-Krahmer 2018), or topic modeling on French Classical and Enlightenment drama (Schöch 2017). This paper is concerned with a subject which can be situated at the intersection of literary studies and linguistics: the relation between syntactic complexity and (sub)genre distinction. In this context, it can be noted that very different research questions are associated with the concept of syntactic complexity on the one hand and the task of genre distinction on the other hand. Nevertheless, an attempt will be made at this point to identify certain aspects of syntactic complexity, as already defined in an earlier study (cf. Hesselbach 2019), on the basis of a corpus of two contemporary French authors and their novels, comprising both crime novels (*roman policier*) and 'high literature' (*littérature blanche*). Subsequently, it may be possible to make statements as to whether certain subgenres of French authors exhibit common features in terms of syntax.

The next section will first discuss which approaches already exist to describe syntactic complexity and which (qualitative as well as quantitative) perspective is used in this study. Section 3 then presents the actual empirical study of two contemporary French authors (Yasmina Khadra and Jean Echenoz), focusing first on the description of the corpus-based method before presenting the results and situating them in the research context. In the last section, after a brief synopsis, further possible research perspectives are presented.

2. Approaches to *Syntactic Complexity*

If one tries to approach the concept of *syntactic complexity*, one will find that there are various ways to approach the issue from a conceptual perspective. The first problem concerns the level of the linguistic system: does the expression refer to the complexity of phrases or entire sentences? In this article we refer only to *syntactic complexity* in the sense of sentence complexity, although looking at the complexity of phrases from a stylistic point of view also opens up interesting research perspectives. Another fundamental question is whether one refers to quantitative or qualitative aspects of complexity. The shortest complex sentence in Spanish as a pro-drop language can consist of only two words, as in Sp. *oigo cantar* 'I hear (somebody) singing' whereas simple sentences can turn out to be fairly long, as the French example in (1) illustrates:[1]

(1) En chemin, dans le crépuscule, elle *nomma* le palais de justice,
la sous-préfecture, la mairie, la prison, la maison natale de
Frédérick Lemaître. (E_Eq_po, 17)[2]
'On the way, in the twilight, she named the courthouse, the sub-prefecture, the town hall, the prison, the birthplace of Frédérick Lemaître.'

The above sentence shows no kind of coordination nor subordination and consists of 23 words. This should make clear that one can make different statements regarding the *type* (qualitative aspects) and *extent* (quantitative aspects) of a construction when speaking of *syntactic complexity*. For this reason, both qualitative and quantitative aspects will be taken into consideration in this article. After a brief overview of different qualitative and quantitative approaches to (measure) *syntactic complexity*, section 3 presents the method and results of the study presented here.

2.1 Qualitative Aspects

Taking a closer look at (not only) French grammars and manuals, complexity is usually understood as a syntactic hierarchical relationship between two sentences/clauses, so that consequently a distinction is made between the *complex* and the *simple* sentence. While examples (2) and (3) are each syntactically simple sentences, (4) and (5) are correspondingly complex sentences (examples taken from Kiesler 2013, 613; originally in: Dubois et al. 1994, s.v. *parataxe*):

(2) Cet homme *est* habile.
'This man is clever.'

(3) Il *réussira*.
'He will succeed.'

[1] Since finite verbs can be regarded as a quantitatively relevant feature of *syntactic complexity*, they are italicized in all given examples, unless another linguistic unit is to be emphasized.

[2] The information at this point refers to texts of the corpus examined. The pattern "author_title_subgenre" is used, so that this refers to the 17th sentence randomly taken from a novel by Jean Echenoz with the title *L'Équipée malaise*, which can be assigned to the subgenre crime novel (*policier* = po).

(4) Cet homme *est* habile et il *réussira*.
'This man is clever and he will succeed.'

(5) Cet homme *réussira*
parce qu'il *est* habile.
'This man will succeed because he is clever.'

Example (4) represents a paratactic structure due to the coordination of the two main clauses, whereas (5) represents a hypotactic construction of main and subordinate clauses, where the hierarchical distinction is made clear by indenting the dependent subordinate clause.[3] Nevertheless, these representations in grammars and manuals are rather prototypical. In actual language use, and thus also in the literary production of novels, hybrid forms of these construction possibilities occur, as Kiesler (2013) points out based on a French corpus. The author presents a typology of complex sentences which distinguishes between (multiple) homogeneous and heterogeneous structures and serves as a basis for this study, as will be explained below (a–g). An example is given from Kiesler (2013), as well as from the corpus of novels analyzed here:

(a) homogeneous parataxis (ho-pa)[4]

A *homogeneous parataxis* (Fr. *phrase parataxique homogène*) is understood to be those cases in which only two main clauses are combined and neither of the two has its own dependent subordinate clauses, as illustrated by examples (6) and (7). It becomes clear that coordination can be carried out both syndetically (i.e. through a conjunction), as shown in example (6), and asyndetically (i.e. without any conjunction), as in (7):

(6)	Je me *plaisais*	et	je *cherchais* à plaire.	(Kiesler 2013, 617)
	'I liked myself	and	I liked to please.'	

[3] Within his theory of *Junktion* ('linkage') Raible (1992) refers to the fact that, at this stage of clause-linkage, there is basically freedom of position of the different clauses, so that no change of meaning is caused in the following correspondences of the given examples: (4') *Cet homme réussira et il est habile* and (5') *Parce que cet homme est habile, il réussira*. Restrictions on the freedom of position exist in cases where an element of the second sentence refers to the first sentence, cf. (4'') *Cet homme est habile et* c'est pourquoi *il réussira* ('This man is clever and *that is why* he succeeds') versus **Cet homme réussira et c'est pourquoi il est habile* ('This man will succeed and that is why he is clever').

[4] Kiesler himself speaks of "phrases parataxiques simples" – "simple paratactic sentences" (2013, 616), but for the sake of a coherent terminology, I prefer to denominate this construction as *homogeneous parataxis*.

(7) Ses doigts *tripotent* dangereusement l'instrument de mort,
le *ramassent*. (K_Do_po, 100)
'His fingers are dangerously touching the instrument of death, picking it up.'

(b) heterogeneous parataxis (he-pa)

In contrast to the previous sentence type, the *heterogeneous parataxis* (Fr. *phrase paratax-ique hétérogène*) consists of a combination of two main clauses, at least one of which contains a dependent subordinate clause.

(8) Obélix ne *veut* pas finir son sanglier,
il *dit*
qu'il n'*a* plus faim! (Kiesler 2013, 614)
'Obelix doesn't want to finish his boar, he says he's not hungry anymore.'

(9) Il *vérifia* la monnaie déposée d'avance sur la table
et
plia son journal
sans quitter des yeux le secrétaire. (E_Me_po, 85)
'He checked the change left on the table and folded his newspaper without taking his eyes off the secretary.'

In example (8) the second main clause (*il dit*) contains a dependent subordinate clause with a finite verb (*qu'il n'a plus faim*). The following example (9) demonstrates that these dependent subordinate clauses can also be realized through infinite verbal structures (*sans quitter des yeux le secrétaire*).

(c) multiple homogeneous parataxis (mu-ho-pa)

A *multiple homogeneous parataxis* (Fr. *phrase parataxique multiple homogène*) presents a coordinated sequence of at least three main clauses, none of which has a dependent subordinate clause, as can be observed in (10) and (11):

(10) Je *regardais*,
je *palpais*,
j'*apprenais* le monde, à l'abri. (Kiesler 2013, 614)
'I watched, I palpated, I learned about the world, sheltered.'

(11) Omar *vibre* des épaules,
passe une grosse langue sur ses lèvres
et
fait tinter ses bagues sur son comptoir. (K_Mo_po, 199)
'Omar shivers his shoulders, runs a big tongue across his lips and jingles his rings on his counter.'

In (10) a sequence of three main clauses (*je regardais*; *je palpais*; *j'apprenais le monde*) can be observed, all of which are linked without any conjunction. Example (11) basically works in a similar way and consists of three main clauses, but in this case the last one (*fait tinter ses bagues sur son comptoir*) is connected by the conjunction *et*.

(d) multiple heterogeneous parataxis (mu-he-pa)
Analogous to the properties of the *heterogeneous parataxis*, a *multiple heterogeneous parataxis* (Fr. *phrase parataxique multiple hétérogène*) is a string of three or more independent main clauses in which at least one dependent subordinate clause must occur, as (12)[5] and (13) illustrate:

(12) [alors il y *avait* des défilés]
[il y *avait* des tas de trucs]
et
[tout ça se *passait*
 d'où je *travaillais*
  parce que c'*était* juste en face] (Kiesler 2013, 614)
'so there were parades and there was a lot of stuff and it all happened from where I worked because it was right across the street'

(13) Les prières s'*émiettent* dans la furie des mitrailles,
les loups *hurlent* chaque soir à la mort,
et
le vent,
 lorsqu'il se *lève*,
livre la complainte des mendiants au croassement des corbeaux.
(K_Hi_bl, 50)
'Prayers crumble in the fury of the machine-gun fire, the wolves howl, and the wind, when it rises, delivers the lament of the beggars to the crows' caws.'

[5] In this example, the bracketing has been taken from Kiesler's original example.

In the case of (12) a string of three main clauses can be seen, the last of which contains two dependent subordinate clauses (*d'où je travaillais*; *parce que c'était juste en face*). Example (13), which is taken from Yasmina Khadra's novel *Les Hirondelles de Kaboul*, also contains three main clauses, the third of which (*le vent livre la complainte des mendiants au croassement des corbeaux*) again contains a dependent subordinate clause, more precisely an adverbial clause (*lorsqu'il se lève*).

(e) simple hypotaxis (sim-hy)

The first type of hypotactic constructions is the so-called *simple hypotaxis* (Fr. *phrase hypotaxique simple*). It consists only of a main clause and a subordinate clause dependent on that main clause, as can be seen in (14) and (15):

(14) Quand le chat n'*est* pas là,
les souris *dansent*. (Kiesler 2013, 619)
'When the cat's away, the mice dance.'

(15) Mais trois heures plus tard,
lorsque Georges *rentra* rue Oberkampf,
Véronique n'*était* toujours pas là. (E_Ch_po, 192)
'But three hours later, when Georges returned to rue Oberkampf,
Véronique was still not there.'

In example (14), the whole construction begins with the subordinate clause (*quand le chat n'est pas là*) which is followed by the main clause (*les souris dansent*), whereas in (15) the adverbial clause (*lorsque Georges rentra rue Oberkampf*) is inserted in the main clause (*mais trois heures plus tard Véronique n'était toujours pas là*).

(f) multiple homogeneous hypotaxis (mu-ho-hy)

Compared to the type of the *simple hypotaxis*, the *multiple homogeneous hypotaxis* (Fr. *phrase hypotaxique multiple homogène*) is also characterized by a main clause with a hypotactic structure, but in this case it has several dependent clauses which all belong to the same type of subordinate clause, as the following two examples illustrate:

(16) Geoffroy *a* un papa très riche
[qui lui *achète* tous les jouets
[qu'il *veut*.]] (Kiesler 2013, 619)
'Geoffroy has a very rich daddy who buys him all the toys he wants.'

(17) Il *subodorait* l'imminence d'une révolution
qui ne *pardonnerait* rien à ceux
qui ne *prendraient* pas le train en marche. (K_Re_bl, 140)
'He sensed the imminence of a revolution that would not forgive those who did not get on board.'

Besides the main clause (*Geoffroy a un papa très riche*), example (16) reveals two dependent subordinate clauses (*qui lui achète tous les jouets*; *qu'il veut*), which can both be characterized as relative clauses. The following sentence taken from Yasmina Khadra's *À quoi rêvent les loups* also consists of one main (*Il subodorait l'imminence d'une révolution*) and two subordinated clauses (*qui ne pardonnerait rien à ceux*; *qui ne prendraient pas le train en marche*). Even if the second subordinate clause does not depend on the first subordinate clause, it still belongs to the same type of subordinate clause, namely a relative clause.

(g) multiple heterogeneous hypotaxis (mu-he-hy)
In case the dependent subordinate clauses of a hypotactic construction belong to different classes, Kiesler speaks of a *multiple heterogeneous hypotaxis* (Fr. *phrase hypotaxique multiple hétérogène*). The two sentences in (18) and (19) can be described as such:

(18) Agnan,
[qui *est* le premier de la classe [...],]
a dit
[que ce *serait* dommage de ne pas avoir arithmétique,
[parce [qu'il *aimait* ça]
et
[qu'il *avait* bien fait tous ses problèmes.]]] (Kiesler 2013, 620)
'Agnan, who is at the top of the class, said that it would be a shame not to have arithmetic, because he liked it, and that he had done all his problems well.'

(19) Lorsque Georges *fut* entré dans ce passage,
quatre personnes au moins s'y *engagèrent* également,
qui toutes s'*intéressaient* à lui. (E_Ch_po, 84)
'When George had entered this passage, at least four people entered as well, all of whom were interested in him.'

In the example taken from Kiesler (2013) a total of four dependent subordinate clauses can be identified, which depend on a main clause. Whereas the first one (*est le premier de la classe*) presents a relative clause modifying the noun *Agnan*, there are two parallelly coordinated *that*-clauses (*que ce serait dommage de ne pas avoir arithmétique*; *qu'il avait bien fait tous ses problèmes*) and one adverbial clause (*parce qu'il aimait ça*). The first subordinate clause (*lorsque Georges fut entré dans ce passage*) in example (19) again shows an adverbial clause (with temporal meaning), whereas the second subordinate clause (*qui toutes s'intéressaient à lui*) can again be described as a relative clause, which defines the noun *personnes* more precisely.

After this typology of complex sentences has been presented, a note must be made about the distinction of simple syntactic constructions. While in the area of complex syntactic structures, Kiesler's typology of complex sentences is used for the analysis, in the field of simple sentences, only those with a finite verbal element and constructions without any finite verb are distinguished. Examples (20) and (21) can be characterized as a *simple sentence* (*simple*), whereas (22) and (23) both represent a *simple sentence without finite verb* (*simple wv*).[6]

(20) Je ne *peux* pas accepter. (E_Ch_po, 119)
'I can't accept.'

(21) Les trois gamins se *sont* déportés sur une autre voiture. (K_Mo_po, 81)
'The three kids fled to another car.'

(22) Rien à faire. (K_Re_bl, 193)
'Nothing to do.'

(23) Et le type? (E_Eq_bl, 40)
'And the guy?'

Thus, these nine types of sentences are used for the qualitative analysis of the sentences of the research corpus (cf. sections 3.1 and 3.2).[7] In the following section, on the other

[6] On the question of whether such constructions can be called *sentences*, Trabant, among others, comments with reference to Bloomfield: "Der Satz (sentence) [sic] wird nämlich deswegen von Bloomfield hervorgehoben, weil er die sprachliche Form ist, die als abgeschlossene Äußerung (*utterance*)—also als Text—auftreten kann. So ist z.B. nicht nur die Äußerung *Poor John ran away* ein Satz, sondern auch *Poor John!* oder *John!*, eben weil sie die abgeschlossene Äußerungen—Texte—ausmachen können" (1981, 8), Eng. "The sentence is emphasized by Bloomfield because it is the linguistic form which can appear as a closed utterance—that is, as text. Thus, for example, not only the utterance *Poor John ran away* is a sentence, but also *Poor John!* or *John!* because they can constitute the completed utterance—text".

[7] These include two different simple sentences, four different types of parataxis, and three types of hypotaxis, so that one can assume the following order from simple to complex: *simple sentence without finite verb* (*simple wv*)—*simple sentence* (*simple*)—*homogeneous parataxis* (*ho-pa*)—*heterogeneous parataxis* (*he-pa*)—*multiple homogeneous parataxis* (*mu-ho-pa*)—*multiple heterogeneous*

hand, the quantitative perspective of studies already carried out will be discussed and an own vector model will be presented.

2.2 Quantitative Aspects

In the past, a large number of quantitative approaches to syntax have been taken, so only a short excerpt will suffice at this point. Using French, Koch (1995) has described a method by which he determines the number of dependent subordinate clauses per main clause and describes the result as *complexité quantitative* ("quantitative complexity"). If one takes another look at examples (18) and (19) and applies Koch's approach, the ratio of the *complexité quantitative* for (18) can be described as 1:4, while for (19) it can be represented as 1:2.

(18) Agnan,
[qui *est* le premier de la classe [...],]
a dit
[que ce *serait* dommage de ne pas avoir arithmétique,
[parce [qu'il *aimait* ça]
et
[qu'il *avait* bien fait tous ses problèmes.]]] (Kiesler 2013, 620)
'Agnan, who is at the top of the class, said that it would be a shame not to have arithmetic, because he liked it, and that he had done all his problems well.'

(19) Lorsque Georges *fut* entré dans ce passage,
quatre personnes au moins s'y *engagèrent* également,
qui toutes s'*intéressaient* à lui. (E_Ch_po, 84)
'When George had entered this passage, at least four people entered as well, all of whom were interested in him.'

This method works very well for complex sentences which only have one main clause. However, for complex constructions containing several main clauses (see above), this method can only be used for the individual main clauses (and the corresponding dependent subordinate clauses).

However, a simple and intuitive way to describe syntactic complexity—not only for linguists—is certainly the measurement of the sentence length which several

parataxis (*mu-he-pa*)—*simple hypotaxis* (*sim-hy*)—*multiple homogeneous hypotaxis* (*mu-ho-hy*)—*multiple hypotaxis* (*mu-he-hy*).

scholars refer to: length can be measured in words (i.e. Sowinski 1999; Szmrecsányi 2004, 1032–33), through the number of immediate constituents (Altmann, Best and Popescu 2014, 94; Altmann and Köhler 2000, 192), through the number of syllables (Best 2005, 300;[8] Fucks 1968, 87) or—for example for sign languages—through the number of characters (Jing 2001). Defining sentence length by counting the words and applying this approach to the examples (18) and (19), the sentence in (18) with 33 words would be more complex than (19) comprising 21 words. As mentioned before, length cannot be the only parameter for measuring complexity, as there can be long sentences without any subordination but also relatively short sentences containing several subordinate clauses. Hence, the degree of embedding must be taken into consideration as well, and therefore represents the main characteristic of syntactic complexity for several scholars (i.e. Givón 2009; Givón and Shibatani 2009). Another way of measuring syntactic complexity is counting the number of nodes dominated (i.e. Johnson 1966; Ferreira 1991), on which Szmrecsányi comments: "Presupposing some notion of formal complexity, counting the number of nodes dominated is conceptually certainly the most direct and intuitively the most appropriate way to assess syntactic complexity. This is because the method reflects how the human parser is supposed to work" (2004, 1033). Szmrecsányi finally proposes a so-called *Index of Syntactic Complexity* (ISC), which is defined as follows:

$$\mathrm{ISC(u)} = 2 \times \mathrm{n\ (u, SUB)} + 2 \times \mathrm{n\ (u, WH)} + \mathrm{n\ (u, VF)} + \mathrm{n\ (u, NP)}$$[9]

Even if this approach of taking several factors into account offers an interesting perspective, it remains unclear, among other things, on which scale the determined values are to be located. The author therefore also gives cause for consideration when introducing this formula: "I would, then, like to suggest the following formula –which, admittedly, is somewhat tentative and ad-hoc [!]– to establish (ISC)" (Szmrecsányi 2004, 1034).

[8] Cf. Best (2005, 300): "Es spricht nun nichts dagegen, Satzlänge auch ganz anders zu messen: nach der Zahl der Silben pro Satz […], oder was anscheinend noch niemand versucht hat, nach der Zahl der Morphe. Auch noch kleinere Einheiten (Laut, Buchstabe) können für Satzlänge genutzt werden […]", Eng. "There is nothing to be said against measuring sentence length in a completely different way: according to the number of syllables per sentence […], or, which apparently no one has ever tried, according to the number of morphs. Even smaller units (phone, letter) can be used for sentence length".

[9] "Let u be the unit of linguistic data under analysis, let ISC(u) be the ISC of the unit of linguistic data under analysis, and let n(u,SUB) be the number of occurrences of SUB in the unit of linguistic data under analysis, etc. According to this formula, ISC of a given unit of data is twice the number of occurrences of subordinating conjunctions and WH-pronouns plus the number of occurrences of verb forms and noun phrases in that unit" (Szmrecsányi 2004, 1035).

The method developed in my dissertation (Hesselbach 2019) takes into account the most important quantitative characteristics of syntactic complexity, namely the sentence length (SL) measured in words, the number of finite verbs (FV) and the maximum depth of embedding (DE) of a syntactic construction. Instead of connecting the determined values in a quotient or a product,[10] they are represented as the following vector: $x = \begin{pmatrix} SL \\ FV \\ DE \end{pmatrix}$. Figure 1 shows this vector presentation schematically.

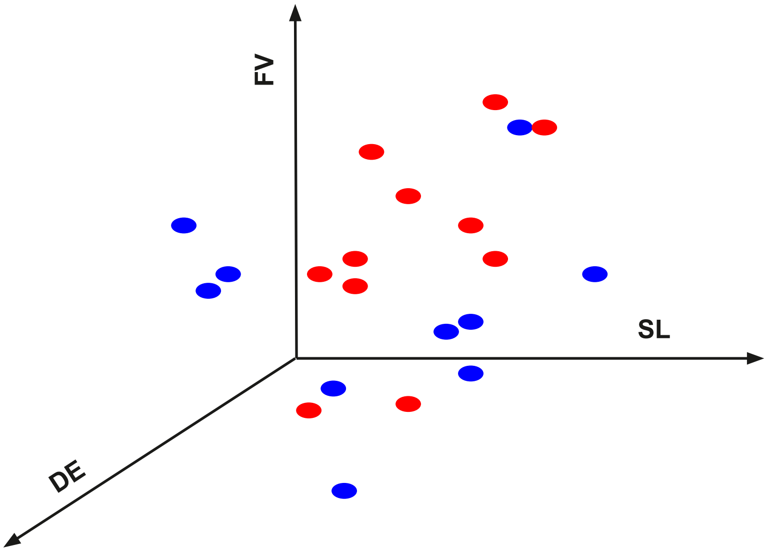


Fig. 1 Schematic illustration of the vector representation (Hesselbach, CC BY).

The advantage of this method—in agreement with Altmann (1978)—is that no mathematical operations are performed to describe syntactic complexity, but that the numerically measurable values of a construction, as just described, are only plotted as vectors.

[10] In this context Altmann speaks of a *sin* which is common in linguistics: “Die üblichste ‘Sünde’ in der Linguistik ist die Bildung von irgendwelchen Quotienten ohne Rücksicht auf die linguistischen und die mathematischen Aspekte des gegebenen Indexes” (1978, 91), Eng. “The most common ‘sin’ in linguistics is the formation of some quotients without regard to the linguistic and mathematical aspects of the given index”.

The visualization then makes several things clear: On the one hand, the different subcorpora can be marked in color, so that a quick overview of the actual ratios can be gained visually. On the other hand, this method can also be used to compare the degree of complexity: the further away from the zero position a data point is located, the more complex the syntactic construction. In addition to determining the qualitative characteristics, this method will now be applied to our corpus of contemporary French novels.

3. A Case Study on Two Contemporary French Authors

Previous studies have focused on the application of this method to a stylistically heterogeneous corpus of modern European Spanish and French (Hesselbach 2019) and to a diachronic corpus of Spanish literature (Hesselbach, in prep.). In contrast, the study presented here focuses on two contemporary French authors and their novels and aims to examine whether determining syntactic complexity, as described above, can help to define subgenre distinction within an author's oeuvre. In the next step, the method and composition of the corpus is explained, before the results of the analysis are presented and interpreted in section 3.2.

3.1 Method

As mentioned before, the aim of the present study is to make a statement as to whether and to what extent the analysis of syntactic complexity can help to define (sub)genre distinctions within an author's oeuvre. Therefore, a corpus-based approach was chosen to investigate different texts. Hence, a corpus of contemporary French-language novels was compiled and digitized. It comprises eight novels written by two different authors (Jean Echenoz and Yasmina Khadra)[11] and covers the period from 1979 to 2002. It is important for the compilation of the corpus that four novels each can be ascribed to the subgenres (*roman*) *policier* and (*littérature*) *blanche*, as both authors have written novels in both subgenres. Table 1 shows the compilation of the corpus indicating the date of publication, the subgenre and the abbreviation (chosen for this study). Since the research question presented at the beginning aims at subgenre distinctions, the texts in this table are clustered due to their belonging to different subgenres (*roman policier* versus *littérature blanche*) and neither chronologically nor alphabetically.

[11] Even though Yasmina Khadra, an Algerian-born writer living in France, is part of this analysis, the regional variety affiliation is not considered significant for an investigation of syntactic complexity—in contrast to questions concerning the lexicon or idiomatic expressions, for example.

Table 1 Corpus of contemporary French novels

	Author	Title of novel	Year of publication	Subgenre	Abbreviation
1.	Echenoz, Jean	*Cherokee*	1983	*policier*	E_Ch_po
2.	Echenoz, Jean	*Le Meridien de Greenwich*	1979	*policier*	E_Me_po
3.	Khadra, Yasmina	*Double blanc*	1998	*policier*	K_Do_po
4.	Khadra, Yasmina	*Morituri*	1997	*policier*	K_Mo_po
5.	Echenoz, Jean	*L'Équipée malaise*	1986	*blanche*	E_Eq_bl
6.	Echenoz, Jean	*Nous trois*	1992	*blanche*	E_No_bl
7.	Khadra, Yasmina	*Les Hirondelles de Kaboul*	2002	*blanche*	K_Hi_bl
8.	Khadra, Yasmina	*À quoi rêvent les loups*	1999	*blanche*	K_Re_bl

From each of the texts, which were already available in digital form as plain-text documents, 200 sentences were randomly selected using *Python*,[12] so that the total corpus examined here has a size of (8×200 =) 1,600 sentences.[13] For each of these sentences, the sentence type, the sentence length, the number of finite verbs and the degree of maximum embedding depth—as described before—were determined.[14] Two example analyses are used to illustrate the procedure:

(12) [alors il y *avait* des défilés] [il y *avait* des tas de trucs]
et [tout ça se *passait*
d'où je *travaillais*
parce que c'*était* juste en face]

Sentence type	mu-he-pa	Sentence length (in words)	29
Number of finite verbs	5	Depth of embedding	2

(16) Geoffroy *a* un papa très riche
[qui lui *achète* tous les jouets
[qu'il *veut*.]]

Sentence type	mu-ho-hy	Sentence length (in words)	15
Number of finite verbs	3	Depth of embedding	2

[12] I would especially like to thank Daniel Schlör for his very valuable support in the application of *Python*.

[13] The research data can be accessed via the following link: https://doi.org/10.5281/zenodo.7458279.

[14] The sentence type as well as the degree of maximum embedding depth were determined manually, while the sentence length and number of finite verbs were analyzed automatically.

After having looked at the two example analyses, the results obtained by analyzing the entire corpus are now presented below.

3.2 Results

The aim of the study presented here is to use corpus data on two different subgenres of contemporary French novels to determine whether the description of syntactic complexity can help to define genre distinctions. In a first step, we will now take a closer look at the relationship between the different types of sentences, before the quantitative aspects of syntactic complexity are considered in section 3.2.2.

3.2.1 Qualitative Aspects

As explained above, both simple and complex sentences can be differentiated even more precisely. Table 2 shows the frequencies of the individual types of sentences for the two subgenres in question. For this purpose, the individual novels which can be assigned to the respective subgenre were combined and the individual frequencies of occurrence were summed up.

Table 2 Distribution of different types of sentences in the corpus

		Jean Echenoz		**Yasmina Khadra**	
		Subgenre		**Subgenre**	
Sentence type		*policier*	*blanche*	*policier*	*blanche*
Simple	simple wv	17	34	30	21
	simple	111	94	193	179
	Σ	128	128	223	200
Complex	ho-pa	61	52	44	50
	he-pa	46	54	25	25
	mu-ho-pa	27	27	13	7
	mu-he-pa	26	24	5	10
	Σ	160	157	87	92
	sim-hy	54	66	76	77
	mu-ho-hy	24	24	5	16
	mu-he-hy	34	25	9	15
	Σ	112	115	90	108
	Σ	400	400	400	400
	Σ	800		800	

If one reads the table from top to bottom, one can recognize the various frequencies and find that in both subgenres the *simple sentence* (with a finite verb) is by far the most common sentence type with 111 (*policier*) and 94 occurrences (*blanche*) for Echenoz and 193 (*policier*) and 179 (*blanche*) for Khadra. It is also noteworthy that the second most common type of sentence (in both subgenres) is the *simple hypotaxis* (with the exception of Echenoz' *policier* novels where the *homogenous parataxis* is predominant). This means that even in the field of complex sentences, the *simplest* types of sentences are very popular with both novelists.[15] If the individual sentence types are ordered according to their frequency, one can get a quick overview of the results in Table 3.

Table 3 Distribution of different types of sentences in the corpus

	Jean Echenoz		Yasmina Khadra	
	policier	*blanche*	*policier*	*blanche*
1.	simple (28 %)	simple (24%)	simple (48%)	simple (45%)
2.	ho-pa (15%)	sim-hy (17%)	sim-hy (19%)	sim-hy (19%)
3.	sim-hy (14%)	he-pa (14%)	ho-pa (11%)	ho-pa (13%)
4.	he-pa (12%)	ho-pa (13%)	simple wv (8%)	he-pa (6%)
5.	mu-he-hy (9%)	simple wv (9%)	he-pa (6%)	simple wv (5%)
6.	mu-ho-pa (7%) mu-he-pa (7%)	mu-ho-pa (7%)	mu-ho-pa (3%) mu-he-hy (3%)	mu-ho-hy (4%) mu-he-hy (4%)
7.	mu-ho-hy (6%)	mu-he-hy (6%) mu-ho-hy (6%) mu-he-pa (6%)	mu-he-pa (1%) mu-ho-hy (1%)	mu-he-pa (3%)
8.	simple wv (4%)	–	–	mu-ho-pa (2%)
9.	–	–	–	–

[15] Another result, which is initially not the focus of interest in the study presented here, is obtained by reading the table from left to right. This gives a general impression of the types of sentences across the genres, and it is shown that the simple sentence types, with a total of (128 +128 + 223 + 200 =) 679 (= 42%) cases, make up almost half of all sentences examined. The sum of the different paratactic constructions (160 + 157 + 87 + 92 = 496 = 31%) is similar to the values of the hypotactic structures (112 + 115 + 90 + 108 = 425 = 27%).

The distribution according to frequency clearly shows that the individual types of sentences are not only in a similar order, but also have comparable percentages for the most part. A clear difference can be seen not between the subgenres, but rather between the two authors: even though the simple sentence is the most common in both subgenres for the two authors, Khadra uses it almost twice as often (48 percent and 45 percent) as Echenoz (28 percent and 24 percent).[16] However, if we look at the characteristics for the particular subgenres for each author separately, we can see that no meaningful qualitative differences can be detected. It now remains to be examined whether this also applies to the corresponding quantitative characteristics.

3.2.2 Quantitative Aspects

In the description of the individual quantitative characteristics, they are first listed individually before they are finally combined in a vector diagram. First, (a) the extent of the sentence length is examined more closely, followed by an analysis of (b) the number of finite verbs and (c) the maximum degree of syntactic embedding.

(a) Sentence length

As described above, in this study the sentence length is determined by the number of words. The following box plot, Figure 2, gives an overview of the different values of sentence length in the samples of the individual texts. Note that the first four texts (on the left side of the box plot) belong to Echenoz's texts (with the subgenres *policier* and *blanche*), while the right half represents the results for both subgenres by Khadra.

The entire corpus comprises a total of 25,081 words, of which 15,273 belong to Echenoz (*policier*: 7,593; *blanche*: 7,680) whereas the other 9,808 words can be related to Khadra's novels (*policier*: 4,595; *blanche*: 5,213). The analysis clearly shows that the two authors do differ in terms of sentence length. Even though the value for the median of the individual text data is less than twenty words in all cases, a comparatively greater tendency toward longer sentences can be observed in Echenoz's texts regardless of the subgenre. The data in the box plot also provides the result that there are more significant differences between the two authors than between the subgenres: when looking at the results for the texts of Echenoz (*E_Ch_po*, *E_Me_po*, *E_Eq_bl*, *E_No_bl*) statistical outliers start at a limit of about 50 words. In contrast, this is already the case for Khadra with about 30 words. The longest sentence (132 words) can also be found in one of Echenoz's texts (*E_No_bl*). The following table shows again clearly

[16] The dominance of the *simple sentence* is highly significant for both authors in both subgenres when looking at the corresponding p-values (based on a χ^2-test) for Echenoz (*policier*: $p < 2.2e\text{-}16$; *blanche*: $p = 1.158e\text{-}12$) and Khadra (*policier*: $p < 2.2e\text{-}16$; Khadra, *blanche*: $p < 2.2e\text{-}16$).

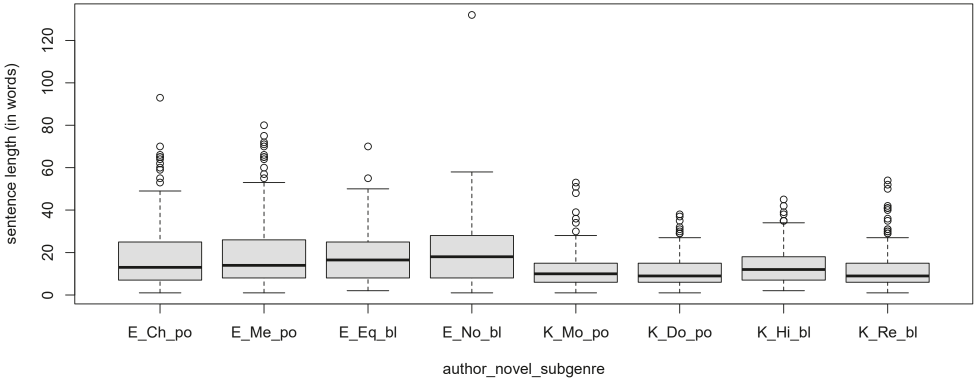


Fig. 2 Distribution of sentence length of the examined texts (Hesselbach, CC BY).

Table 4 Average sentence length (in words) by author

	Jean Echenoz	Yasmina Khadra
policier	19.0	11.5
blanche	19.2	13.0

that these differences manifest themselves between the two authors and not between the subgenres.[17]

As mentioned above, the analysis of sentence length as a quantitative criterion of syntactic complexity cannot help to distinguish genres, but it can help to distinguish authors. In the following, the feature of the number of finite verbs will be considered in more detail.

(b) Number of finite verbs

In a previous study it was found that scientific texts (in French) show an average of 1.89 finite verbs/sentence (Hesselbach 2019, 261). For narrative texts, (finite) verbs are of particular importance, since they are used to drive the plot of the story. It is therefore of great interest to see whether narrative texts (of different literary subgenres) show a value that differs from this previous result. The box plot in Figure 3 illustrates the ratios for the entire corpus.

[17] A two-sample t-test was performed to determine the statistical significance of the results: with p = 1.02e-27, it can be concluded that this distribution is highly statistically significant.

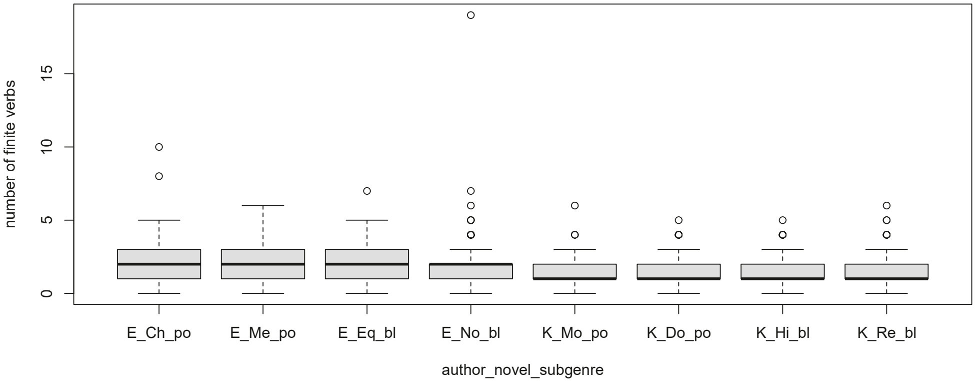


Fig. 3 Distribution of the number of finite verbs (per sentence) of the examined texts (Hesselbach, CC BY).

Table 5 Average number of finite verbs by author

	Jean Echenoz	Yasmina Khadra
policier	2.0	1.4
blanche	1.9	1.5

Again, it becomes clear from the box plot that the criterion of finite verbs also does not give any information about the affiliation of a novel to a certain subgenre. The analysis of the data shows that the distribution is very robust: out of a total of 1,600 sentences, there are only 16 sentences which can be considered statistical outliers. Table 5 shows the mean values for the texts analyzed, broken down by subgenre and author.

It is obvious that Echenoz's values are about 0.5 points higher than Khadra's and almost identical with the results of the previous study mentioned above. Once again, considering the results of Echenoz's texts, one can again ask the question of whether the number of finite verbs are rather an author-specific feature, since they are the only ones in the box plot which can be regarded as more complex (at least for *E_Ch_po*, *E_Me_po*, *E_Eq_bl*) compared to the data received from Khadra's novels, who tends to use less finite verbs in his fiction than Echenoz.[18] Finally, the maximum degree of syntactic embedding is determined in the following.

[18] Again, a two-sample t-test was performed: with p = 9.59e-17, it can be concluded that this distribution is highly statistically significant.

(c) Maximum degree of syntactic embedding

The last quantitative aspect deals with the depth of embedding of a syntactic construction, which for many authors constitutes the criterion par excellence for syntactic complexity (see section 2.2.). For each of the 1,600 sentences, the value for the deepest level of embedding was determined, so that the results can be seen in Figure 4.

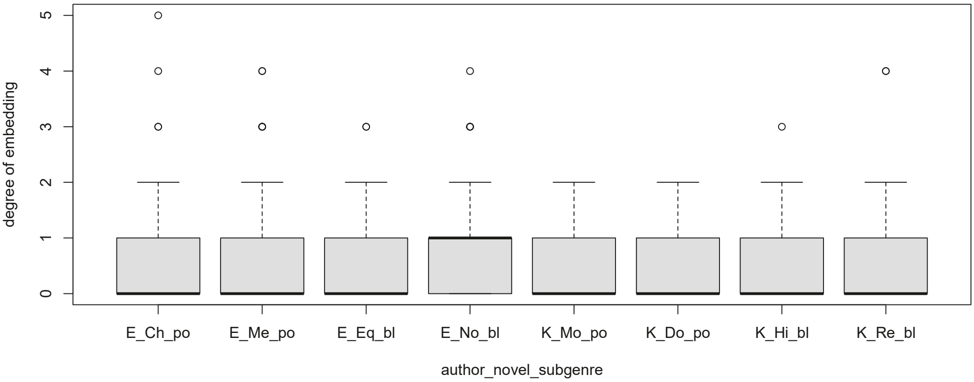


Fig. 4 Distribution of the maximum degree of embedding (per sentence) of the examined texts (Hesselbach, CC BY).

As can be seen from the illustration, there are only extremely few sentences with a depth of embedding of 3 (and more) in the entire corpus, namely a total of only 10, which are visible here as statistical outliers. The data clearly shows—at least for this corpus—that the two authors rely rather on syntactically *flat* constructions, regardless of the literary subgenre. This is interesting in so far as syntactically demanding, i.e. complex, constructions are not necessarily a characteristic of *littérature blanche*, as one might expect. To verify this on the basis of larger datasets would certainly be a rewarding research project. A closer look at the results reveals that Echenoz's sentences again display a more noticeable complexity feature, as can be seen in Table 6.[19]

After having taken a closer look on those complexity features and the results of this study, it has become clear that Echenoz's sentences generally display a greater

[19] Again, a two-sample t-test was performed: with p = 1.29e-12, it can be concluded that this distribution is highly statistically significant.

Table 6 Average maximum degree of embedding by author

	Jean Echenoz	Yasmina Khadra
policier	0.6	0.3
blanche	0.6	0.4

degree of complexity than those of Khadra, and that syntactic complexity must be understood here as an author-specific, rather than a subgenre-specific, feature.

(d) Vector representation

In this subsection, the individually determined quantitative values are now to be put in relation to each other, so that—as described in 2.2.—a vector representation can be generated. Table 7 shows the average quantitative value of the syntactic complexity as a vector for the two investigated subgenres.

Table 7 Average vector values by author and subgenre

	Jean Echenoz	Yasmina Khadra
policier	$\begin{pmatrix} 19.0 \\ 2.0 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} 11.5 \\ 1.4 \\ 0.3 \end{pmatrix}$
blanche	$\begin{pmatrix} 19.2 \\ 1.9 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} 13.0 \\ 1.5 \\ 0.4 \end{pmatrix}$

Therefore, it can be concluded that the two subgenres do not differ fundamentally in terms of quantitative characteristics, as the main differences can be found between the two authors. If one now wants to visualize the results obtained, the individual sentences can be highlighted in color and then plotted as vectors. Blue marked data points represent the sentences from the crime novels (*roman policier*), whereas the red points reflect the sentences from both authors referring to the *littérature blanche*. The following plots in Figure 5, which were generated with *R*,[20] give a very detailed impression of the relation between the two subgenres.

The representations clearly show that the overwhelming majority of the data can be found in the compacted cluster and that only individual, extremely complex sentences, which are furthest away from the zero position, become visible as outliers.

[20] The three-dimensional version of the plot shown here can be seen on the following website: https://rpubs.com/RobertHesselbach/985657.

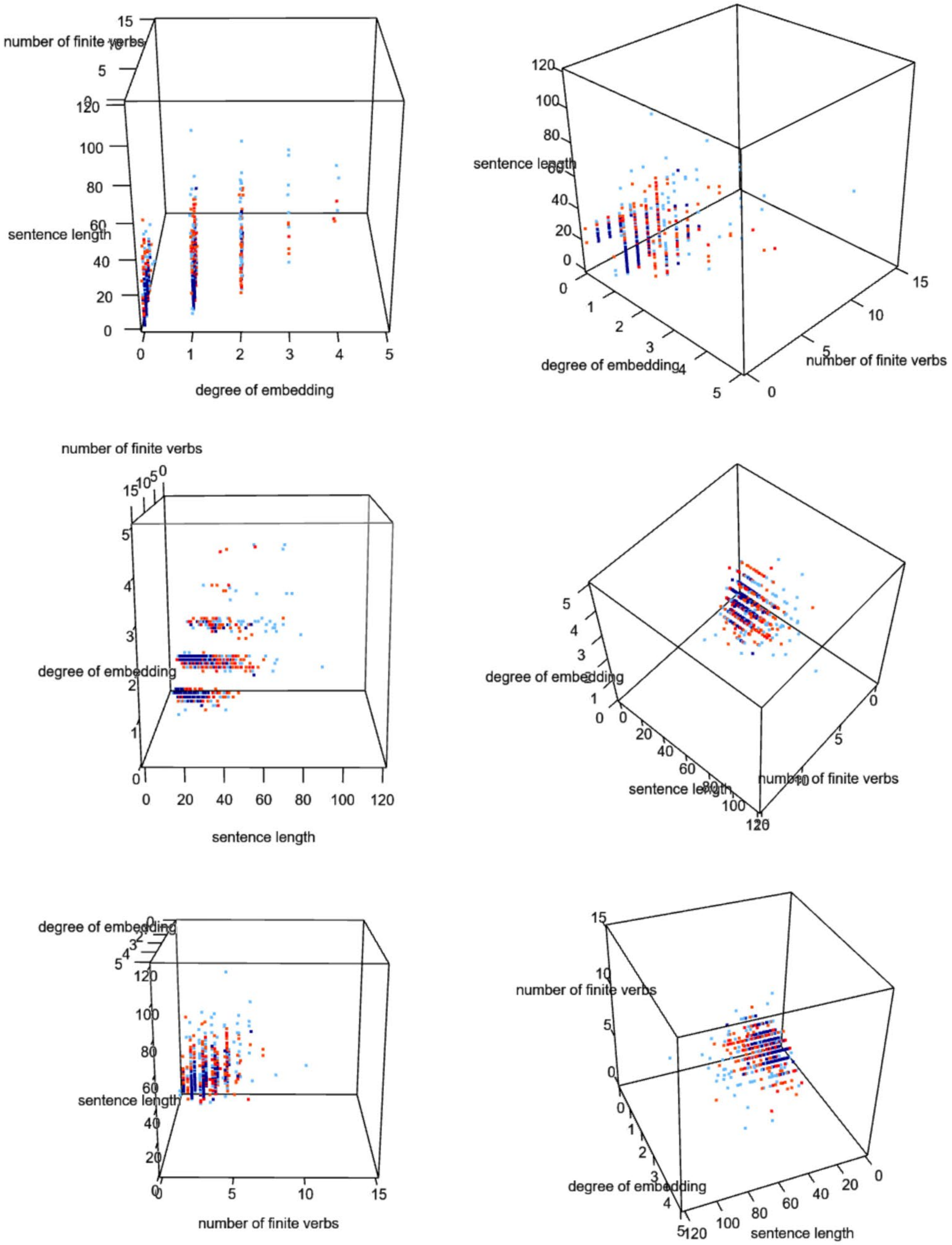


Fig. 5 Visualization of syntactic complexity in vector space for Echenoz's and Khadra's data (Hesselbach, CC BY).

4. Conclusion

The goal of the study presented here was to determine whether the analysis of the syntactic complexity of the sentences of a text can help to distinguish between different subgenres or rather between authors writing in different subgenres. To achieve this, both quantitative and qualitative characteristics were evaluated using a corpus of two contemporary French authors. It is astonishing that the results suggest that there are no significant differences in both quantitative and qualitative terms. On the contrary, a different conclusion can be formulated: the high degree of congruence of both types of sentences and numerically ascertainable values describes an extremely robust distribution which bears witness to the fact that the degree of syntactic complexity for the narrative texts of both subgenres analyzed must be described as almost identical. The corpus analyzed here is admittedly too small to be able to make definitive statements about the distinction between *roman policier* and *littérature blanche* or even further statements with regard to other narrative genres. Nevertheless, further studies (on the same as well as on other genres) with larger amounts of data should provide a better understanding of the relationship between syntactic complexity and (sub)genre distinction.

With regard to syntactic complexity, however, it can be concluded from the available data that its extension seems to be an author-specific characteristic, as we have seen in the analysis of Jean Echenoz's texts, who relies on more complex sentences (in numeric terms) than Yasmina Khadra does. Finally, the combination of quantitative and qualitative results could open up interesting perspectives in the future with regard to the study of the style of individual authors.

Appendix

The research data is available on Zenodo: https://doi.org/10.5281/zenodo.7458279.

Acknowledgements

The research group was funded by the German Ministry for Education and Research (BMBF), and I would like to thank them for their support. A further thank you goes to Fotis Jannidis, Christof Schöch, Ulrike Henny-Krahmer, José Calvo Tello, Daniel Schlör, Steffen Pielström, Julian Schröter and Barbara Gersitz for being great colleagues at the University of Würzburg. I am also grateful to Christopher Geekie (Erlangen-Nürnberg) for his valuable advice.

ORCID®

Robert Hesselbach https://orcid.org/0000-0001-9758-8290

References

Altmann, Gabriel. 1978. “Zur Verwendung der Quotiente in der Textanalyse.” In *Glottometrika*, edited by Gabriel Altmann, 91–106. Bochum: Brockmeyer.

Altmann, Gabriel, and Reinhard Köhler. 2000. “Probability Distributions of Syntactic Units and Properties.” *Journal of Quantitative Linguistics* 7 (3): 189–200.

Altmann, Gabriel, Karl-Heinz Best, and Ioan-Iovitz Popescu. 2014. *Unified Modeling of Length in Language*. Lüdenscheid: RAM-Verlag.

Best, Karl-Heinz. 2005. “Satzlänge.” In *Quantitative Linguistik*, edited by Gabriel Altmann, Reinhard Köhler and Rajmund G. Piotrowski, 298–304. Berlin/Boston: De Gruyter.

Calvo Tello, José. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld: transcript.

Dubois, Jean et al. 1994. *Dictionnaire de linguistique et des sciences du langage*. Paris: Larousse.

Ferreira, Fernanda. 1991. “Effects of Length and Syntactic Complexity on Initiation Times for Prepared Utterances.” *Journal of Memory and Language* 30 (2): 210–33.

Fucks, Wilhelm. 1968. *Nach allen Regeln der Kunst: Diagnosen über Literatur, Musik, bildende Kunst – die Werke, ihre Autoren und Schöpfer*. Stuttgart: Deutsche Verlags-Anstalt.

Gianitsos, Efthimios Tim, Thomas J. Bolt, Pramit Chaudhuri, and Joseph P. Dexter. 2019. “Stylometric Classification of Ancient Greek Literary Texts by Genre.” *Proc. of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 52–60. https://www.aclweb.org/anthology/W19-2507.pdf

Givón, Talmy. 2009. *The Genesis of Syntactic Complexity – Diachrony, ontogeny, neuro-cognition, evolution*. Amsterdam: John Benjamins.

Givón, Talmy, and Masayoshi Shibatani, eds. 2009. *Syntactic Complexity – Diachrony, Acquisition, Neuro-Cognition, Evolution*. Amsterdam: John Benjamins.

Henny-Krahmer, Ulrike. 2018. “Exploration of Sentiments and Genre in Spanish American Novels.” *DH 2018*. https://dh2018.adho.org/exploration-of-sentiments-and-genre-in-spanish-american-novels/

Hesselbach, Robert. 2019. *Diaphasische Variation und syntaktische Komplexität – eine empirische Studie zu funktionalen Stilen des Spanischen mit einem Ausblick auf das Französische*. Berlin, Boston: De Gruyter.

Hesselbach, Robert. (in prep.). “Sobre la complejidad sintáctica de textos literarios del español a través del tiempo.”

Hettinger, Lena, Isabella Reger, Fotis Jannidis, and Andreas Hotho. 2016. "Classification of Literary Subgenres." In *DHd 2016 Digital Humanities. Konferenzabstracts*, 154–58. Leipzig: Universität Leipzig. http://www.dhd2016.de/abstracts/vortr%C3%A4ge-049.html.

Jing, Zhuo. 2001. "Satzlängenhäufigkeiten in chinesischen Texten." In *Häufigkeitsverteilungen in Texten*, edited by Karl-Heinz Best, 202–10. Göttingen: Peust & Gutschmidt.

Johnson, Neal. 1966. "On the Relationship between Sentence Structure and the Latency in Generating the Sentence." *Journal of Verbal Learning and Verbal Behavior* 5 (4): 375–80.

Kiesler, Reinhard. 2013. "Pour une typologie des phrases complexes." *Zeitschrift für romanische Philologie* 129 (3): 608–28.

Koch, Peter. 1995. "Subordination, intégration syntaxique et 'oralité'." In *La subordination dans les langues romanes. Actes du colloque international, Copenhague 5.5.–7.5.1994*, edited by Hanne Leth Andersen and Gunver Skytte, 13–42. Copenhagen: Munksgaard.

Raible, Wolfang. 1992. *Junktion – eine Dimension der Sprache und ihre Realisierungsformen zwischen Aggregation und Integration*. Heidelberg: Winter.

Schöch, Christof. 2017. "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Sowinski, Bernhard. 1999. *Stilistik: Stiltheorien und Stilanalysen*. 2nd ed. Stuttgart: Metzler.

Szmrecsányi, Benedikt. 2004. "On Operationalizing Syntactic Complexity." In *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis. Louvain-la-Neuve, March 10–12, 2004*, edited by Gérard Purnelle, Cédrick Fairon, and Anne Dister, 1032–39. Louvain-la-Neuve: Presses universitaires de Louvain.

Trabant, Jürgen. 1981. "Wissenschaftsgeschichtliche Bemerkungen zur Textlinguistik." In *Beiträge zur Linguistik des Französischen*, edited by Thomas Kotschi, 1–20. Tübingen: Narr.

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

Clémence Jacquot, Ilaria Vidotto, and Laetitia Gonon

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

Keywords *PhraseoRom*, stylistic annotation, *motifs*

1. Introduction

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

[1] Agence Nationale de la Recherche & Deutsche Forschungsgemeinschaft.

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

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

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

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

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

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

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

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

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

2. Investigating the Novel Genre through Extended Phraseology

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

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

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

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

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

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

Working on the specific language of the novel requires an investigation of its generic boundaries and the *values* that derive from them (see Jouve 2010). This dictates the inclusion of works in the corpus that, by editorial tradition and ideological representation of the novel's subgenres, are categorized by French literature specialists as *paralittérature* (see Couégnas 1992; Boyer 2002), i.e. popular fiction.

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

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

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

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

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

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

3. Stylistic Corpus Annotation: Methodological and Epistemological Issues

3.1 Traditional Stylistics vs. *Digital Stylistics*

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

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

3.2 Stylistic Annotation

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

3.2.1 Step 1: Extracting RLTs

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

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

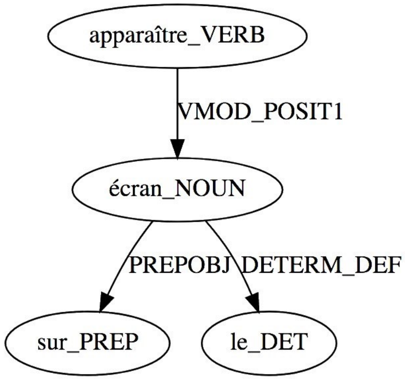


Fig. 1 Example of the RLT extraction <*apparaître sur l'écran*> ('appear on screen'), (Jacquot, Vidotto, Gonon, CC BY).

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

3.2.2 Step 2: Selecting and Semantically Tagging LSCs

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

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

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

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

3.2.3 Step 3: The Discursive Functions

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

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

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

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

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

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

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

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

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

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

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

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

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

[7] M. Villard, *Cœur sombre*, 1997 (CRIM). N.B. This excerpt from French novel and all the following excerpts mentioned in this article have been translated by the authors and appear in quotation marks.

[8] J.-Ch. Rufin, *Sauver Ispahan*, 1998 (GEN).

[9] Then follows a descriptive sequence, triggered by the *motif*.

[10] J.-C. Grangé, *Le Vol des cigognes*, 1994 (CRIM).

[11] O. Gay, *Les Talons hauts rapprochent les filles du ciel*, 2012 (POL).

[12] K. Giébel, *Juste une ombre*, 2012 (CRIM).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[18] M. Chattam, *Le Cycle de l'homme 1*, *Les Arcanes du chaos*, 2013 (CRIM).

[19] A. H. Japp, *Cinq filles, trois cadavres, mais plus de volant*, 2009 (CRIM).

4. Project Perspectives and Annotation Issues

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

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

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

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

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

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

Before concluding this section on stylistic annotation, a more precise definition of the concept of *motif* is called for. Up to this point, it has provisionally been defined as a "relevant grouping of LSCs." As we have seen, LSCs already provide interesting information for work on differentiating fiction genres, but their description does not take into consideration the textual dimension of the corpus, i.e. their role in structuring texts. This is where the concept of *motif* comes in by making it possible to integrate its discursive component into the phraseological dimension of this analysis.

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

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

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

6. Stylistic Annotation and the Granularity of *Motifs*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

7. Conclusion: Stylistic Analysis as Starting and End Point

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

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

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

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

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

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

— Upstream: it creates the literary coherence of the corpora and provides elements of literary problematization of the data, for example diegetic stereotypy versus linguistic stereotypy. Furthermore, stylistic annotation contributes to the process of building up a corpus by conceptualizing and complexifying *motifs* through the identification of DFs.

— Downstream: stylistic analyses facilitate the study of extracted data and stimulate critical reflection on conventional boundaries (literary criticism, academic work, publishing house collections, etc.) between subgenres by shedding light on the very definitions—linguistic, phraseological, stylistic—of the boundaries.

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Complexity and Style of Modern Spanish Literary Texts

Katharina Dziuk Lameira

Abstract This article discusses whether text complexity can be seen as a dimension of authorial style, and if so, which linguistic features are suitable for the description of both complexity and style. First, the notion of text complexity will be defined, followed by a presentation of its evolution from readability studies. In addition, a comparison regarding definitions of text complexity and literary style will be followed by various parameters of text complexity used in a research project on the linguistic complexity of modern Spanish literary texts conducted at the University of Kassel, presented together with the tools and methods used for their analysis. In conclusion, an investigation concerning a selection of parameters and data from the project regarding a stylistic point of view will be offered, prior to the results being discussed and future research steps being proposed. First analyses show that metaphorical density, sentence length, clause length, the index of subordination, the density of the noun phrase, and logarithmic average (word) frequency and combinations of some of those parameters could be suitable for the description of some of the novels analyzed in this article.

Keywords text complexity, authorial style, readability, Spanish, literary texts, metaphor identification

1. Introduction

Text complexity is defined as the interaction of different textual features that influence text difficulty and can be measured and observed objectively (Dziuk Lameira 2023). Following Dahl's definition of relative complexity (2004), a text can be seen as more complex the more it deviates from typical patterns.[1] Analogously, different definitions

[1] *Pattern* refers here to a variety of phenomena such as linguistic patterns like phrasemes or patterns regarding textual genres.

Dziuk Lameira, Katharina: Complexity and Style of Modern Spanish Literary Texts, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 279–297.
DOI: https://doi.org/10.17885/heiup.1157.c19376

for literary style have been proposed, e.g. as the interaction of textual features or deviation from patterns in texts.[2] These parallel definitions lead to the questions of if and to what extent text complexity can be seen as a dimension of literary style and vice versa. Given that many of the variables investigated in a current research project conducted at the University of Kassel by Dziuk Lameira on linguistic complexity of modern Spanish texts are based on textual features that are often used in stylistics as well, the data from the project will be explored in this article from a stylistic point of view.

2. Text Complexity

2.1 The Notion of *Text Complexity*

Text complexity is investigated here from a linguistic point of view. It can be defined as a property of a text that emerges from the interaction of different text features and levels that can be measured and observed objectively (Dziuk Lameira 2023). *Text complexity* and *text difficulty* are seen here as different concepts. *Text difficulty* is considered a consequence of the interaction of text complexity and extralinguistic parameters, such as reader or task characteristics (e.g. motivation, cognitive capabilities or type of task). The definition offered here regarding linguistic text complexity also excludes the difficulty of content that can also lead to text difficulty.

2.2 From Readability to Text Complexity

Research on text complexity can be traced back to the first readability formulae that have been developed mostly by psychologists and educational researchers since the 1920s in order to predict the comprehensibility of texts for specific groups of readers. Mikk describes the standard way of finding a readability formula as follows:

> To elaborate a readability formula, a sample of texts, representative of the texts in the area of intended formula application, should be taken. The comprehension level of the texts is measured by some experiments and the texts are analysed to establish the values of the hypothesised measures for the text comprehensibility. The comprehension level and the comprehensibility measures are tied in a formula by multiple regression analysis. The most valid compre-

[2] See Fix (2009) for an overview on deviation and pattern in rhetoric and stylistics.

> hensibility measures intercorrelations of which are low [sic] are included in the readability formula as the result of the analysis (2005, 913).

Which comprehensibility measures are valid can vary depending on text type and target group, e.g. Mikk found over 200 valid comprehensibility measures for popular scientific texts read by students (2000). However, many studies have shown that word difficulty (mostly measured by word length and/or word frequency) and sentence difficulty (mostly measured by sentence length), account for most of the variance in readability measures, as the addition of other variables could scarcely improve the prediction of comprehensibility (Entin and Klare 1978; Klare 1974; 1984). For this reason, most readability formulae make use of those variables. Readability formulae are often criticized for only relying on superficial textual features neglecting reader and task characteristics (although those can also be integrated into a formula, e.g. Mikk and Elts 1999). The terminological shift from *readability* to *text complexity* illustrates the attempt to integrate quantitative and qualitative text features and reader and task characteristics to a model of text complexity.[3]

2.3 Complexity: A Complex Notion

Complexity itself is a complex notion, as there is no generally accepted definition of complexity and the existing definitions vary depending on the discipline or the object of investigation. The Santa Fe Institute defines complexity as follows: "In general, the complexity of a system emerges from the interactions of its interrelated elements as opposed to the characteristics of those elements in and of themselves" (Santa Fe Institute 2018). Rescher defines it from a philosophical point of view in the following way: "Complexity is first and foremost a matter of the number and variety of an item's constituent elements and of the elaborateness of their interrelational structure, be it organizational or operational" (1998, 1). Both definitions underline the importance of interrelation and interaction in complex systems. According to Rescher (1998)[4], the notion of complexity encompasses various modes (see Table 1).

First, Rescher mentions the EPISTEMIC MODES of complexity: *descriptive complexity*, which refers to the length of a description of a system; *generative complexity* or the number of instructions necessary to generate a system; and *computational complexity* referring to the time and resources needed to solve a problem.

[3] See e.g. the Common Core State Standards model of Text Complexity (CCSSO 2010).

[4] Rescher's definition and systematization of complexity have been part of the discussion about complexity in linguistics since their reception by Karlsson, Miestamo, and Sinnemäki (2008).

Table 1 Modes of Complexity according to Rescher (1998, 9)

Epistemic modes	
Formulaic complexity	*Descriptive complexity*
	Generative complexity
	Computational complexity
Ontological modes	
Compositional complexity	*Constitutional complexity*
	Taxonomic complexity (or heterogeneity)
Structural complexity	*Organizational complexity*
	Hierarchical complexity
Functional complexity	*Operational complexity*
	Nomic complexity

Secondly, the author lists the ONTOLOGICAL MODES of complexity, which are compositional, structural, and functional complexity.

Compositional complexity is divided into *constitutional complexity*, which refers to the number of elements that constitute an object (relating to texts this could be the number of words or paragraphs); and *taxonomic complexity (*or *heterogeneity)*, which refers to the different types of elements in a system (e.g. modes and tenses used in a text).

Moreover, there are two types of structural complexity: *organizational* and *hierarchical complexity*. *Organizational complexity* alludes to the number of different ways to organize the different elements of an object; whereas *hierarchical complexity* refers to the degree of elaborateness of the hierarchical relations in the system (e.g. levels of syntactic subordination).

The last type that Rescher mentions is functional complexity, which encompasses *operational* and *nomic complexity* (1998, 9). *Operational complexity* relates to the number of different possible functions and states that can arise during a process. The higher the number of possible states, the less predictable the behavior of a system is (which leads to higher complexity) (Rescher 1998, 12–13). *Nomic complexity* finally refers to the structures and laws that govern a system (ibid., 13).

2.4 Absolute and Relative Complexity

In the study of linguistic complexity, the differentiation between *absolute* and *relative complexity* is often made. According to Miestamo, absolute complexity is theory-oriented and objective whereas relative complexity is receiver-oriented and subjective (2006, 2008). The relative notion of complexity can be illustrated by the question:

"Complex to whom?" (Miestamo 2006, 3; Kusters 2003, 6) and Miestamo argues that this form of complexity should be called "cost" or "difficulty" (2008, 25–26).

Dahl uses the term *relative complexity* in a different sense:

> An entity E would have a certain complexity relative to a description or theory T measured by the length of the additional description necessary to characterize E provided that T is already given. [...] A theory of a class of entities may specify (or predict) the properties that are common to all the members of the class. However, it may go beyond that and also specify properties that are typical of or 'normal' for the members of the class—what holds in a default or prototypical case. The description of each member may then be considerably simplified, given that only deviations from the normal case have to be specified. An interesting consequence that now appears is that an entity which deviates from the default case in more respects will tend to be more complex, in this sense (2004, 25–26).

According to him, a deviation from the normal case leads to a longer description and thereby to higher complexity. He exemplifies this by saying that when describing a person, we don't have to mention that "she has two legs, two arms and one head" (Dahl 2004, 25), because we know what human beings look like in general. Transferred to text complexity, this could mean that the complexity of a text is higher the more it deviates from the typical text of its class, because the description of its properties is longer.

In a similar way, Merlini Barbaresi (2011) argues that texts can be seen as complex systems that are characterized on different levels by "markedness/naturalness," in which markedness leads to higher degree of complexity (for a criticism on the use of the term *markedness* see Haspelmath 2006).

2.5 Text Complexity and Text Difficulty

Text complexity can be defined as a property of a text that emerges from the interaction of different text features and levels that influence text difficulty and can be measured and observed objectively, whereas text difficulty can be seen as the consequence of the interaction of text complexity and extralinguistic parameters that can be perceived subjectively (Dziuk Lameira 2023). Extralinguistic parameters are reader characteristics such as age, educational background, reading experience, motivation etc.; and task characteristics such as the translation of a text versus the retrieval of information. Thus, *text complexity* can be seen as the cause and *text difficulty* as the effect.

According to Rescher, complexity and cognitive difficulty are two different concepts that are coordinated:

> As an item's complexity increases, so do the cognitive requisites for its adequate comprehension, although, of course, cognitive ineptitude and mismanagement can manage to complicate even simple issues. All the same, our best practical index of an item's complexity is the effort that has to be expended in coming to cognitive terms with it in matters of description and explanation (1998, 1).

Another important issue is how linguistic text complexity can be measured, or if it is even possible to measure text complexity. If we assume that it is possible, there are two possibilities: First, text complexity can be measured directly by means of measurement categories like the type and number of relations between elements etc.[5]

Second, complexity could be measured indirectly by measuring its effect (e.g. the cost caused by the complexity). In the case of texts, the cost of complexity could be the difficulty that a reader experiences reading the text.

However, according to the Santa Fe Institute, which is known for its studies on complex (adaptive) systems, "complex behavior generally cannot be reduced to, or derived from, the sum of the behavior of the system's components" (Santa Fe Institute 2018). Although the definition of complex adaptive systems is not perfectly fitted to texts, many complexity definitions emphasize the important role of interrelations in complex systems. A model of text complexity has to be able to measure or at least describe these interrelations. An interesting approach comes from Stede (2018), who suggests that the interplay of features on different text levels can be analyzed by annotating the structure of texts in a multilevel annotation in order to find correlations between the different text levels. The approach is called *level-oriented text linguistics* ("Ebenen-orientierte Textlinguistik") and presents different tools for the analysis of separated text levels as well as a database that comprises all annotations (Stede 2018). In the case that those multilevel annotations don't exist, the qualitative text analyses (e.g. semantic text analysis according to Gardt 2012) can help to comprehend the nature of interactions in texts.

2.6 Text Complexity and Literary Style

The comparison of definitions of the notions of complexity and literary style show two parallel branches:

1. the definition of text complexity and literary style as ensembles of textual features
2. the definition of text complexity and literary style as deviation

[5] See Rescher's modes of complexity (section 2.3)

As mentioned above, text complexity is defined here as a property of a text that emerges from the interaction of different text features and levels that influence text difficulty and can be measured and observed objectively. Another emergent feature of literary texts is their style. The notions of complexity and style of a text are defined in similar ways. According to Herrmann, van Dalen-Oskam, and Schöch "[s]tyle is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively" (2015, 44). The emergent character of *style* is not mentioned by the authors, but should be considered, as well. The main difference between the two notions is the effect, which is difficulty in the case of complexity, and the perception of an (individual) style in the case of style.

The second possible definition of *style* is style as deviation from typical patterns or reader expectations,[6] which connects back to Dahls' definition of complexity as a deviation from the normal case.

As the definitions of *style* and *complexity* are very similar, it is valid to ask if linguistic text complexity can be seen as a dimension of literary style and vice versa.

This can be understood in two ways:

- the overall complexity of a text could be seen as a dimension of style
- the parameters (text features) that are suited to measure or describe complexity could also be suited to identify a style

One difference between *style* and *complexity* is that the description of style is often not made gradually but, in the case of author identification, with the aim being to attribute a text to a specific author. Complexity on the other hand is always a matter of graduality or degree.

As the description of the overall complexity of a text can vary depending on the parameters chosen by the researcher, the only way to see if the overall complexity of a text can be related to its style would be to study this question indirectly by comparing the difficulty of texts judged by readers or experts with a facet of style, e.g. authorship or register. Another possibility would be to study isolated or combined text features for their suitability to discriminate a certain facet of style. It is this second possibility that is being tested in this paper.

[6] See Fix (2009) for an overview on deviation and pattern in rhetoric and stylistics.

3. Analyzing Text Complexity

3.1 Project on Linguistic Complexity Profiles of Modern Spanish Texts

In a current research project conducted at the University of Kassel (Germany) by Dziuk Lameira, first the complexity of a text collection consisting of 30 Spanish literary text excerpts from novels published between 2004 and 2017 is analyzed quantitatively and qualitatively regarding lexical, semantic, syntactical, morphological, and textual features to provide a complexity profile for each text. One half of the novels were read in the B1 Spanish course and the other half in the B2 course at the University of Kassel. The choice of the novels was based on didactic considerations. In a second step, six representative texts were chosen from the corpus for an online-questionnaire developed by Friedrich (2017) that was presented to German-speaking students of Spanish philology which were grouped according to their Spanish proficiency level (CEFR[7] levels A2–C1). The students rated the difficulty of the given texts and answered questions concerning the difficulty of those texts. The mean values of these text ratings were used for a statistical analysis to identify text features contributing to text complexity. In this paper, various parameters that are usually used to measure text complexity will be tested for their suitability for distinguishing different authorial styles.

The analyzed text excerpts were taken from the following novels (three excerpts per novel):

- Javier Cercas: *El impostor* (2014)
- Juan Gabriel Vásquez: *El ruido de las cosas al caer* (2010)
- Almudena Grandes: *El lector de Julio Verne* (2012)
- Javier Marías: *Así empieza lo malo* (2014)
- Eduardo Mendoza: *Riña de gatos. Madrid 1936* (2010)
- Juan José Millás: Hay algo que no es como me dicen (2004)
- Edurne Portela: *Mejor la ausencia* (2017)
- Carme Riera: *Naturaleza casi muerta* (2011)
- Andrea Stefanoni: *La abuela civil Española* (2014)
- Andrés Trapiello: *Ayer no más* (2012)

The excerpts were compiled based on the following criteria: They are between 308 and 353 words long and are understandable without context in order to be used in the questionnaire. Furthermore, they contain no direct speech (with or without *inquit*

[7] Common European Framework of Reference for Languages.

formulae) and no parts of the original text are omitted or changed. Additionally, they contain at least one metaphor and can be assigned to narrative and/or descriptive text types. The analyzed text compilation contains 9,759 words in total.

3.2 Quantitative Complexity Parameters

The quantitative analysis was carried out by means of the program TRUNAJOD (Véliz and Karelovic) and different readability formulae like Fernández Huerta (1959), Gutiérrez (1972), Szigriszt Pazos (1993), INFLESZ (Barrio-Cantalejo et al. 2008), legibilidad µ (Muñoz Baquedano and Muñoz Urra 2019) etc., which are validated or adapted for Spanish texts and available online.[8]

The program TRUNAJOD (Véliz and Karlovic) can determine the readability of texts in Spanish and is compatible with the tagger Connexor Machinese Syntax. It calculates the following indices:

- LO (*longitud de la oración*): sentence length
- LC (*longitud de la cláusula*): clause length
- IS (*índice de subordinación*): index of subordination
- DeP *(índice de densidad proposicional*): propositional density
- DeL (*índice de densidad léxica*): lexical density
- DiL (*índice de diversidad léxica*): lexical diversity
- DFN (*densidad de la frase nominal*): density of the noun phrase
- FP (*frecuencia promedio de palabras*): average word frequency
- FPL (*frecuencia promedio logarítmica*): logarithmic average frequency

3.3 Semantic Complexity Parameters

Semantic text complexity consists of predictors based on semantic information. It is contested whether semantic text features are suitable parameters for the measurement or description of text complexity and readability. François and Fairon developed a model of readability for learners of French as a foreign language based on support vector machines which allows a better prediction of readability than classic readability formulae (2012, 466). According to the authors "the information carried by semantic predictors is largely correlated with that of lexico-syntactical ones" (ibid., 475). Their explanation is that "semantic and lexical predictors are correlated because the methods used for the parameterization of the semantic factors heavily rely on lexical

[8] E.g. https://legible.es or LEXILE.

information. This is the case for the LSA,[9] as well as for the propositional approach of the content density" (ibid.). This observation leads to the question of how the semantic dimension of text complexity can be operationalized.

3.4 Metaphorical Text Complexity

In order to explore possibilities to operationalize the semantic dimension of text complexity, the field of metaphorical text complexity was focused on in the project mentioned above.

The following parameters were thus suggested for the measurement of metaphorical text complexity (see also Dziuk Lameira 2019; Dziuk Lameira 2023):

- METAPHORICAL DENSITY: The number of metaphor related words divided by the number of lexical units[10] multiplied by 100 (MD = MRW[11]/lexical units × 100)
- METAPHORICAL VARIETY: The number of different concept combinations (conceptual metaphors) divided by the number of metaphor related words (MV = Number of different concept combinations/MRW)
- NNMRW: The percentage of non-nominal metaphor related words within the total of metaphor related words (Percentage of non-nominal MRWs)
- EXTENDED METAPHORS: The number of extended metaphors (EM) divided by the number of lexical units multiplied by 100 (EM/lexical units × 100)
- DEGREE OF CONVENTIONALIZATION: The percentage of lexicalized and new metaphors per text

Other parameters that are analyzed qualitatively include: CO-TEXTUALIZATION (Skirl 2009), the interaction with similes,[12] and the revitalization of metaphors (Goatly 1997). Metaphors were counted as proposed by MIPVU (Metaphor Identification

[9] Latent Semantic Analysis (LSA), also known as latent semantic indexing, is a method developed by T. K. Landauer for the semantic analysis of document collections. The objective is to extract latent concepts, which are relevant terms and co-occurrences, from the documents. In LSA analysis, documents are represented in the form of a term-document matrix (vector space model), common stop words are filtered out, frequently occurring terms are specifically weighted, and a singular value decomposition is performed (a method of linear algebra to reduce the number of dimensions per document) (Glück and Rödel 2016, 390).

[10] Unit of analysis when applying the Metaphor Identification Procedure Vrije Universiteit (MIPVU) (Steen et al. 2010).

[11] Metaphor related words (see Steen et al. 2010).

[12] This parameter could be as well seen as a subparameter of co-textualization.

Procedure Vrije Universiteit, Steen et al. 2010) in order to increase the objectivity of metaphor identification.

The procedure was developed by Pragglejaz Group (2007) as MIP (Metaphor Identification Procedure) and further developed by Steen et al. (2010) under the name of MIPVU. The method consists of the following steps (ibid., 25–26):

1. Find metaphor related words (MRWs) by examining the text on a word-by-word basis.
2. When a word is used indirectly and that use may potentially be explained by some form of cross-domain mapping from a more basic meaning of that word, mark the word as metaphorically used (MRW).
3. When a word is used directly and its use may potentially be explained by some form of cross-domain mapping to a more basic referent or topic in the text, mark the word as direct metaphor (MRW, direct).
4. When words are used for the purpose of lexico-grammatical substitution, such as third person personal pronouns, or when ellipsis occurs where words may be seen as missing, as in some forms of coordination, and when a direct or indirect meaning is conveyed by those substitutions or ellipses that may potentially be explained by some form of cross-domain mapping from a more basic meaning, referent, or topic, insert a code for implicit metaphor (MRW, implicit).
5. When a word functions as a signal that a cross-domain mapping may be at play, mark it as a metaphor flag (MFlag).
6. When a word is a new-formation coined, examine the distinct words that are its independent parts according to steps 2 through 5.

Although first attempts to automatize metaphor identification exist (e.g. Berber Sardinha 2010; Rai et al. 2016; Rai and Chakraverty 2017), they are still not ready for application and therefore the metaphor identification has to be carried out manually.

4. Analysis

4.1 Preliminary Results

In the following, the complexity parameters calculated by the software TRUNAJOD and one parameter for metaphorical complexity (MD: metaphorical density) have been tested for their ability to discriminate different authorial styles. As the text compilation used for the analyses contains 30 excerpts from ten different novels written by different authors, *authorial style* was here defined as the affiliation of a text excerpt to a novel. Since the text compilation includes three excerpts per novel, the question is whether

Table 2 Kruskal Wallis Test

	MD	LO	LC	IS	DeP	DeL	DiL	DFN	FP	FPL
Kruskal Wallis H	19,916	19,957	17,434	17,730	9,713	10,729	9,959	18,528	12,742	18,841
df	9	9	9	9	9	9	9	9	9	9
Asymptotic significance	*0.018*	*0.018*	*0.042*	*0.038*	0.374	0.295	0.354	*0.030*	0.175	*0.027*

the chosen quantitative variables (MD: metaphorical density, LO: sentence length, LC: clause length, IS: index of subordination, DeP: propositional density, DeL: lexical density, DiL: lexical diversity, DFN: density of the noun phrase, FP: average frequency and FPL: logarithmic average frequency) show a significant difference between novels.

Table 2 shows the results of the Kruskal Wallis Test. The null hypothesis that the distribution of a variable is the same across the novels is rejected for the variables MD, LO, LC, IS, DFN and FPL. Hence, there are significant differences between novels for these variables.

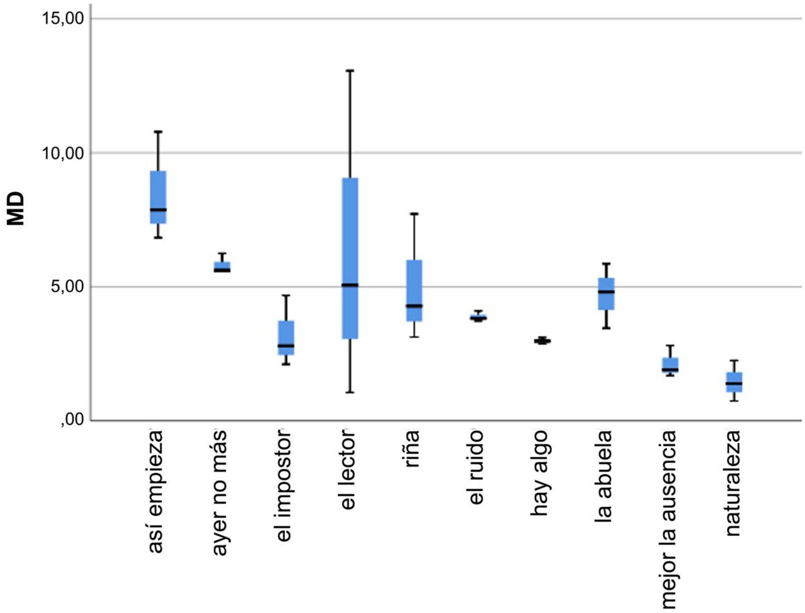


Fig. 1 Boxplot comparison for variable MD (Dziuk Lameira, CC BY).

The boxplot diagrams in Figure 1 visualize the spread within novels as well as differences between medians for the variable MD (metaphorical density). Although tendencies can be recognized, the pairwise comparison of novels for MD using t-test only shows a significant difference between the novels *Naturaleza casi muerta* and *Así empieza lo malo* (see Table 3). This shows that metaphorical density can be a valid criterion for the differentiation of novels and therefore should be further investigated.

Table 3 Pairwise comparison of novels for the variable MD

Sample 1-Sample 2	N	Test Statistic	Std. error	Std. Test Statistic	Sig.	Adj. Sig.[a]
naturaleza-así empieza	3	24,000	7,188	3,339	0.001	0.038

a Significance values have been adjusted by the Bonferroni correction for multiple tests.

When using the Bonferroni-Holm correction for multiple tests the only significant difference remains between the novels *Naturaleza casi muerta* and *Así empieza lo malo*.

At the same time, Figure 1, as well as Figure 2, shows the high spread of the MD and LO values within some of the novels, e. g. the three excerpts from *El lector de Julio Verne* written by Almudena Grandes show a high spread for the variable MD (metaphorical density) whereas the three excerpts from the novel *El impostor* by Javier Cercas show a high spread for the variable LO (sentence length). (the same could be observed for the other analyzed variables). The high deviation paired with the small sample size per novel influences the suitability of the variables.

The results for the Pearson test show a very high correlation for the variables LO (sentence length) and IS (index of subordination) ($r = 0.95$, $p < 0.001$) as well as for the variables FP (average frequency) and FPL (logarithmic average frequency) ($r = 0.875$, $p < 0.001$). The reason for the correlation between LO and IS could be that longer sentences have a higher probability of containing more clauses. FP and FPL could be correlated because both variables measure word frequency. Because of the correlation, one of each correlating variables was discarded from the following cluster analysis. In the case of FP and FPL, FPL, the logarithmic average frequency, was maintained, because high frequency words have a lower impact when calculating the index. In the case of the correlating variables LO and IS, LO, the sentence length, was maintained, because many studies have shown its importance as a predictor of text comprehensibility (Entin and Klare 1978; Klare 1974; 1984).

The dendrogram in Figure 3 shows the result of the cluster analysis using the variables MD (metaphorical density), LO (sentence length), LC (clause length), DeP (propositional density), DeL (lexical density), DiL (lexical diversity), DFN (density of the noun phrase) and FPL (logarithmic average frequency). The cluster analysis was performed with squared Euclidean distance, z-score standardization, and Ward linkage. Looking at the distribution of the text excerpts when 10 clusters are defined (see

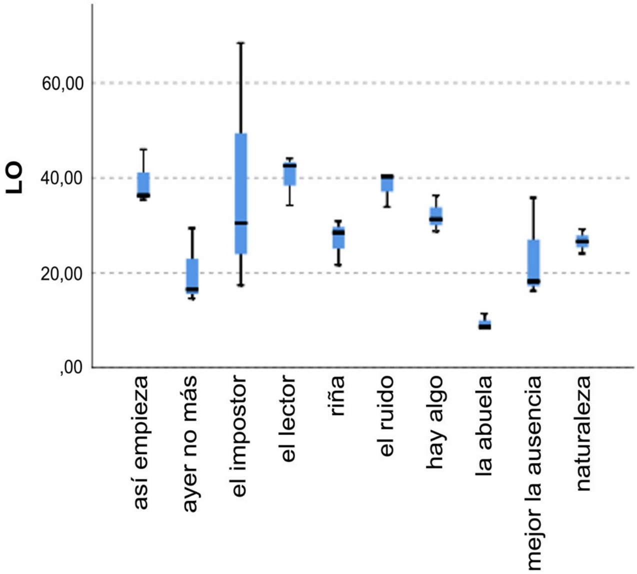


Fig. 2 Boxplot comparison for variable LO (Dziuk Lameira, CC BY).

Table 4 Overview of clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Así empieza 1	Ayer no más 1	Ayer no más 2	Ayer no más 3	Impostor 3
Así empieza 2	Impostor 1	La abuela 2	Mejor la ausencia 2	
Así empieza 3	Impostor 2	La abuela 3		
Cluster 6	**Cluster 7**	**Cluster 8**	**Cluster 9**	**Cluster 10**
El lector 1	El lector 2	El lector 3	Rina 1	La abuela 1
	Hay algo 2	El ruido 3	Rina 2	Mejor la ausencia 1
	Naturaleza 1		Rina 3	
	Naturaleza 2		El ruido 1	
	Naturaleza 3		El ruido 2	
			Hay algo 1	
			Hay algo 3	
			Mejor la ausencia 3	

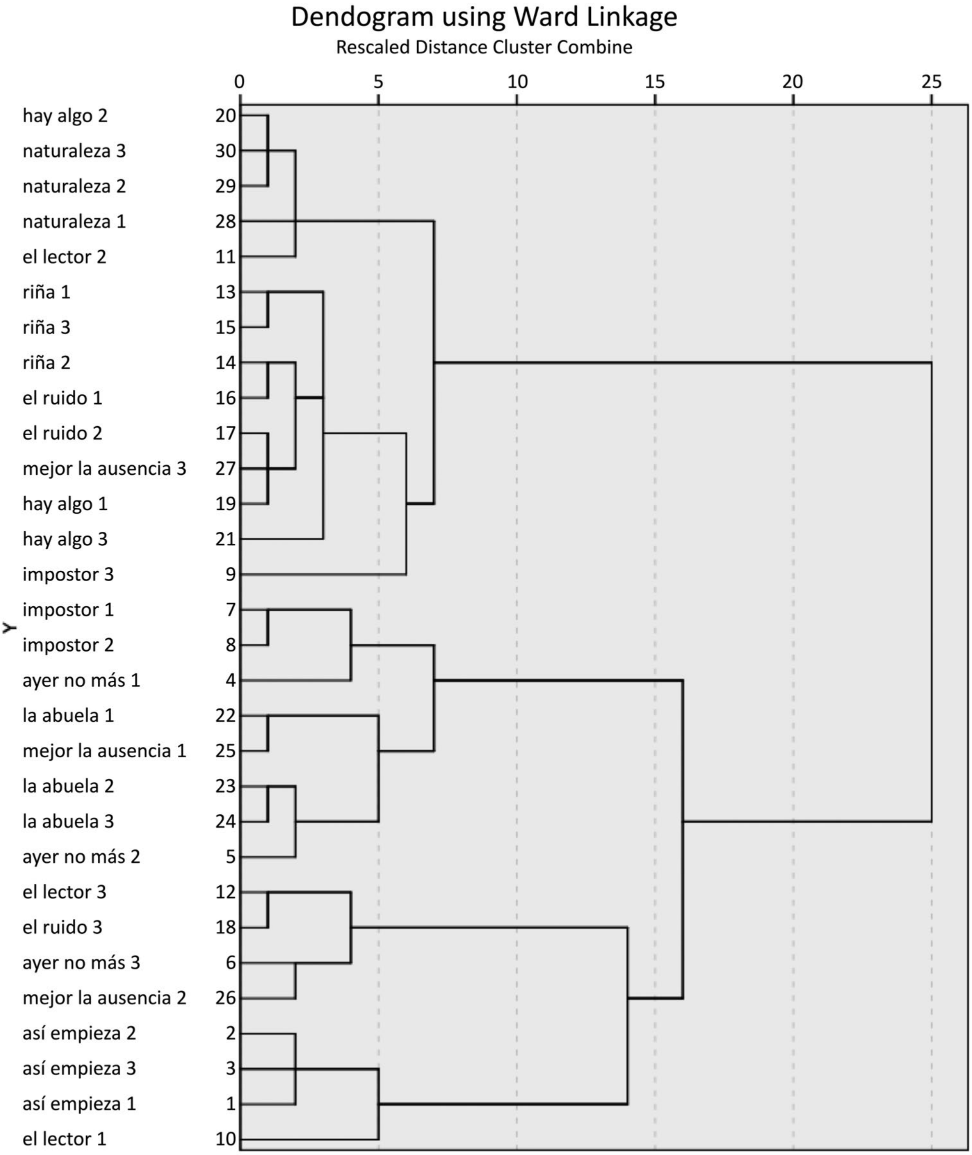


Fig. 3 Dendrogram (Dziuk Lameira, CC BY).

Table 5 Clusters 1, 7 and 9

	Cluster 1 (high MD and high FPL)			Cluster 7 (low MD and low FPL)			Cluster 9 (high LC and low FPL)		
	N	Mean	Standard deviation	N	Mean	Standard deviation	N	Mean	Standard deviation
MD	3.00	8.49	1.18	5.00	1.71	0.43	8.00	3.83	0.63
LO	3.00	39.23	3.39	5.00	32.06	3.64	8.00	31.40	1.96
LC	3.00	9.00	0.49	5.00	9.48	0.30	8.00	10.68	0.35
DeP	3.00	43.33	1.20	5.00	42.20	0.58	8.00	39.25	0.62
DeL	3.00	48.33	0.67	5.00	50.40	0.40	8.00	50.88	0.58
DiL	3.00	64.00	0.58	5.00	66.60	0.87	8.00	64.50	0.71
DFN	3.00	6.93	0.90	5.00	4.06	0.62	8.00	3.21	0.15
FPL	3.00	269.53	32.97	5.00	122.54	5.30	8.00	127.88	10.12

Table 4), the excerpts from the novels *Así empieza lo malo* (Cluster 1), *Naturaleza casi muerta* (Cluster 7), and *Riña de gatos* (Cluster 9) are grouped together. The excerpts from *Así empieza lo malo* were the only ones grouped on their own (see Figure 3).

In the following, the clusters 1, 7 and 9 (see table 5) will be described further in order to identify parameters or combinations of parameters which are potentially typical of the authors whose text excerpts were grouped together in the same cluster.

Cluster 1, which includes only the three text excerpts from the novel *Así empieza lo malo*, as well as two excerpts from other novels, is characterized by a high metaphorical density (MD) and a high logarithmic average frequency (FPL) compared to the other clusters. It should be further investigated if the combination of a high metaphorical density and a relatively high logarithmic average word frequency is characteristic for the novel *Así empieza lo malo*, or Javier María's authorial style in general.

In contrast to cluster 1, cluster 7, which includes all three text excerpts from the novel *Naturaleza casi muerta* by Carme Riera as well as two other novel excerpts, has low values for the variables MD (metaphorical density) and FPL (logarithmic average frequency). The combination of a low metaphorical density and a low logarithmic average frequency could therefore be potentially typical of Carme Rieras novel.

Cluster 9 includes all three excerpts from the novel *Riña de gatos. Madrid 1936* by Eduardo Mendoza as well as five other excerpts. It is characterized by a high clause length (LC) and a low FPL (logarithmic average frequency), which could be typical for the style of the novel *Riña de gatos. Madrid 1936*.

The analysis of the clusters showed that the variables DeP (propositional density), DeL (lexical density), DiL (lexical diversity), DFN (density of the noun phrase) showed little deviation throughout the text excerpts. When the cluster analysis was performed without the variables DeP (propositional density), DeL (lexical density), DiL (lexical diversity), DFN (density of the noun phrase) only using the variables MD (metaphorical density), LO (sentence length), LC (clause length), and FPL (logarithmic average frequency), the same text excerpts were grouped together. Thus, the analysis shows that the parameters MD, LO, LC and FPL are more suitable for the discrimination of different styles than the other parameters.

4.2 Discussion

This analysis has shown that some parameters that are used to measure text complexity are significantly different between some of the novels. Therefore, those parameters can potentially be used for stylistic analysis. This is the case for the variables MD (metaphorical density), LO (sentence length), LC (clause length), IS (index of subordination), DFN (density of the noun phrase), and FPL (logarithmic average frequency). Also, the combination of some parameters could be typical of the style of some authors (e.g. a high metaphorical density and a high logarithmic average word frequency for Javier Marías). This should be further investigated using more text excerpts from the same novels as well as from other novels by the same authors.

The relatively short length of the text excerpts of approximately 300 words could explain cases of high standard deviation for the variable LO ('sentence length'). Given that a large amount of sentences included in the text excerpts exceed 100 words, whereas a similar number of sentences consist of less than five words, the average sentence length can vary greatly depending on the selected sample. Therefore, the variable LO is not suitable for the analysis and comparison of relatively short text excerpts.

The cluster analysis did not group the text excerpts consistently according to their novel. Future investigation should include more text excerpts per novel to perform this kind of analysis.

Additionally, the spread of the values for the different variables varies depending on the novel. More samples should be analyzed in order to determine if certain variables are more consistent within particular novels or novels by the same author.

References

Barrio-Cantalejo, Inés María, Pablo Simón-Lorda, M.C. Puerta Melguizo, Isabel Escalona, María Isabel Marijuán, and Pablo Hernando. 2008. "Validation of the INFLESZ scale to evaluate readability of texts aimed at the patient." *Anales del sistema sanitario de Navarra* 31 (2): 135–52.

Berber Sardinha, Tony. 2010. "A Program for Finding Metaphor Candidates in Corpora." *ESPecialist* 31 (1): 49–67.

Council of Chief State School Officers (CCSSO). 2010. *Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects. Appendix A*, Washington, DC. https://achievethecore.org/content/upload/corestandards_appendix_a_text_complexity_ela.pdf (Accessed July 16, 2023)

Dahl, Östen. 2004. *The Growth and Maintenance of Linguistic Complexity.* Studies in Language Companion Series. Amsterdam: John Benjamins.

Dziuk Lameira, Katharina. 2019. "Complejidad Semántica: El Ejemplo de la Metáfora." In *Competencia textual y complejidad textual. Perspectivas transversales entre didáctica y lingüística*, edited by Angela Schrott and Bernd Tesch, 125–45. Berlin: Peter Lang.

Dziuk Lameira, Katharina. 2023. *Textkomplexität und Textverständlichkeit: Studien zur Komplexität spanischer Prosatexte.* Berlin, Boston: De Gruyter.

Entin, Eileen B., and George R. Klare. 1978. "Factor Analyses of Three Correlation Matrices of Readability Variables." *Journal of Reading Behavior* 10 (3): 279–90.

Fernández Huerta, José. 1959. "Medidas sencillas de lecturabilidad." *Consigna* 214: 29–32.

Fix, Ulla. 2009. "Muster und Abweichung in Rhetorik und Stilistik." In *Rhetorik und Stilistik. Ein internationales Handbuch historischer und systematischer Forschung*, vol. 2, edited by Ulla Fix, Andreas Gardt, and Joachim Knape, 1300–15. Berlin, Boston: De Gruyter.

François, Thomas, and Cédrick Fairon. 2012. "An 'AI Readability' Formula for French as a Foreign Language." In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 466–77. Stroudsburg, PA: Association for Computational Linguistics.

Friedrich, Marcus. 2017. *Textverständlichkeit und ihre Messung: Entwicklung und Erprobung eines Fragebogens zur Textverständlichkeit.* Münster, New York: Waxmann.

Gardt, Andreas. 2012. "Textsemantik. Methoden der Bedeutungserschließung." In *Geschichte der Sprache und Sprache der Geschichte. Probleme und Perspektiven der historischen Sprachwissenschaft des Deutschen. Oskar Reichmann zum 75. Geburtstag*, edited by Jochen A. Bär and Marcus Müller, 61–82. Berlin: Akademie-Verlag.

Glück, Helmut, and Michael Rödel (edd.). 2016. *Metzler Lexikon Sprache*. Stuttgart: Metzler.

Goatly, Andrew. 1997. *The Language of Metaphors*. London, New York: Routledge.

Gutiérrez de Polini, Luisa Elena. 1972. "Investigaciones sobre lectura en Venezuela. Informe presentado a las *Primeras Jornadas de Educación Primaria*." Caracas: Ministerio de Educación.

Haspelmath, Martin. 2006. "Against Markedness (and What to Replace It With)." *Journal of Linguistics* 42 (1): 25–70.

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

Karlsson, Fred, Matti Miestamo, and Kaius Sinnemäki. 2008. "Introduction: The Problem of

Language Complexity." In *Language Complexity: Typology, Contact, Change*, edited by Fred Karlsson, Matti Miestamo, and Kaius Sinnemäki, vii–xiv. Amsterdam: John Benjamins.

Klare, George R. 1974. "Assessing Readability." *Reading Research Quarterly* 10 (1): 62–102.

Klare, George R. 1984. "Readability." In *Handbook of Reading Research,* vol. 1, edited by P. David Pearson. 681–744. London: Routledge.

Kusters, Wouter. 2003. *Linguistic Complexity*. Utrecht: LOT.

Merlini Barbaresi, Lavinia. 2011. "A 'Natural' Approach to Text Complexity." *Poznań Studies in Contemporary Linguistics* 47 (2): 203–36.

Miestamo, Matti. 2006. "On the feasibility of complexity metrics." In *FinEst Linguistics, Proceedings of the Annual Finnish and Estonian Conference of Linguistics*, edited by Krista Kerge and Maria-Maren Sepper, 11–26. Tallinn: Tallinn University Press.

Miestamo, Matti. 2008. "Grammatical Complexity in a Cross-Linguistic Perspective." In *Language Complexity: Typology, Contact, Change*, edited by Fred Karlsson, Matti Miestamo, and Kaius Sinnemäki, 23–41. Amsterdam: John Benjamins.

Mikk, Jaan. 2000. *Textbook: Research and Writing*. Frankfurt am Main, Berlin: Peter Lang.

Mikk, Jaan. 2005. "Text Comprehensibility." In *Quantitative Linguistics: An International Handbook*, edited by Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski, 909–21. Berlin, Boston: De Gruyter.

Mikk, Jaan, and Jaanus Elts. 1999. "A Reading Comprehension Formula of Reader and Text Characteristics." *Journal of Quantitative Linguistics* 6 (3): 214–21.

Muñoz Baquedano, Miguel, and José Muñoz Urra. 2019. *Legibilidad Mu*. Viña del Mar, Chile. http://www.legibilidadmu.cl (Accessed April 15, 2019).

Pragglejaz Group. 2007. "MIP: A Method for Identifying Metaphorically Used Words in Discourse." *Metaphor and Symbol* 22 (1): 1–39.

Rai, Sunny, and Shampa Chakraverty. 2017. "Metaphor Detection Using Fuzzy Rough Sets." In *International Joint Conference on Rough Sets*, 271–79. Cham: Springer International Publishing.

Rai, Sunny, Shampa Chakraverty, and Devendra K. Tayal. 2016. "Supervised Metaphor Detection Using Conditional Random Fields." In *Proceedings of the Fourth Workshop on Metaphor in NLP*, 18–27. San Diego, CA: Association for Computational Linguistics.

Rescher, Nicholas. 1998. *Complexity: A Philosophical Overview.* New Brunswick, NJ: Transaction Publishers.

Santa Fe Institute. 2018. *Complexity*. https://www.complexityexplorer.org/explore/glossary/11-complexit (Accessed July 16, 2023)

Skirl, Helge. 2009. *Emergenz als Phänomen der Semantik am Beispiel des Metaphernverstehens. Emergente konzeptuelle Merkmale an der Schnittstelle von Semantik und Pragmatik.* Tübingen: Narr.

Stede, Manfred. 2018. *Korpusgestützte Textanalyse: Grundzüge der Ebenen-orientierten Textlinguistik*. Tübingen: Narr Francke Attempto.

Steen, Gerard J., Aletta G. Dorst, Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Tryntje Pasma. 2010. *A Method for Linguistic Metaphor Identification.* Amsterdam: John Benjamins.

Szigriszt Pazos, Francisco 1993. *Sistemas predictivos de legibilidad del mensaje escrito: Fórmula de Perspicuidad* (Doctoral dissertation) Madrid: Universidad Complutense de Madrid.

Véliz, Mónica, and Bruno Karelovic. "Trunajod." Universidad de Concepción, Chile. http://www.udec.cl/~trunajod/.

Applying General Impostors Method to the Ferrante Case

Michele A. Cortelazzo, George K. Mikros,
and Arjuna Tuzzi

Abstract Elena Ferrante is the *nome de plume* of an anonymous writer who is highly successful on the international stage and whose success far exceeds that of other authors of contemporary Italian literature. In this study, we approach Ferrante's authorship investigation as a verification problem since we cannot be sure whether the real author behind Ferrante's pseudonym is among the candidates we have considered in previous studies. For this reason, we applied the General Impostors (GI) method using the Cosine Delta distance in both a corpus of 150 novels written by 40 authors (39 candidates and Elena Ferrante) and a non-literary corpus of 113 texts signed by 14 different entities (12 authors, a collective author, and Elena Ferrante). In the literary corpus, Starnone emerged as the most likely author of Ferrante's novels. Results were quite different in the second case: Starnone was not the only possible author since, in many non-literary texts, Raja, Martone as well as the E/O publishing house staff and publishers, seem to have authorial contributions. The GI method not only confirmed previous results but also improved our knowledge of this case since it provides a measure of the attribution strength.

Keywords Ferrante, authorship verification, stylometry, General Impostors method

1. Introduction

Elena Ferrante is the pen name of an Italian writer whose novels became a global phenomenon. Today she is perhaps the best-known Italian author on the international stage, and this result seems both deserved and peculiar given that Elena Ferrante is a secretive author. Elena Ferrante's identity has been kept secret for the last 30 years with

Cortelazzo, Michele A., Mikros, George K. and Tuzzi, Arjuna: Applying General Impostors Method to the Ferrante Case, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 299–313. DOI: https://doi.org/10.17885/heiup.1157.c19377

the strong support of her publishers, Sandro Ferri and Sandra Ozzola, the owners of the E/O publishing house.

In 2018, Elena Ferrante was the author of seven novels: *L'amore molesto*, *I giorni dell'abbandono*, *La figlia oscura*, *L'amica geniale. Infanzia, adolescenza*, *Storia del nuovo cognome. L'amica geniale volume secondo*, *Storia di chi fugge e di chi resta. L'amica geniale volume terzo* and *Storia della bambina perduta. L'amica geniale volume quarto*. The last four books represent episodes of the best-selling series of novels *L'amica geniale* (*My Brilliant Friend*).

Elena Ferrante has also written a children's story (*La spiaggia di notte*), and she is also the main contributor to a collection of non-literary texts (interviews, essays, and letters) published in a book entitled *La Frantumaglia* (Ferrante 2016). In 2019 a new collection of non-literary texts, *L'invenzione occasionale* (*Incidental Inventions*), appeared: it includes all columns that she published during her year-long collaboration with *The Guardian* in 2018 (Ferrante 2019). Moreover, a new novel was published by the E/O publishing house in November 2019 with the title *La vita bugiarda degli adulti*. This new novel is not the fifth episode of the famous *L'amica geniale* saga, though it is a further story set in Naples. In 2021 Elena Ferrante described in the essay "I margini e il dettato" the pleasure of reading and writing (Ferrante 2021).

Beyond the obvious, intriguing issue of her real identity, Elena Ferrante represents a relevant research object from both the stylistic and the stylometric standpoints. Ferrante's authorship problem is a complex research task as she is an active author not only in literature but also in non-fiction prose. Moreover, since it is a pseudonym, we cannot exclude the existence of processes of collective writing and/or ghostwriters behind the pen name. To better understand her linguistic production and model her writing style, we need to examine not only her literary works but also her articles, essays, and books that are primarily journalistic or autobiographical and represent a completely different genre.

Since this study aims at comparing results with the ones achieved in previous studies, we consider two corpora that have already been exploited:

1) a large literary corpus (Tuzzi and Cortelazzo 2018a, b, c);
2) a corpus of non-literary texts (Cortelazzo, Mikros, and Tuzzi 2018).

2. Corpora Used in This Study

To examine Ferrante's fiction style, we utilized a corpus of contemporary Italian literature that contains 150 novels from 40 different authors,[1] most of them written between 1987 and 2016 and totaling 9,837,851 tokens[2] and 159,149 types. The corpus consists of texts of variable length (Mdn = 50,840.5 words, Mean = 65,586, St.Dev = 39,120, Min = 8,129, Max = 199,839 words) and Ferrante is represented by all her seven novels (635,819 tokens, 33,158 types) which are also variable in size (Mdn = 97,893 words, Mean = 90,831, Min = 36,784, Max = 142,215). The corpus is composed of 50 books from 13 female authors (including Ferrante) and 11 authors from the Campania region (including Ferrante) with 46 books. It contains not only the authors suspected to be behind Ferrante's name but also a wider range of authors that offer a more varied picture of literary production. In that sense, the specific corpus can be used to explore Ferrante's position in the larger framework of contemporary Italian literature and to model author profiles with a more generic coverage.

The selected novels and novelists are all ascribed to one (or more) of these categories:

— Elena Ferrante's novels.
— Novels written by authors from the same area (Naples and its surroundings).
— Novels written by novelists suspected to be Elena Ferrante.
— Blockbusters (best sellers, award-winning novels).
— Novels written by authors who enjoyed the praise of literary criticism.

Furthermore, we compiled a second corpus of Ferrante's non-fiction texts along with a comparable non-fiction corpus with some of the candidate authors behind Ferrante's pseudonym (Cortelazzo, Mikros, and Tuzzi 2018). This non-fiction corpus is composed of 113 texts (143,695 word tokens[3] and 19,020 word types, Mdn = 779 words, Mean = 1,272, St.Dev = 1,406, Min = 228, Max = 8,987). It includes letters, interviews, and additional material written by different authors that can be compared with a selection of texts of *La Frantumaglia* by Elena Ferrante (last Italian version 2016) (Ferrante 2016).

[1] The authors contained in this corpus are: Affinati, Ammaniti, Bajani, Balzano, Baricco, Benni, Brizzi, Carofiglio, Covacich, De Luca, De Silva, Faletti, Ferrante, Fois, Giordano, Lagioia, Maraini, Mazzantini, Mazzucco, Milone, Montesano, Morazzoni, Murgia, Nesi, Nori, Parrella, Piccolo, Pincio, Prisco, Raimo, Ramondino, Rea, Scarpa, Sereni, Starnone, Tamaro, Valerio, Vasta, Veronesi, Vinci.

[2] Calculations performed with Taltac software (Bolasco 2010) ver. 2.10.

[3] Calculations performed with Taltac software (Bolasco 2010) ver. 2.10.

The subcorpus of non-Ferrante texts contains 86 texts (87,458 tokens, 14,308 types, Mdn = 723.5 words, Mean = 1,017, St.Dev = 965, Min = 228, Max = 4,777), and it is composed mainly of articles in newspapers and magazines, essays published in various media, interviews, letters, and texts posted on the Web. 78 of these texts were written by 12 authors (Laura Buffoni, Gianrico Carofiglio, Sandro Ferri, Goffredo Fofi, Marcella Marmo, Mario Martone, Sandra Ozzola, Valeria Parrella, Francesco Piccolo, Anita Raja, Clara Sereni, Domenico Starnone) and eight by a collective subject (E/O) that represents the editorial staff of E/O publishing house (Sandro Ferri and Sandra Ozzola are the owners of E/O).

The subcorpus of Ferrante's non-fiction works includes 27 texts signed by Elena Ferrante and it is distributed across six essays, seven interviews and 14 letters (56,237 tokens, 21,293 types, Mdn = 1,001 words, Mean = 2,083, St. Dev = 2,138, Min = 298 words, Max = 8,987 words).

3. Methodology

3.1 Ferrante and the Need for Applying Authorship Verification Methods

The authorship identification in Elena Ferrante's case is a complex task since it involves several unknown parameters regarding the exact nature of the problem. It can be viewed as an authorship attribution case, i.e., a closed-class classification problem where a standard text classification task can be used. Machine learning algorithms can be trained on a corpus where texts belong to known authors, and the model developed can be further verified using a hold-out set. Then it can predict the authorship in the collection of texts of unknown authorship. When we apply this pipeline to Ferrante's fiction corpus (Mikros 2018), we get as the most probable author behind Elena Ferrante, Domenico Starnone, with an accuracy of over 96 percent. However, this approach is based on a very unstable assumption, namely, that the rest of the 39 authors (including Starnone) who are represented in the fiction corpus are indeed a set of authors which includes, beyond any doubt, the real author of the Ferrante novels. However, we do not have any external evidence that this is the case. A high accuracy attribution further confirmed by different stylometric methods applied to the same data (Tuzzi and Cortelazzo 2018a; Savoy 2018b) is a sign of increased consensus and high reliability. However, we still cannot exclude the possibility that the real author behind Ferrante is someone outside our initial large corpus of contemporary Italian literature.

This suspicion can be further supported by the puzzling author profiling results we obtained when we examined the non-fiction corpus and tried to evaluate Ferrante's gender and age. Author profiling methods were used with considerable success in the fiction corpus defining Ferrante as a male author over 60 and coming from the Naples area (the sole candidate with these characteristics was Domenico Starnone) (Mikros 2018). However, when the same methods were applied to the non-fiction corpus, the results were inconclusive (Cortelazzo, Mikros, and Tuzzi 2018). The gender, age, and region profiling results can be found in Figure 1.

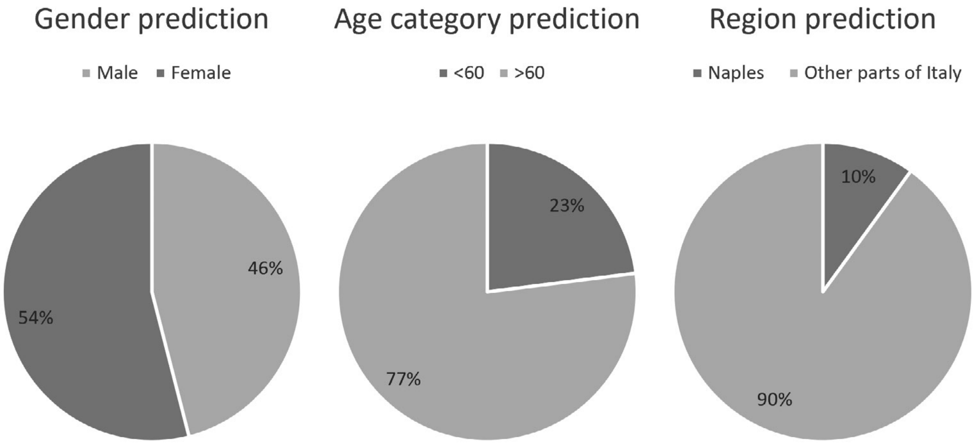


Fig. 1 Pie charts visualizing profiling results (gender, age, region) in Ferrante's non-fiction texts. The reported percentages correspond to the portion of the profiling characteristic predicted in Ferrante's texts. E.g., from the 27 texts signed by Ferrante, the algorithm predicted that 12 (46 percent) were written by a male author. (Cortelazzo, Mikros, Tuzzi, CC BY).

As shown in Figure 1, nearly half of the non-fiction texts are attributed to a female (46 percent) and the other half to a male (54 percent). A somewhat less but still intense variation can be seen in the age profiling, where 77 percent of the texts are attributed to a person over 60 years old and 23 percent to someone less than 60 years old. A more stable attribution appears with the region profiling since 90 percent of the texts are classified as belonging to someone from Naples. A reasonable hypothesis that emerges from these results is that the non-fiction Ferrante texts represent a collective work of more than one author employing authors of both genders and some variation in age.

The above research outcomes reinforce the need for employing explicitly designed methodologies for handling open authorship problems and do not require the existence of real authors inside the training corpus. This specific set of methods is designated by the general term of authorship verification. They can be classified into two broad categories (Potha and Stamatatos 2017):

- Intrinsic methods: perform analysis only on the documents under investigation and handle the verification problem as a one-class classification task. These methods are robust since they do not require external resources and fast since they analyze only a few documents. Examples of these approaches can be found in Jankowska, Milios, and Kešelj (2014), Halvani, Winter, and Pflug (2016), and Mikros and Perifanos (2011).
- Extrinsic methods: these methods analyze an additional set of external documents and transform the verification problem into a binary classification task. They are usually more effective, especially when the set of external documents has been carefully compiled. Characteristic examples of this approach are Koppel and Winter (2014), Seidman (2013), and Kestemont et al. (2016a).

The verification problem is considered the most challenging among the authorship identification tasks and, over the recent years, has attracted considerable research attention, including the organization of two PAN competitions in 2014 (Stamatatos et al. 2014, Stamatatos et al. 2015).

3.2 General Impostors Method

Among the various approaches proposed for solving the verification problem, we selected the GI method, and we used the version implemented in the *stylo* R package (Eder, Rybicki, and Kestemont 2016). The GI method is based on earlier work of Koppel's Many Candidates method (Koppel, Schler, and Argamon 2011), which was further enhanced and tested using different variations by Koppel and Winter (2014) and Seidman (2013) and optimized by Kestemont et al. (2016b). As Kestemont et al. (2016b, 88) explain:

> The general intuition behind the GI is not to assess whether two documents are simply similar in writing style, given a static feature vocabulary, but rather, it aims to assess whether two documents are significantly more similar to one another than other documents, across a variety of stochastically impaired feature spaces (Stamatatos 2006; Eder 2012) and compared to random selections of so-called distractor authors (Juola 2015), also called "imposters."

The procedure is based on a bootstrapped approach in which repeated samples of stylometric features (usually words or n-grams) are used in distance-based comparisons between an anonymous text and a random selection of *impostor* documents the original author did not write. The score calculated represents not only how different the anonymous text is from the other texts of the candidates but also how consistent the stylistic differences are between them.

The process is based on a second-order metric termed O2 in Kestemont et al. (2016b) in the sense that the distance metric is further processed and transformed into a proportion metric which is the final metric used in the algorithm. More specifically, the algorithm finds the distance between the vector of an anonymous document to the centroid vector of a list of the documents of candidate authors. It also finds the distance between the vector of the anonymous text and a list of random non-relevant to the authorship problem texts. Then the GI algorithm starts a bootstrapping procedure in which it samples a random subset of the linguistic features used and a random subset of impostors. In each iteration, it determines whether the vector of the anonymous document is closer to the vector of the candidate author's texts or the vector of the distractors' texts. GI then calculates the proportion of times the vector of the anonymous document was found closer to the vector of the candidate author compared to the vector of the impostors' documents. The proportion is normalized in the 0–1 scale, and since it is based on the distance metric, which should be first calculated, is considered a second-order metric.

The GI method is a versatile technique as the researcher can use a variety of distances, including some well-established in authorship research like Delta (Burrows 2002), Cosine Delta (Evert et al. 2017), Min-Max (Kestemont et al. 2016a), etc. Moreover, it was the winning method in the PAN authorship identification contest in 2013 (Seidman 2013) and 2014 (Khonji and Iraqi 2014), performing considerably better than other authorship verification methods.

4. Results and Discussion

We applied the GI method to Ferrante's authorship problem and used both the fiction and the non-fiction corpus. Although there is a consensus among recent stylometric research (Tuzzi and Cortelazzo 2018a; Cortelazzo and Tuzzi 2020; Savoy 2018b) that the stylometric profile of Domenico Starnone is the closest to Ferrante's among other candidates in the fiction corpus, we used the GI method to evaluate this claim further. Moreover, authorship verification methods have not yet been applied to the Ferrante case, and this research would fill in the existing gap in the stylometric quest to reveal her authorship.

We used the GI method, utilizing the 5,000 most frequent words as stylometric features and applying the Cosine Delta (aka Würzburg) distance as this seems to be experimentally more robust across many languages and varied sizes of most frequent words sets (Evert et al. 2017). We decided not to use any impostors' texts and to focus on the texts of the 39 Italian authors of the fiction corpus, treating each one as a possible candidate for being the Ferrante author. In this way, we did not make any assumptions regarding Ferrante's authorship and let the algorithm examine each of the 39 authors as candidates with equal probabilities of being the real author. Although this procedure is time-consuming, the results obtained are far more valid since the proposed methodology can be considered an alternative cross-validation strategy that checks attribution scores across all possible author pairs with the questioned document.

Since GI scores fall into the normalized range of 0–1, there is also a need to determine the attribution threshold, i.e., the score which marks a positive or a negative attribution to the questioned document. E.g., in our dataset, the "Letter to Ozzola (no. 2)" written by Ferrante was assigned a score of 0.43 when tested against the collective authorship of the E/O publishing house. What sort of evidence is 0.43? Can it be translated to a specific attribution or not? Is it high or low? The only way for us to answer these questions is to thoroughly examine a given corpus to calculate the average proximity between any texts written by the same author and the average proximity between a text by a given author and any text written by someone else. This procedure will define a margin where a classifier is (on average) wrong. *stylo* has adopted a score-shifting algorithm (Kestemont et al., 2016b), based on the c@1 measure of the classifier's performance (Peñas and Rodrigo 2011).

Using the algorithm mentioned above, we calculated the optimal decision thresholds for rejecting or accepting the attribution using the Cosine Delta distance through the relative function offered by the Impostors method implemented in the *stylo* package. Given our corpus, the lower value was calculated as 0.49 and the upper value as 0.51. This means that GI scores under 0.49 can be considered *low* and cannot be used as evidence for attributing a tested text to a specific author. Moreover, GI scores over 0.51 are considered high enough to be translated as positive evidence for attributing the test document to the specific author.

We compared each of Ferrante's books (7 novels) with all the books by 39 authors (143 novels). In this task, Domenico Starnone was indicated as the author of these novels with a probability GI score of 1, which can be explained as a perfect match across all features' subsets. This result further confirms all the previously reported stylometric research stating that Domenico Starnone's stylometric profile is the closest to Elena Ferrante's writings.

After establishing the validity of the GI method in the fiction corpus, we tested Ferrante's authorship in the non-fiction corpus. Since the non-fiction corpus is smaller than the fiction one, we used the 2,000 most frequent words as stylometric features and applied the Cosine Delta distance. Moreover, we calculated the optimal decision thresholds for the decline and the acceptance of the attribution using the cosine distance. Using the relative optimization algorithm employed in the *stylo* package, the lower value was calculated to be 0.4 and the upper value to be 0.52, i.e., any GI score over 0.52 produced in a comparison of a known authorship text and the anonymous text could indicate that the author of the known authorship text is also the author of the unknown text.

We compared each of Ferrante's texts (27 letters, interviews, essays) with the rest of the non-fiction texts written by 12 authors and one collective writer (staff of the E/O publishing house). The non-literary Ferrante texts (included in *La Frantumaglia*) seem like they may have been written by multiple authors. Among them are Starnone, Raja, Martone, Parella, Ozzola, and the rest of the staff from the E/O publishing house. Other candidates (Buffoni, Carofiglio, Ferri, Fofi, Marmo, Piccolo) seem entirely irrelevant to the writing of these documents, and seven texts out of 27 do not have an exact author match.

These results confirm our previous study in the same corpus using profiling methods, and closed-class supervised machine learning algorithms (Cortelazzo, Mikros, and Tuzzi 2018). To reduce the algorithm's search space and explore in detail the most probable candidates, we repeated the whole procedure maintaining E/O, Martone, Parrella, Raja, and Starnone in the candidates' pool and recalculating the GI decision boundaries and the GI scores for each of Ferrante's non-fiction texts. We used the same feature set (2,000 most frequent words), and the GI threshold values were calculated as 0.38 (rejection threshold) and 0.45 (attribution threshold). Table 1 reports the GI scores for each text and each candidate.

The GI scores calculated confirm our initial attribution. Using the GI verification method, all of Ferrante's non-fiction texts have been assigned to one or more than one author (six out of 27 have been assigned to two authors and one of them to three). Both Starnone and Raja seem to have written a number of these texts. The extended staff of the E/O publishing house now has ten attributions, confirming our suspicion that part of *La Frantumaglia* is a collective production of the staff of the E/O publishing house. The distributed authorship hypothesis can be visualized in Figure 2, which displays the GI scores.

In Figure 2, we used a stacked bar chart to standardize the magnitude of GI scores which were over 0.45 and are considered above the threshold of positive authorship attribution. E.g., if two authors had GI scores above the attribution threshold in one text, these two scores were normalized in relative percentages. For example., in *lettero2,* there was a GI score of 0.52, attributing this text to the E/O publishing house team, and a GI score of 0.65 that attributes the text to Martone. The stacked bar converted these scores to 44 percent and 56 percent correspondingly so that the bar adds up to 100 percent, and we can compare the relative magnitude of each GI score across the

Table 1 GI scores calculated for the Ferrante non-fiction texts and the attributed candidate authors. The attribution scores can be seen in bold under the candidate's column.

Testing documents	E/O	Martone	Parrella	Raja	Starnone
essay01	0.06	0.42	0.01	**0.77**	0.18
essay02	0.06	**0.93**	0.12	0.37	0.26
essay03	0.06	**0.95**	0.11	0.52	0.08
essay04	0.2	0.25	0.01	**0.86**	0.05
essay05	0.01	**0.79**	0.13	**0.49**	0.23
essay06	0.29	0.05	**0.63**	0	**0.61**
interview01	**0.92**	0.04	0.02	0.01	0.14
interview02	**0.86**	0	0.02	0.3	0.09
interview03	0.43	0.39	0	**0.68**	0
interview04	0.26	0.06	**0.93**	0	0.29
interview05	0.12	0.38	0	**0.87**	0.11
interview06	0.34	0.07	0.02	**0.85**	0.07
interview07	**0.46**	0.07	0.19	**0.47**	**0.53**
letter01	0.18	0.05	0.07	0	**0.89**
letter02	**0.52**	**0.65**	0.22	0	0.11
letter03	**0.86**	0.17	0.01	0	**0.52**
letter04	0.43	0.19	0.18	0	**0.56**
letter05	**0.65**	0.15	0.19	0	0.33
letter06	**0.96**	0.16	0.07	0	0.11
letter07	**1**	0.13	0	0	0.1
letter08	0.2	0.12	0.22	**0.45**	0.24
letter09	0.15	0.08	**0.86**	0	**0.54**
letter10	**0.66**	0.28	0.32	0	0.34
letter11	**0.57**	0.17	0.23	0	0.43
letter12	0.32	0.3	0.24	0	**0.59**
letter13	0.4	0.27	0.06	0.13	**0.47**
letter14	0.26	0.28	0.06	0	**0.67**

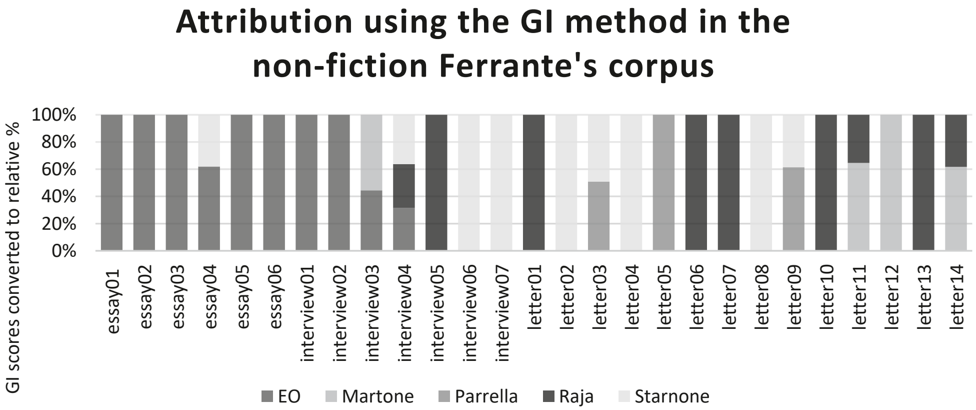


Fig. 2 Stacked bar chart of the GI scores obtained. The attributions of the Ferrante texts can be seen using the legend of the graph. (Cortelazzo, Mikros, Tuzzi 2023, CC BY).

various Ferrante's texts on a uniform scale way since all scores have now been transformed to the scale 0–100 percent.

Given the previous discussion and the results obtained by applying the GI method in Ferrante's non-fiction texts,[4] we can safely infer that Ferrante's non-fiction texts do not represent a homogeneous stylometric profile and could be attributed to various people working for the public relations of the Ferrante brand name.

5. Conclusion

Elena Ferrante's authorship remains a very interesting stylometric problem and one of the most complex cases of cross-genre attribution, as she is an active author in fiction and non-fiction texts. In this study, we tried to complement previous stylometric research and use an authorship verification technique called the GI method. The rationale behind this approach is that we need to approach Ferrante's authorship case as an open-class problem where the research question does not imply a set of predetermined candidates but leaves space for possibilities other than the ones we might have in mind.

[4] The script used for applying the GI method to the Ferrante's texts is available on GitHub: https://github.com/gmikros/GI-Method-in-Ferrante-texts.

In both corpora (fiction and non-fiction), the GI method confirmed previous research results, but it also improved our knowledge since it provided a measure of the attribution strength. Domenico Starnone's stylometric profile was identified as the single undisputed match compared to the stylometric profile of Ferrante's novels. These results confirm and enrich results obtained by previous studies collected in Tuzzi and Cortelazzo (2018a) (cf. Eder 2018; Juola 2018; Lalli, Tria, and Loreto 2018; Mikros 2018; Ratinaud 2018; Rybicki 2018; Savoy 2018a). However, this clear-cut picture did not emerge when we examined the authorship of the non-fiction corpus. The GI method attributed some of the texts collected in *La Frantumaglia* to the staff or the owners of the E/O publishing house. For some other texts, Domenico Starnone, Anita Raja, Mario Martone, and Valeria Parrella were identified as possible authors. Moreover, in specific Ferrante non-fiction texts, we can detect patterns of co-writing as we observe attribution GI scores to more than one author.

Authorship verification methods are less accurate than the supervised classification pipelines, but, in our case, they can be used to complement the published research on this topic. They can shed light on research questions that a closed-class classification algorithm cannot answer. Both the mixed authorship signal detected in some of Ferrante's non-fiction texts and the distributed authorship hypothesis as part of an organized public communication project supporting Ferrante's name can be reliably investigated under the authorship verification framework. Enlarging available corpora with new works signed by Elena Ferrante could be the starting point for further investigations and new research questions.

ORCID®

Georgios Mikros https://orcid.org/0000-0002-4093-5973
Arjuna Tuzzi https://orcid.org/0000-0003-3795-5567

References

Bolasco, Sergio. 2010. *TaLTaC2.10 Sviluppi, esperienze ed elementi essenziali di analisi automatic dei testi*. Milano: LED.

Burrows, John F. 2002. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17 (3): 267–87.

Cortelazzo, Michele A., George Mikros, K., and Arjuna Tuzzi. 2018. "Profiling Elena Ferrante: A Look Beyond Novels." In *JADT 2018: Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*, edited by Domenica Fioredistella Iezzi, Livia Celardo, and Michelangelo Misuraca, 165–73. Rome: UniversItalia.

Cortelazzo, Michele A., and Arjuna Tuzzi. 2020. "A chi assomiglia Elena Ferrante? Un profilo stilometrico aggiornato." *Italica Wratislaviensia* 11 (1): 123–41.

Eder, Maciej. 2012. "Computational Stylistics and Biblical Translation: How Reliable Can a Dendrogram Be?" In *The Translator and the Computer*, edited by Tadeusz Piotrowski and Łukasz Grabowski, 155–70. Wrocław: WSF Press.

Eder, Maciej. 2018. "Elena Ferrante: A Virtual Author." In *Drawing Elena Ferrante's Profile. Workshop Proceedings, Padova, 7 September 2017*, edited by Arjuna Tuzzi and Michele A. Cortelazzo, 31–45. Padova: Padova University Press.

Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. "Stylometry with R: A Package for Computational Text Analysis." *R Journal* 8 (1): 107–21.

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. "Understanding and Explaining Delta Measures for Authorship Attribution." *Digital Scholarship in the Humanities* 32 (2): ii4–ii16. https://doi.org/10.1093/llc/fqx023.

Ferrante, Elena. 2016. *La Frantumaglia*. Rome: Edizioni E/O.

Ferrante, Elena. 2019. *L'invenzione occasionale*. Rome: Edizioni E/O.

Ferrante, Elena. 2021. *I margini e il dettato*. Rome: Edizioni E/O.

Halvani, Oren, Christian Winter, and Anika Pflug. 2016. "Authorship Verification for Different Languages, Genres and Topics." *Digital Investigation* 16: S33–S43. https://doi.org/10.1016/j.diin.2016.01.006.

Jankowska, Magdalena, Evangelos Milios, and Vlado Kešelj. 2014. "Author Verification Using Common N-Gram Profiles of Text Documents." In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 387–97. Dublin, Ireland: Dublin City University and Association for Computational Linguistics.

Juola, Patrick. 2015. "The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions." *Digital Scholarship in the Humanities* 30 (1): i100–i113. https://doi.org/10.1093/llc/fqv040.

Juola, Patrick. 2018. "Thesaurus-Based Semantic Similarity Judgements: A New Approach to Authorial Similarity?" In *Drawing Elena Ferrante's Profile. Workshop Proceedings, Padova, 7 September 2017*, edited by Arjuna Tuzzi, and Michele A. Cortelazzo, 48–59. Padova: Padova University Press.

Kestemont, Mike, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. 2016a. "Authorship Verification with the Ruzicka Metric." In *Digital Humanities 2016: Conference Abstracts*, 246–49. Krakow: Jagiellonian University & Pedagogical University.

Kestemont, Mike, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. 2016b. "Authenticating the Writings of Julius Caesar." *Expert Systems with Applications* 63: 86–96. https://doi.org/10.1016/j.eswa.2016.06.029.

Khonji, Mahmoud, and Youssef Iraqi. 2014. "A Slightly-Modified GI-Based Author-Verifier with Lots of Features (ASGALF) − Notebook for PAN at CLEF 2014." In *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15–18 September, Sheffield, UK, September 2014*, edited by Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij, 977–83. CEUR Workshop Proceedings (CEUR-WS.org).

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. 2011. "Authorship Attribution in the Wild." *Language Resources and Evaluation* 45 (1): 83–94. https://doi.org/10.1007/s10579-009-9111-2.

Koppel, Moshe, and Yaron Winter. 2014. "Determining If Two Documents Are Written by the Same Author." *Journal of the Association for Information Science and Technology* 65 (1): 178–87. https://doi.org/10.1002/asi.22954.

Lalli, Margherita, Francesca Tria, and Vittorio Loreto. 2018. "Data-Compression Approach to Authorship Attribution." In *Drawing Elena Ferrante's Profile. Workshop Proceedings, Padova, 7 September 2017*, edited by Arjuna Tuzzi, and Michele A. Cortelazzo, 61–83. Padova: Padova University Press.

Mikros, George, K. 2018. "Blended Authorship Attribution: Unmasking Elena Ferrante Combining Different Author Profiling Methods." In *Drawing Elena Ferrante's Profile. Workshop Proceedings, Padova, 7 September 2017*, edited by Arjuna Tuzzi, and Michele A. Cortelazzo, 85–95. Padova: Padova University Press.

Mikros, George K., and Kostas Perifanos. 2011. "Authorship Identification in Large Email Collections: Experiments Using Features That Belong to Different Linguistic Levels." *Proceedings of PAN 2011 Lab, Uncovering Plagiarism, Authorship, and Social Software Misuse held in conjunction with the CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation, 19–22 September 2011, Amsterdam.*

Peñas, Anselmo, and Alvaro Rodrigo. 2011. "A Simple Measure to Assess Non-response." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 1415–24. Portland, Oregon: ACL.

Potha, Nektaria, and Efstathios Stamatatos. 2017. "An Improved Impostors Method for Authorship Verification." In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017*, edited by Gareth J.F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, and Nicola Ferro, 138–44. Cham: Springer.

Ratinaud, Pierre. 2018. "The Brilliant Friend(s) of Elena Ferrante: A Lexicometrical Comparison between Elena Ferrante's Books and 39 Contemporary Italian Writers." In *Drawing Elena Ferrante's Profile. Workshop Proceedings, Padova, 7 September 2017*, edited by Arjuna Tuzzi and Michele A. Cortelazzo, 97–110. Padova: Padova University Press.

Rybicki, Jan. 2018. "Partners in Life, Partners in Crime?" In *Drawing Elena Ferrante's Profile. Workshop Proceedings, Padova, 7 September 2017*, edited by Arjuna Tuzzi, and Michele A. Cortelazzo, 111–22. Padova: Padova University Press.

Savoy, Jacques. 2018a. "Elena Ferrante Unmasked." In *Drawing Elena Ferrante's Profile. Workshop Proceedings, Padova, 7 September 2017*, edited by Arjuna Tuzzi, and Michele A. Cortelazzo, 123–41. Padova: Padova University Press.

Savoy, Jacques. 2018b. "Is Starnone Really the Author behind Ferrante?" *Digital Scholarship in the Humanities* 33 (4): 902–18. https://doi.org/10.1093/llc/fqy016.

Seidman, Shachar. 2013. "Authorship Verification Using the Impostors Method." In *Working Notes for CLEF 2013 Conference, Valencia, Spain, September 23–26, 2013*, edited by Pamela Forner, Roberto Navigli, Dan Tufis, and Nicola Ferro. Valencia: CEUR.

Stamatatos, Efstathios. 2006. "Authorship Attribution Based on Feature Set Subspacing Ensembles." *International Journal on Artificial Intelligence Tools* 15 (5): 823–38. https://doi.org/10.1142/S0218213006002965.

Stamatatos, Efstathios, Walter Daelemans, Ben Verhoeven, Benno Stein, Martin Potthast, Patrick Juola, Miguel A. Sánchez-Pérez, and Alberto Barrón-Cedeño. 2014. "Overview of the Author Identification Task at PAN 2014." In *Working Notes for CLEF 2014 Conference*, edited by Linda Cappellato, Nicola Ferro, Martin Halvey, Wessel Kraaij, Nicola Ferro, Martin Halvey, and Wessel Kraaij, 877–97. CEUR Workshop Proceedings (CEUR-WS.org).

Stamatatos, Efstathios, Walter Daelemans, Ben Verhoeven, Patrick Juola, Aurelio López-López, Martin Potthast, and Benno Stein. 2015. "Overview of the Author Identification Task at PAN 2015." In *Working Notes of CLEF 2015 – Conference and Labs of the Evaluation Forum*, edited by Linda Cappellato, Nicola Ferro, Gareth J. F. Jones, and Eric San Juan. CEUR Workshop Proceedings (CEUR-WS.org).

Tuzzi, Arjuna, and Michele A. Cortelazzo, eds. 2018a. *Drawing Elena Ferrante's Profile. Workshop Proceedings, Padova, 7 September 2017*. Padova: Padova University Press.

Tuzzi, Arjuna, and Michele A. Cortelazzo. 2018b. "It Takes Many Hands to Draw Elena Ferrante's Profile." In *Drawing Elena Ferrante's Profile. Workshop Proceedings, Padova, 7 September 2017*, edited by Arjuna Tuzzi and Michele A. Cortelazzo, 9–29. Padova: Padova University Press.

Tuzzi, Arjuna, and Michele A. Cortelazzo. 2018c. "What Is Elena Ferrante? A Comparative Analysis of a Secretive Bestselling Italian Writer." *Digital Scholarship in the Humanities* 33 (3): 685–702. https://doi.org/10.1093/llc/fqx066.

About the Authors

Douglas Biber is Regents' professor emeritus (Applied Linguistics) at Northern Arizona University. His research efforts have focused on corpus linguistics, English grammar, and register variation. Previous books include *Register, Genre, and Style* (Cambridge, 2009/2019), the co-authored *Grammar of Spoken and Written English* (John Benjamins 2021), treatments of methodological issues in corpus linguistics (Cambridge 1998, 2015, 2020, 2022), and studies of grammatical complexity (Cambridge 2016, Routledge 2022) and register variation (Cambridge 1988, 1995, 2018; Benjamins 2006).

José Calvo Tello works as a researcher and specialist librarian at the Göttingen State and University Library. He obtained his doctorate in Humanities at the University of Würzburg, Germany, with the title *The Novel in the Spanish Silver Age: A Digital Analysis of the Genre Using Machine Learning* (transcript 2021). His research focuses on the application and development of statistical and computational methods applied to Romance literature and library data.

Michele A. Cortelazzo is a professor emeritus, former full professor in Italian linguistics, at the University of Padua and "Accademico ordinario" at the Accademia della Crusca. His research focuses on modern Italian and Italian for special purposes (medical, scientific and – in particular – political and institutional and administrative Italian). In the light of the results of his research, he has promoted clear and effective writing policies in the public administration. Over the last years, he has expanded his interests to the application of quantitative methods for the identification of similarities between texts.

Andreas van Cranenburgh is an assistant professor in Digital Humanities and information science at the Computational Linguistics Department of the Faculty of Arts at the University of Groningen. His work focuses on the automated analysis of sentence and text structure, and applying those analyses for the *distant reading* of literature, revolving around the question of what distinguishes literary language from other kinds of language.

Álvaro Cuéllar is a postdoctoral researcher currently employed at the University of Vienna, where he specializes in the application of Digital Humanities to Spanish Golden Age literature. His research interests are diverse and include authorship, dating, transcription, rhythmic analysis, orthographic modernization, etc. Notably, he has made significant contributions to the field by uncovering new attributions for relevant dramatists, such as the discovery of *La francesa Laura*, an unknown play by Lope de Vega, which has garnered international attention.

Katharina Dziuk Lameira studied French and Spanish language and literature at the University of Duisburg-Essen and obtained her PhD in Romance studies at the University of Kassel. Her research interests include text complexity, text linguistics, cognitive linguistics, metaphor and second language acquisition. She is currently completing teacher training in Stuttgart, Germany.

Jesse Egbert is associate Professor of applied linguistics at Northern Arizona University. He specializes in register variation, corpus linguistics, and legal interpretation. He is a founding general editor of *Register Studies.* His most recent book is *Designing and evaluating language corpora: A practical framework for corpus representativeness* (Cambridge 2022).

Laetitia Gonon is an associate professor in French language and stylistics at the Université de Rouen, Normandie. She is currently working on phraseology in nineteenth-century novels, in connection with the press and newspapers, and in popular narratives between the nineteenth and the twenty-first centuries.

Ulrike Henny-Krahmer is junior professor for Digital Humanities at the University of Rostock. She wrote her PhD thesis on "Genre Analysis and Corpus Design: Nineteenth-Century Spanish-American Novels (1830–1910)" at the University of Würzburg and has a background in Latin American studies, which she studied at the universities of Cologne and Lisbon. Her research focuses on digital scholarly editing, digital text analysis, and evaluation and sustainability of Digital Humanities research output.

Laura Hernández-Lorenzo is currently a Juan de la Cierva postdoctoral researcher at the Spanish National Distance University (Spain). Previously she was a postdoctoral researcher at POSTDATA ERC project (UNED, Spain), at the Institute of Polish Language (Krakow) and at the University of Seville. She holds a PhD in Spanish literature, which has been awarded as best PhD in Digital Humanities by the BBVA Foundation and the Spanish Association of Digital Humanities. Her research focuses on the application of Digital Humanities, mainly quantitative, computational stylistics and stylometry methodologies to Spanish literature, and especially to Spanish poetry.

Robert Hesselbach studied English/American and Romance studies at the Universities of Würzburg, Austin/TX (USA) and Munich. He earned his PhD in Romance linguistics (University of Würzburg, Germany) with a thesis on syntactic complexity.

He works as a researcher/lecturer at the Friedrich-Alexander-Universität Erlangen-Nürnberg (Germany), where he is currently researching the grammar of Spanish and French, political social media discourse, and the presence of Romance regional and/or minority languages in the digital space.

Clémence Jacquot is an associate professor of literature and holds a PhD in French language and literature from the Université Paris-Sorbonne. Her research focuses on the stylistics of literary texts, with a particular emphasis on genre comparisons.

George Mikros is currently a professor at the MA Program of Digital Humanities at the Department of Middle Eastern Studies at the Hamad Bin Khalifa University in Qatar. Since 1999 and till 2019, he has been a professor of computational and quantitative linguistics at the University of Athens, Greece. He was the founder and the director of the Computational Stylistics Lab at the same institution. Since 2013 he is also adj. professor at the Department of Applied Linguistics at the University of Massachusetts, Boston, USA. He held the position of research associate at the Institute for Language and Speech Processing and was part of research groups that have developed core language resources and NLP tools for Modern Greek. Since 1999 he has held the position of teaching associate at the Hellenic Open University, and from 2016 till 2019, he was the director of the undergraduate program "Spanish Language and Culture." Prof. Mikros has authored 5 monographs and more than hundred papers published in peer-reviewed journals, conference proceedings, and edited volumes. In 2007, he was elected a member of the Council of the International Association of Quantitative Linguistics (IQLA). From 2018 to 2021 he served as its president. He has been the keynote speaker in many international conferences, workshops, and summer schools related to Digital Humanities and quantitative linguistics. His main research interests are computational stylistics, quantitative linguistics, computational linguistics, and forensic linguistics.

Nanette Rißler-Pipka is a digital humanist, literary scholar, and specialist in French and Spanish literature. She is co-managing director of DARIAH-DE and National Coordinator of Germany for DARIAH-ERIC. As such she is also part of the coordination committee of the Association for Research Infrastructure in the Humanities and Cultural Studies (*Geistes- und kulturwissenschaftliche Forschungsinfrastrukturen – GKFI*). She holds a master's degree in comparative literature, Romance languages and economics, as well as a PhD, and a habilitation in Romance literature (University of Siegen). During several visiting professorship positions (Eichstätt-Ingolstadt, Tübingen) and the collaboration with the Helmholtz Association at the Karlsruhe Institute of Technology (KIT) she focused her research on Digital Humanities, digital research infrastructure and on connecting the Romance languages and literature to digital methods.

Jan Rohden received his doctorate in Romance studies from the Universities of Bonn, Florence, and Paris IV (Sorbonne) as part of a trinational graduate program before completing his master of arts in library and information science at Humboldt University of Berlin in 2020. Professionally and academically, Rohden has been engaged with Digital Humanities and research data in various positions since 2016. His research interests include fin-de-siècle literature, Petrarchism, research data, and digital stylometry.

Christof Schöch is professor of Digital Humanities at the University of Trier, Germany, and co-director of the Trier Center for Digital Humanities. He works in the area of computational literary studies, with a focus on analyses of French literature, and pleads for Open Science in the Humanities. Find out more at: https://christof-schoech.de/en.

Julian Schröter is a professor of digital literary studies at the Ludwig-Maximilians-Universität München. His research interests include the methodology and epistemology of computational literary studies, genre theory and interpretation theory. He is currently working on the history of German nineteenth-century novellas. This project, which is a habilitation project at the University of Würzburg, has been funded by the DFG and carried out as a Walter-Benjamin fellowship at the School of Information Sciences at the Universities of Illinois at Urbana-Champaign and the Antwerp University.

Arjuna Tuzzi is a full professor of social statistics at the Department of Philosophy, Sociology, Education and Applied Psychology at the University of Padua, Italy. Her main research interests concern statistical analysis of textual data, data collection tools in social research, statistical methods for the evaluation of university systems, and political-institutional communication. She teaches text mining, social research methods and social statistics for undergraduates and graduates in communication studies and for PhD students in social sciences and in linguistics, philology and literary studies. She is the director of the IQLA-GIAT Summer School in quantitative analysis of textual data, and one of the founders of the Interdisciplinary Text Analysis Group (GIAT). She has been the president of the International Quantitative Linguistics Association for two offices (2014–2018).

Ilaria Vidotto is *première assistante diplômée* in linguistics and stylistics at the University of Lausanne. Her PhD thesis, *Proust et la comparaison vive,* was published in 2020 by Classiques Garnier. Her publications focus on authors of 19th and 20th century French literature (Proust, Balzac, Aragon, Camus, Duras, Radiguet), and on stylistic and rhetorical issues. Her current research focuses more particularly on juvenile works as a stylistic and socio-poetic category.

⇨ Digital Stylistics is an area of research at the intersection of Literary Studies, Linguistics, Digital Humanities, and Computational Literary Studies. It is concerned with the computational and statistical analysis of literary style and of style in language use. This volume brings together research in Digital Stylistics from Romance Studies and beyond, contributing to new methods and applications in different language contexts and literatures. All the research results are based on the empirical, computational analysis of literary corpora chosen to analyze major genres or subgenres of poetry, drama, and prose from the nineteenth to the twenty-first century.

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

ISBN 978-3-96822-201-1
9 783968 222011