Digital Stylistics in Romance Studies and Beyond

February 27th – March 2nd 2019 | University of Würzburg, Germany

Keynotes: Prof. Douglas Biber (Applied Linguistics, Northern Arizona University)
Prof. Glenn Roe (Digital Humanities, Sorbonne Université)

Local Organizers: CLiGS research group (José Calvo Tello, Ulrike Henny-Krahmer, Robert Hesselbach, Daniel Schlör)

Digital Stylistics in Romance Studies and Beyond

Date:
27th of February – 2nd of March 2019

Venue:
University of Würzburg, Germany

Keynotes:
Prof. Douglas Biber (Applied Linguistics, Northern Arizona University)
Prof. Glenn Roe (Digital Humanities, Sorbonne Université)

Local organizers: CLiGS-research group
José Calvo Tello
Ulrike Henny-Krahmer
Dr. Robert Hesselbach
Daniel Schlör

Scientific committee:
Prof. Dr. Elisabeth Burr (Leipzig)
Prof. Dr. Karina van Dalen-Oskam (Amsterdam)
Prof. Dr. Maciej Eder (Kraków)
Prof. Dr. Hanno Ehrlicher (Tübingen)
Prof. Dr. Fotis Jannidis (Würzburg)
Dr. Borja Navarro-Colorado (Alicante)
Prof. Dr. Christof Schöch (Trier)
Prof. Dr. Angela Schrott (Kassel)
# Table of Content

Welcome address by the CLiGS group ................................................................. 4
Welcome address by Prof. Dr. Alfred Forchel ................................................. 5
Welcome address by Prof. Dr. Roland Baumhauer ......................................... 6
Welcome address by Prof. Dr. Fotis Jannidis ................................................... 7
Presentation of the organizing committee ....................................................... 8
Venues ............................................................................................................. 10
Program .......................................................................................................... 14
Presentation of the keynote speakers and abstracts ....................................... 17
Abstracts ...................................................................................................... 19
Bus time tables .............................................................................................. 54
Additional information .................................................................................. 60

#dsrom
Welcome address by the CLiGS group

Dear participants,
Ladies and Gentlemen,

we are very happy to welcome you to the conference “Digital Stylistics in Romance Studies and Beyond” at the University of Würzburg. When we were planning this conference, one point of debate was its disciplinary and thematic scope. For one thing, we wished to provide a forum for the discussion of digital stylistics in general, to bring together experts and those interested in this interdisciplinary field concerned with the computational analysis of literary and language style. On the other side, our intention was to bring the field forward with a conference focusing on one of the aspects of stylistics that are at the heart of the CLiGS project: the study of genre and the work with corpora in Romance languages. We decided to place an emphasis on the linguistic and cultural areas of the Romania, but not exclusively, and thematically open because we believe that the dialogue between researchers concerned with diverse research objects is fruitful for the methodological developments that digital stylistics focuses on. We are delighted that investigators from so many different countries working on French, Spanish, Italian, German, Dutch and English literary and linguistic corpora followed our invitation and look very much forward to the scholarly debate and exchange.

The project CLiGS – “Computational Literary Genre Stylistics” – started in its preliminary phase already in 2014 and fully in 2015 and has been funded by the German Federal Ministry for Education and Research (BMBF) since then. This conference takes place in the final phase of the project and we wish to take the opportunity to present and discuss our work, focused on French drama and Spanish novels, on text classification and models of genre, with all the participants. From the beginning, the CLiGS project has been part of the chair of Computational Literary Studies at the University of Würzburg, benefitting from the expertise in Digital Humanities built up here since 2003.

This conference would not be possible without the support of different organizations and people. Therefore, we would like to thank the German Federal Ministry for Education and Research (BMBF), the University of Würzburg, in particular the Chair of Computational Literary Studies and the Graduate School of Life Sciences, which will host the conference in their new building, and all the mentors of the research group (especially Prof. Christof Schöch, Prof. Brigitte Burrichter, Prof. Angela Schrott, Prof. Hanno Ehrlicher, Prof. Andreas Hotho) and all our colleagues at the Institute.

We wish all participants an interesting stay at our university and also in the city of Würzburg, both rich in tradition, with their local as well as outgoing flair, and many enriching encounters.

Ulrike Henny-Krahmer, José Calvo Tello, Robert Hesselbach and Daniel Schlör
Members of the CLiGS group
Welcome address by Prof. Dr. Alfred Forchel
President of the Julius-Maximilians-University of Würzburg

Ladies and Gentlemen,
dear guests,

Julius-Maximilians-Universität (JMU) Würzburg was originally founded in 1402 with the four traditional faculties of theology, law, medicine and philosophy. Today, JMU has more than 28,000 students in 10 faculties with a wide range of subjects.

With digitalization in virtually all sectors of life, the use of computational methods in academic research and teaching is more important than ever – also in the human and social sciences. The digitization of the cultural heritage is a main subject investigated in the Digital Humanities and has reached a point where new types of research issues can be addressed and are required.

The e-Humanities junior research group “Computational Literary Genre Stylistics” (CLiGS) is affiliated to the Chair of Computational Literary Studies at the University of Würzburg and aims to provide a methodological linkage between new techniques of quantitative analysis of literary texts and the fundamental issues of literary studies in the domain of genre theory and stylistics.

Romance languages offer a unique potential for stylistic analyses with digital methods that has not been explored widely yet. By bringing together international scholars working in the field of stylistic analysis in order to present and discuss innovative methods, digital tools and current research projects, the conference “Digital Stylistics in Romance Studies and Beyond” contributes significantly to the progress of scientific research in this field.

I wish all participants an interesting and stimulating meeting with many insights into new research perspectives, a lively and fruitful exchange with colleagues and international scientists, and a pleasant stay in Würzburg. I hope you also find time to enjoy some marvelous spots of our beautiful city.

Prof. Dr. Alfred Forchel
President
Welcome address by Prof. Dr. Roland Baumhauer
Dean of the Faculty of Arts

Ladies and Gentlemen,

I would like to welcome all participants to the Conference “Digital Stylistics in Romance Studies and Beyond” at the Faculty of Arts of the University of Würzburg.

I am especially delighted that so many colleagues of European and American universities have made it possible to come to Würzburg in order to participate at the conference. I think this also expresses the quality of the e-Humanities junior research group “Computational Literary Genre Stylistics” (CLiGS) that is funded by the German Ministry for Education and Research (BMBF) and affiliated to the Chair of Computational Literary Studies and the Center of Digital Editing at our University.

I am grateful for the effort the local organizing committee has been putting in the organization of this event. Many thanks to them! Finally, I would like to wish you diverse and productive discussions, so that meeting may lead to new knowledge and ideas, and establishes new contacts as well. I hope that you all have a pleasant stay at our university, in our town and the lovely region of Lower Franconia.

Roland Baumhauer
Dean
Welcome address by Prof. Dr. Fotis Jannidis
Chair of Computational Literary Studies

Ladies and Gentlemen,

The BMBF-funded Junior Research Group on computer-aided literary genre stylistics (CLiGS) has developed within a few years into an important element of the Würzburg working group on computational literary studies. The institutional framework for this interdisciplinary working group is the BMBF project Kallimachos, in which a permanent research cooperation between philologists, representatives of the digital humanities and computer scientists has been established. Only a few weeks ago, the new Centre for Philology and Digitality ‘Kallimachos’ at the University of Würzburg was established as the new long-term home of this cooperation. In about three years, it will also move into its own new building and then accommodate relevant researchers and, last but not least, young researchers.

The young scientists of the CLiGS Junior Research Group impressively embody the spirit of interdisciplinary cooperation that characterizes the field of computational literary studies. During the still ongoing constitutional phase of the research field, this includes the systematic overtaxing of all those involved, since it has yet to be negotiated which knowledge sets are now clearly part of the field and which can be dismissed as excessive. If we bear in mind that the work on the dissertation represents a stressful period of socialisation for most people, even when the knowledge stocks within the field of research are clearly defined, then the actual performance of the participants becomes really clear: they have faced this challenge with great enthusiasm for their topics, with a steep increase in competence and with great creativity. And they have developed into widely recognized researchers in the field – a fact that is particularly evident in the appointment of the first head of the junior research group, Christof Schöch, to a professorship for Digital Humanities, but which also applies to the younger members.

All in all, their achievements have set high standards for those coming after them. This is also expressed in the ambitious program of the conference ‘Digital Stylistics in Romance Studies and Beyond’.

I wish all participants of the conference fruitful days and hope you also have time to take a look at the beautiful city.

Fotis Jannidis
Chair of Computational Literary Studies
Presentation of the organizing committee

Dr. Robert Hesselbach
Robert Hesselbach studied English/American and Romance Philology at the universities of Würzburg and Austin/TX (USA). He earned a Master’s degree in Spanish and English Linguistics/Literature and passed the (first) State examination to teach at public Grammar Schools. He holds a PhD in Romance Linguistics from the University of Würzburg and is the head of the CLiGS research group since February 2018. His main research interests are Romance Syntax, Corpus and variational Linguistics, Linguistic Stylistics, History of the Romance Languages, Lexicology and Phonetics/Phonology. Currently he investigates the correlation between the degree of syntactic complexity and genre distinction in modern French (crime) novels. Besides his academic research he is the 2nd Vice-President of the German Association of Romance Scholars (Deutscher Romanistenverband) and runs – together with Christof Schöch (Trier) and Lars Schneider (Munich) – the internet platform www.romanistik.de.

For more information: https://cligs.hypotheses.org/team/robert-hesselbach

Ulrike Henny-Krahmer
Ulrike Henny-Krahmer studied Regional Sciences of Latin America at the University of Cologne and the University of Lisbon. From 2011 to 2015, she worked at the Cologne Center for eHumanities (CCeH) where she was involved in various projects concerned with digital editions and archives. Currently, she is a member of the CLiGS group at the University of Würzburg and working on her PhD project with the working title “Genres as text categories: a stylistic analysis of nineteenth century Spanish American novels (1830-1910)”.

For more information: https://cligs.hypotheses.org/team/ulrike-henny-krahmer
José Calvo Tello
José Calvo Tello studied Spanish Philology and learned programming and mark up languages. He has worked both in linguistic and editorial and corpus building projects. Currently he analyzes the subgenres of the novels of the Spanish Silver Age in his PhD at the University of Würzburg, applying Machine Learning techniques to linguistic features, evaluating the results through metadata.

For more information:
https://cligs.hypotheses.org/team/jose-calvo-tello

Daniel Schlör
Daniel Schlör holds a master’s degree in computer-science from the University of Würzburg and works as PhD student in the junior research group Computational Literary Genre Stylistics (CLiGS). His research interests include data- and text-mining, machine-learning and in general the application of computer-science methods to answer questions from the field of digital-humanities.

For more information:
https://cligs.hypotheses.org/team/daniel-schloer
Venues

(1) **Würzburg Residence: Inauguration and Keynotes**  
(Toscanasaal; Residenzplatz 2, 97070 Würzburg)  
[https://en.wikipedia.org/wiki/W%C3%B6rzburg_Residence](https://en.wikipedia.org/wiki/W%C3%B6rzburg_Residence)

The “Toscanasaal” is located in the south wing of the Würzburg Residence. Once you are in front of the Residence you can access by entering the Residence to the right and go to the second yard. You will find the “Toscanasaal” on the 2nd floor (elevators are available).
(2) Bürgerspital-Weinstuben: Get-together after the Inauguration
(Theaterstr. 19, 97070 Würzburg)
https://en.wikipedia.org/wiki/B%C3%BCrgerspital_zum_Heiligen_Geist
The Conference will be held at the newly built Graduate School of Life Sciences (GSLs). Once you get off the bus at the stop “Am Hubland” you just need to cross the little bridge. If you get off at “Philosophisches Institut” it is a 3-5 minutes walk up the street. The lecture hall is located on the ground floor.
In the old building of the “Alter Kranen” you will find several restaurants, the “Locanda”, the “Beef 800°” and the brewery/restaurant “Brauereigasthof Alter Kranen”. You can access the the restaurant via the street (Mainkai) or via the beergarden that is connected to the restaurant.
### Program

#### Wednesday, 27\textsuperscript{th} of February

**Würzburg Residence (Toscanasaal)**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>17-18</td>
<td>Registration (Residence, in front of Toscanasaal)</td>
</tr>
<tr>
<td>18-20</td>
<td>Conference opening: Robert Hesselbach &amp; Christof Schöch &amp; Baris Kabak (Vice-President of the University of Würzburg)</td>
</tr>
<tr>
<td></td>
<td>Keynote</td>
</tr>
<tr>
<td></td>
<td>Douglas Biber (Northern Arizona University)</td>
</tr>
<tr>
<td></td>
<td>Using corpus-based analysis to study fictional style: A multi-dimensional analysis of variation among and within novels</td>
</tr>
</tbody>
</table>

#### Thursday, 28\textsuperscript{th} of February

**Graduate School of Life Sciences, Campus Hubland Nord**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-9.30</td>
<td>Simon Gabay (Neuchâtel)</td>
</tr>
<tr>
<td></td>
<td>Français vs francois: does linguistic normalisation affect stylometric results?</td>
</tr>
<tr>
<td>9.30-10</td>
<td>Andreas von Cranenburgh (Groningen)</td>
</tr>
<tr>
<td></td>
<td>Dutch weak and strong pronouns as a stylistic marker of literariness</td>
</tr>
<tr>
<td>10-10.30</td>
<td>Sascha Diwersy (Montpellier) / Olivier Kraif (Grenoble)</td>
</tr>
<tr>
<td></td>
<td>Patterns and Novels – an outline of the PhraseoRom project and its provisional results</td>
</tr>
<tr>
<td>10.30-11</td>
<td>Coffee break</td>
</tr>
<tr>
<td>11-11.30</td>
<td>Martin Wynne (Oxford)</td>
</tr>
<tr>
<td></td>
<td>Exploring Rhetoric in the Electronic Enlightenment</td>
</tr>
<tr>
<td>11.30-12</td>
<td>Arjuna Tuzzi (Padova) / George Mikros (Athens) / Michele A. Cortelazzo (Padova)</td>
</tr>
<tr>
<td></td>
<td>Applying General Imposters’ method to the Ferrante’s case</td>
</tr>
<tr>
<td>12-14</td>
<td>Lunch break</td>
</tr>
<tr>
<td>14-14.30</td>
<td>Katharina Dziuk Lameira (Kassel)</td>
</tr>
<tr>
<td></td>
<td>Complexity and Style of Spanish literary Texts</td>
</tr>
<tr>
<td>14.30-15</td>
<td>Fotis Jannidis (Würzburg)</td>
</tr>
<tr>
<td></td>
<td>Text complexity and style</td>
</tr>
<tr>
<td>15-15.30</td>
<td>Coffee break</td>
</tr>
<tr>
<td>15.30-16</td>
<td>Julian Schröter (Würzburg)</td>
</tr>
<tr>
<td></td>
<td>The challenge of exploring the style of the German Novelle as a virtually orderless genre</td>
</tr>
<tr>
<td>16-16.30</td>
<td>Daniel Schlör (Würzburg)</td>
</tr>
<tr>
<td></td>
<td>Preparation of a Text Type Dataset</td>
</tr>
<tr>
<td></td>
<td>Bootstrapping Rare Classes for the Annotation Process</td>
</tr>
<tr>
<td>18-20</td>
<td>City tour</td>
</tr>
<tr>
<td></td>
<td>Meeting point: fountain in front of the Residence</td>
</tr>
<tr>
<td>Time</td>
<td>Speaker</td>
</tr>
<tr>
<td>--------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>9-9.30</td>
<td>Jan Rohden (Göttingen)</td>
</tr>
<tr>
<td>9.30-10</td>
<td>Laura Hernández Lorenzo (Sevilla)</td>
</tr>
<tr>
<td>10-10.30</td>
<td>Anne-Sophie Bories (Basel)</td>
</tr>
<tr>
<td>10.30-11</td>
<td></td>
</tr>
<tr>
<td>11-11.30</td>
<td>Jonathan Armoza (New York)</td>
</tr>
<tr>
<td>11.30-12</td>
<td>Nanette Rißler-Pipka (Karlsruhe)</td>
</tr>
<tr>
<td>12-14</td>
<td></td>
</tr>
<tr>
<td>14-14.30</td>
<td>Ulrike Henny-Krahmer (Würzburg)</td>
</tr>
<tr>
<td>14.30-15</td>
<td>José Calvo Tello (Würzburg)</td>
</tr>
<tr>
<td>15-15.30</td>
<td></td>
</tr>
<tr>
<td>17-19</td>
<td>Keynote (at Toscanasaal, Würzburg Residence)</td>
</tr>
<tr>
<td>20.00</td>
<td>Conference Dinner (Alter Kranen)</td>
</tr>
<tr>
<td>Time</td>
<td>Speaker(s)</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------------</td>
</tr>
<tr>
<td>9-9.30</td>
<td>Alvaro Cuéllar González (Kentucky)</td>
</tr>
<tr>
<td>9.30-10</td>
<td>José Manuel Fradejas Rueda (Valladolid)</td>
</tr>
<tr>
<td>10-10.30</td>
<td>Coffee break</td>
</tr>
<tr>
<td>10.30-11</td>
<td>Clémence Jacquot (Montpellier) / Ilaria Vidotto (Grenoble) / Laetitia Gonon (Grenoble)</td>
</tr>
<tr>
<td>11-11.30</td>
<td>Simone Rebora (Verona) / Massimo Salgaro (Verona)</td>
</tr>
<tr>
<td>11.30-12</td>
<td>Final discussion &amp; information on publication etc.</td>
</tr>
</tbody>
</table>
Presentation of the keynote speakers and abstracts

Prof. Douglas Biber
Prof. Biber holds a Ph.D. from the University of Southern California (1984) and was appointed as an Assistant Professor at the Department of Linguistics at the University of Southern California (1984-1990). Since 1990 he works as a Professor at the Department of Applied Linguistics at the University of Northern Arizona (since 2000 as a Regents’ Professor). He has published various books and articles focusing on Corpus Linguistics and register variation and style. For more information, please visit the following website: https://dougbiber.weebly.com/

Abstract

Using corpus-based analysis to study fictional style: A multi-dimensional analysis of variation among and within novels

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 talk supports this claim through an MD analysis of linguistic variation in a 5-million word corpus of fictional prose in English. The linguistic dimensions are interpreted in functional terms, and then applied to describe the fictional style of individual novels, focusing on the ways in which particular novels are distinctive relative to the range of variation among all novels. Although fictional prose might be stereotypically regarded as ‘literate narrative’, the MD analysis shows that both the ‘oral’-’literate’ opposition as well as the ‘narrative’-’non-narrative’ opposition are fundamentally important in this discourse domain.

In conclusion, the talk briefly illustrates how the MD framework can be applied to an additional issue for stylistic analysis: describing the patterns of linguistic variation that occur within the scope of an individual novel. For this purpose, novels are automatically segmented into topically coherent ‘discourse units’, based on the distribution of vocabulary in the novel (see Biber, Connor, Upton 2007). Subsequently, the linguistic dimensions of variation can be applied to track shifts in linguistic style across discourse units within a novel.
Prof. Glenn Roe
Prof. Roe holds a Ph.D. in French Literature from the Department of Romance Languages and Literatures at the University of Chicago. He was a (Senior) Lecturer in Digital Humanities at the Centre for Digital Humanities Research at the Australian National University (2013-2015; 2016-2018). From 2017 to 2018 he worked as a Visiting Professor at the Labex OBVIL (Observatoire de la vie littéraire) at Sorbonne Université (Paris). In 2018 he was appointed as a Professor for Digital Humanities (Professeur des universités) at Sorbonne Université in Paris. He is the author/editor of various books and articles focusing on the use of digital methods for literary analysis. His current research projects focus on new computational approaches with traditional literary and historical research questions, such as the concept of ‘authorship’ in Enlightenment France, and its relationship with ‘authority’ over the long 18th century. For more information, please visit the following website: http://www.glennroe.net/

Abstract
Voltaire’s Style: A Study in Digital Methods

Stylometry is one of the oldest, and better-known approaches to computational literary studies. Primarily concerned with issues of author attribution and small-scale stylistic/linguistic analysis, quantitative studies of literary ‘style’ have in recent years become almost ubiquitous in digital humanities circles. But, as digital collections grow larger by the day, and new data-intensive digital methods are developed, perhaps it is time to revisit the stylometric notion of ‘style’ in the era of ‘big data’. Taking Voltaire as a test-case, whose complete works and letters stretch to over 22,000 documents and 13 million words, this talk will outline attempts to leverage new digital tools and methods to reconsider our notion of style as it pertains to the 18th-century’s most famous writer, stressing the often-overlooked aesthetic of innovation inherent to many of his writings.
Abstracts

Jonathan Armoza (New York)

Probabilistic Matrix Factorization for Digital Humanities: Modeling the Parts of Speech of Emily Dickinson’s Fascicles

Research into data modeling methods called collaborative filtering has been bolstered by the tech community. These methods are meant to characterize sparse data sets, enabling clustering of records and provide plausible inference of missing values. Their outputs are recognizable in the recommendation sections of internet commerce sites like Amazon, or streaming services like Netflix. One such method called Nonnegative Matrix Factorization (NMF) has proven successful in these tasks. By factorizing a matrix of records, one can approximate two separate factor matrices that hypothetically produced the data set, thus estimating latent factors that led us to an observed reality. In the case of Netflix, it determines a quantified representation of the hidden attributes of subscribers that led them to rate certain movies, or of the hidden attributes of movies that led subscribers to rate them. Subscribers or movies can then be grouped and characterized per their latent commonalities. NMF can be extended via probabilistic modeling, a realm of methods familiar to digital humanists (e.g. topic modeling).

These methods can be applied to aspects of the humanities’ objects of study that have digital representations, including the parts of speech (POS) of texts. In this talk, I explain probabilistic matrix factorization (PMF) and demonstrate how this method and an accompanying modeling program “Nimfa” can be used to characterize the POS of the “fascicle” books of Emily Dickinson – which act as bibliographic and chronological structure for this exploration. By using a POS tagger over her poems, and then running their POS counts through PMF, latent patterns emerge that may have informed Dickinson’s poem arrangement. Here, PMF exposes "hidden poems" whose POS qualities tell us about the actual poems that are associated with them. By examining how actual poems statistically vary from those hidden ones, I identify POS dynamics that Dickinson included and positioned in each fascicle. Suggestive visuals are included to help the audience navigate the output of the factorization, showing how clustering can be juxtaposed with bibliographic arrangement. By conclusion, I show how this latent POS usage has also identified thematic patterning in Dickinson’s poems.

*****

Anne-Sophie Bories (Basel)

A Tempo for Negritude in Césaire’s Cahier

This research stemmed from my efforts to develop a tool for exploring French free verse in a large digitized corpus. In the search for a descriptive device light enough to be reasonable easily put in place, flexible enough to adapt to various poets, and specific enough that it would capture the specificities of French free verse, I decided to focus on the stability of syllables. French strict verse
is particular for two reasons. First, it is exclusively syllabic, without feet, merely requiring a caesura for its longer metres. Second, the number of syllables in lines, which alone sets the metre, is neither the actual number of syllables nor the number of pronounced syllables. The verse language, in French, has its own rules for counting syllables, and maintains many syllables that are muted by the standard pronunciation. This lends French strict verse a particular tempo, slower, with a higher syllable to word ratio, and this tempo is perceived by the readers not only as a marker of verse but as a marker of poetry.

The method I developed in collaboration with Bertrand Gaiffe of the ATILF laboratory allows me to tag all syllables in the corpus as belonging to one of three categories: stable (syllables pronounced our counted in all cases), muted (syllables muted in all cases) and unstable (syllables muted by the standard language but counted by the verse language). My idea was to measure the tempo of various free verse poets, and see how it differed between poets, and how it compared to strict verse tempo.

Aimé Césaire is the best known French caribbean poet, and the inventor of the word “negritude”. His *Cahier d’un retour au pays natal* is a bewildering, daring and powerful cry, both in its subject (the very long poem progresses from a furious reprobation of his Martinique island, to an exhilarated rebellion of the *nègre*).

Applying my tool to Césaire’s *Cahier* posed specific challenges, because the *Cahier* mixes lines of free verse with paragraphs of prose, with no boundary, no separation of any kind between the two forms.

The results of this analysis walked me to an exploration of the metrical and phonetical environment of the word *nègre*, racist and mostly pejorative in French, and of its creole counterpart, the word *nèg*, which has very different and sometimes laudatory connotations. In this paper, I shall first offer an overview of my method, then give a short insight into my findings regarding Césaire’s shifting use of the word *nègre*, and their implications in terms of stylistics and poetics.

*****

José Calvo Tello (Würzburg)

*Subgenre Classification: Linguistic vs. Literary Features*

In the field of computational genre analysis simple lexical information has been proven to be hard to beat as the best type of feature. This was the result already in one of the first papers about the topic, by Kessler, Numberg, and Schütze (1997), with insignificant advantages when using grammatical annotation. Other researchers like Santini (2011) or Underwood (2014) have pointed out the usefulness of lexical information even when other features were extracted. Some recent works in the field that compared more complex features such as topic modelling, of literary characters or sentiment analysis did not surpassed the baseline of the lexical frequency (Hettinger et al. 2016; Henny-Krahmer 2018). Other kind of features such as curated literary metadata has been used only in very tentative manners (Underwood 2014; Wilkens 2016).
How is it possible that complex concepts such as literary genres are best predicted with simple linguistic features like frequency of tokens? One of the reason might lay in the differentiation between complex features and complex models (Müller and Guido 2016, 30–31; Alpaydin 2010, 31–32). Frequencies of tokens are easy to compute when compared to other techniques in which much preprocessing is needed (character networks, topics), but they can be used by thousands. A model of thousands of pieces (features) is a more complex model than one with only a dozen parts. Even when tokens constitute simple features (easy to compute), the researchers normally create complex models with them (normally around 1,000 and 5,000 features).

In this work I will compare the results of the classification of subgenres using different types of features:

- lexical frequencies as baseline
- literary metadata extracted manually
- mixed features using stylistic and lexical information

For that I will use the *Corpus of Novels of the Spanish Silver Age* (CoNSSA) with novels published between 1880 and 1939 by Spanish writers. This corpus has been annotated manually with complex literary metadata such as information about the protagonist (age, gender, social status), place and period of the plot, and other literary phenomena (narrator, type end, auto-referentiality, etcetera). Besides, the texts have been textually (for example difference between narrative passages and direct speech) and linguistically annotated (using the NLP tool Freeling). The entire data (text, metadata and annotation) has been encoded in TEI, allowing the extraction of complex features. The subgenre classes have been extracted from several sources such as National Library, Amazon, ePubLibre and own annotation, concluding into twenty consistent semantic subgenres (among other dialogue novel, historical novel, war novel, humorist novel or adventure novel).

These classes (in binary form) will be classified using different classification algorithms (SVM, Logistic Regression, Random Forest) and ways of weighting the features (relative frequency, TF-IDF, z-scores, binary). The results of the different features will be reflected not only in terms of accuracy of the prediction, but also in the terms of their complexity and how explanatory are for humans.

References


The Dutch language (along with other languages) has full and reduced versions of some of its personal pronouns (see Table 1). Full pronouns are also called emphatic or strong, while in Dutch the reduced pronouns are weak pronouns (as opposed to clitics in Romance languages, which 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 2010). 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 refer to a less salient referent; strong pronouns are required when expressing contrast or in comparisons.

The distinction also turns out to correlate with the perceived literariness of novels. In the project The Riddle of Literary Quality, readers from the general public rated 401 recent Dutch-language novels 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.

We focus on cases where there may be a choice between weak and strong, so we only consider pronouns with both versions. We exclude the neuter pronoun het since it is also a determiner; similarly, “haar” is also a possessive pronoun; while “m” is not common in written language. This leaves us with the following regular expressions:

Strong: (mij|jij|jou|mij|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. We first look at the overall frequency of both pronoun types; for example, more literary novels might focus more on ideas than people. See Fig. 1 for the results. There is indeed a negative correlation of pronoun frequency and literariness.

Fig. 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% of pronouns are weak across the 401 novels). While we already found a strong correlation for pronoun frequency, Figure 2 shows that the strong/weak distinction yields a stronger correlation, demonstrating that the distinction is stylistically meaningful. A straightforward explanation would be that weak pronouns are a proxy for informality. However, if the result is due to the linguistic properties related to discourse salience and contrast, the effect may signal a more complicated discourse structure.

Additionally, we see a striking set of outliers in Fig. 2 with a substantially higher proportion of strong pronouns (>35%). The outliers are listed in Table 2; except for Forsyth and Mitchell, they
are by Dutch authors (the corpus also includes translated novels) and rated as highly literary (except for Forsyth, a thriller).

<table>
<thead>
<tr>
<th></th>
<th>Strong</th>
<th>Weak</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st sg</td>
<td>(ik), mij</td>
<td>me</td>
</tr>
<tr>
<td>2nd sg</td>
<td>jij, jou</td>
<td>je</td>
</tr>
<tr>
<td>3rd sg fem</td>
<td>zij, haar</td>
<td>ze</td>
</tr>
<tr>
<td>3rd sg masc</td>
<td>(hij), hem</td>
<td>'m</td>
</tr>
<tr>
<td>3rd sg neut</td>
<td>het</td>
<td>'t</td>
</tr>
<tr>
<td>1st pl</td>
<td>wij, (ons)</td>
<td>we</td>
</tr>
<tr>
<td>2nd pl</td>
<td>(jullie)</td>
<td>-</td>
</tr>
<tr>
<td>3rd pl</td>
<td>zij, hen/hun</td>
<td>ze</td>
</tr>
</tbody>
</table>

Table 1: Strong and weak personal pronouns in Dutch. The Strong column shows subject and object forms. The pronouns in parentheses have no weak counterpart.

<table>
<thead>
<tr>
<th></th>
<th>Strong %</th>
<th>Weak %</th>
<th>Both %</th>
<th>Strong prop.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Springer, Quadriga</td>
<td>1.43</td>
<td>0.98</td>
<td>2.41</td>
<td>59.3</td>
</tr>
<tr>
<td>Mitchell, The thousand autumns of ..</td>
<td>0.74</td>
<td>0.56</td>
<td>1.29</td>
<td>56.8</td>
</tr>
<tr>
<td>Kooten, Verrekijker</td>
<td>0.99</td>
<td>0.85</td>
<td>1.84</td>
<td>53.7</td>
</tr>
<tr>
<td>Dewulf, Kleine dagen</td>
<td>1.28</td>
<td>1.60</td>
<td>2.87</td>
<td>44.4</td>
</tr>
<tr>
<td>Bernlef, De een zijn dood</td>
<td>0.98</td>
<td>1.40</td>
<td>2.38</td>
<td>41.2</td>
</tr>
<tr>
<td>Verhulst, De laatste liefde van..</td>
<td>0.75</td>
<td>1.28</td>
<td>2.02</td>
<td>36.8</td>
</tr>
<tr>
<td>Abdolah, Koning</td>
<td>0.68</td>
<td>1.19</td>
<td>1.87</td>
<td>36.4</td>
</tr>
<tr>
<td>Forsyth, The Cobra</td>
<td>0.40</td>
<td>0.70</td>
<td>1.09</td>
<td>36.3</td>
</tr>
<tr>
<td>Japin, Vaslav</td>
<td>1.11</td>
<td>1.95</td>
<td>3.05</td>
<td>36.2</td>
</tr>
</tbody>
</table>

Table 2: Novels that are outliers with respect to the proportion of strong pronouns.
Figure 1: Percentage of pronouns with respect to all words, correlated against literary ratings.

Figure 2: Percentage of strong pronouns with respect to both pronoun types, correlated against literary ratings.

References

*****
Álvaro Cuéllar González (Kentucky)

Presentation of the Estilometría TSO Project: Stylometry Applied to Spanish Golden Age Theatre

In recent years, stylometry, highly developed by the Computational Stylistics Group, has been applied to numerous authorial problems. However, in my judgement, it has not been sufficiently utilized with one of the most fruitful territories in this discipline: the Spanish Golden Age Theatre. It is a corpus with thousands of plays in which many cases of doubtful authorship abound. Furthermore, the authorial problems of the theatre are not restricted to peripheral works, but rather to famous masterpieces such as *El condenado por desconfiado* and *El burlador de Sevilla*.

In the cases where stylometry has been applied to try to shed light on this problem, only a few dozen works have been used to carry out the analyses, which I find insufficient. For this reason, I with the help of the University of Valladolid professor, Germán Vega García-Luengos, who is expert in Golden Age Theatre, have been working for two years in creating a corpus of more than 800 plays by more than 30 authors of Spanish Golden Age Theatre. In doing this research, we have counted on the help of research teams and independent researchers from around the world, who have sent us many of these plays. The preparation of the corpus has not been free of complications and decisions: we have to homogenize the 17th century orthography of the texts, decide whether to include stage directions or not, and eliminate everything outside the main texts.

Upon working with the considerable corpus, we can use it to analyze authorial problems. Firstly, we check that stylometry really works with Spanish Golden Age Theatre. To get this, we have done some experiments with works of undoubted authorship using cross validation. The results have been very satisfactory: above a 95 percent success rate. After having demonstrated the effectiveness of the stylometry with our undisputed corpus, we are authorized to use it with real authorship problems. Until now, we have carried out some analyses with promising results that still have to be completed and refined by the traditional philology. One conclusion seems clear, stylometry is an irreplaceable tool to point out cases that deserve to be investigated. When we work with our entire corpus, unexpected connections begin to appear and, if we analyze them carefully, they usually lead us to stimulating issues of authorship. The process proves contrary to the usual one, we do not usually start with a specific authorial problem, but rather we let the corpus speak to us and show unforeseen connections, which we later analyze. Finally, we offer the results for free to anyone who asks for it, we want our project to be an additional tool for the different research teams.

In conclusion, my paper intends to present for the first time in public the Estilometría TSO project, which while it still continues to develop, I firmly believe that it is already mature enough to be a valuable resource for the literary research community.
The aim of our talk is to give an overview of the ANR-DFG founded project *PhraseoRom*, its main research topics, its resources / data and methodology, and some preliminary results.

The central hypothesis of the *PhraseoRom* project is that lexico-grammatical patterns are distinctive features of discourse genre, and that it is thus possible to develop a structural and functional typology of lexico-syntactic constructions (LSCs) that provides a viable basis for genre description and comparison. This claim has several implications, which will be spelt out as follows.

The first part of our talk will deal with the theoretical background to the notion of LSC and the extent to which it draws on Sinclair’s work on compound lexical items (2004:30 ff.), Hunston and Francis’ *Pattern Grammar* (2000) and Hoey’s *Theory of Lexical Priming* (2005).

The second part of the talk will be concerned with methodological issues. In a first step, we will introduce the concept of Recurring Lexico-syntactic Trees (RLT, cf. Kraif & Diwersy 2014:389 ff.; Tutin & Kraif 2017; see Appendix 2) as a technique for extracting potential instances of LSCs from annotated corpora. In a second step, we will point out several levels of structural and functional description involved in the linguistic analysis of the extracted patterns and the identification of the constructions they represent.

In the third part, the focus will be on the implementation of the outlined theoretical and methodological principles within the framework of the *PhraseoRom* project. After a brief presentation of the project’s working corpora – which at present comprise more than 2500 British, French and German twentieth-century novels (see Appendix 1) – we will expose the categorisation scheme underlying the typology of novel specific LSCs in terms of their semantic structure and discursive (narrative) function. Elements of this typology will be illustrated by the results of some case studies such as Dyka, Novakova & Siepmann (2017) or Gonon et al. (2018). The remainder of part 3 will deal with our current experiments on the classification of novelistic genres by means of textometric methods (cf. Lebart, Salem & Berry 1998), in particular correspondence analysis (CA) and the computation of occurrence specificities. For instance, we will report on the results of a series of CA analyses based on authors with “multi- genre” portfolios.

The talk will conclude with an outlook on *PhraseoRom*’s upcoming milestones, and especially the finalisation of the platform providing online access to the project’s data and tools.

References


Table 1: Working corpora of the PhraseoRom project

<table>
<thead>
<tr>
<th>Genre</th>
<th>French</th>
<th></th>
<th></th>
<th>English</th>
<th></th>
<th></th>
<th>German</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#Texts</td>
<td>#Authors</td>
<td>#Tokens</td>
<td>#Texts</td>
<td>#Authors</td>
<td>#Tokens</td>
<td>#Texts</td>
<td>#Authors</td>
<td>#Tokens</td>
</tr>
<tr>
<td>FANTASY</td>
<td>104</td>
<td>43</td>
<td>13323976</td>
<td>142</td>
<td>41</td>
<td>17814599</td>
<td>110</td>
<td>41</td>
<td>16666339</td>
</tr>
<tr>
<td>LITERARY FICTION</td>
<td>445</td>
<td>170</td>
<td>34334554</td>
<td>147</td>
<td>89</td>
<td>17442917</td>
<td>90</td>
<td>41</td>
<td>9068325</td>
</tr>
<tr>
<td>HISTORICAL</td>
<td>114</td>
<td>39</td>
<td>14868273</td>
<td>116</td>
<td>48</td>
<td>17764261</td>
<td>149</td>
<td>38</td>
<td>25454115</td>
</tr>
<tr>
<td>CRIME</td>
<td>194</td>
<td>84</td>
<td>17859351</td>
<td>115</td>
<td>55</td>
<td>13561947</td>
<td>97</td>
<td>35</td>
<td>11117326</td>
</tr>
<tr>
<td>ROMANCE</td>
<td>112</td>
<td>40</td>
<td>9802410</td>
<td>96</td>
<td>35</td>
<td>13295474</td>
<td>208</td>
<td>44</td>
<td>18689013</td>
</tr>
<tr>
<td>SCIENCE FICTION</td>
<td>147</td>
<td>39</td>
<td>13173618</td>
<td>116</td>
<td>39</td>
<td>12962491</td>
<td>92</td>
<td>38</td>
<td>8347856</td>
</tr>
<tr>
<td>TOTALS</td>
<td>1116</td>
<td>415</td>
<td>103362182</td>
<td>732</td>
<td>307</td>
<td>92841689</td>
<td>746</td>
<td>237</td>
<td>89342974</td>
</tr>
</tbody>
</table>

Figure 1: RLT extracted from the French crime fiction corpus

Katharina Dziuk Lameira (Kassel)

Complexity and Style of Spanish literary Texts

This presentation discusses whether text complexity can be seen as a dimension of style, and if so, which linguistic features are suitable for the description of both, complexity and style. Text
complexity is defined as the interaction of different text features that influence text difficulty and can be measured and observed objectively. Text difficulty, however, is considered a consequence of the interaction of text complexity and extralinguistic parameters, such as educational background or reading experience. Following Dahl’s definition of relative complexity (2004) a text is more complex the more it deviates from typical patterns. Analogously style can be seen as deviation from patterns in texts. The project integrates different notions of complexity to a model of text complexity for Spanish learners. The research study consists of two steps. First, the complexity of a text collection consisting of about 30 Spanish literary text excerpts is analyzed quantitatively and qualitatively regarding lexical, semantic, syntactical, morphological and textual features to provide a complexity profile for each text. The text collection consists of literary text excerpts from novels used in Spanish courses at the University of Kassel and was compiled based on didactic considerations. The text excerpts consist of logical units, contain no direct speech and are grouped by length. In a second step, a set of representative texts will be chosen from the corpus for an online questionnaire that will be presented to: German-speaking students of Spanish philology which are grouped according to their Spanish proficiency level (CEFR levels A2-C1), and native speakers of Spanish studying philology. Both groups will rate the difficulty of the given texts and answer questions concerning the style of the texts. The mean values of these text ratings will be used for a regression analysis to identify text features contributing to text complexity. Questions which will be addressed are amongst others: What is the current state of research regarding the intersection between complexity and style, which stylistic aspects have an impact on text difficulty and which methods can be used to analyze the complexity of texts?

References

*****

José Manuel Fradejas Rueda (Valladolid)
The authorship and the redactions of the Primera Partida in the light of stylometry

Stylometric analysis with the aid of stylo (Eder, Rybicki, Kestemont 2016) helped us to detect the different “hands”, i.e. sources, hidden in some Castilian medieval works (Fradejas Rueda, in press)
as the _Libro de la montería_ (before 1350) and the _Libro de las aves que cazan_ (1386). In this paper we propose to analyze with the help of stylo one of the most complex works produced in the Kingdom of Castile in the Middle Ages: the _Siete Partidas_ (7P).

The 7P is complex legal text attributed to Alfonso X (1252-1284) divided in seven part, hence its title _Siete Partidas_. It was composed between 1256 and 1272 over a long drafting process, and was established as an effective code of law, although secondary, in 1348. However, the textual fixation would only be achieved in 1491, when the _editio princeps_ was published, but to be replaced by a new edition in 1555, edition that was instituted as the legal version to be used in any court of law. (The 7P is a legal code still in force in the US, Brazil and some Spanish-American republics.)

There is no critical edition of the text, and perhaps it will never be achieved if the intended aim is the Ur-text due to the long process of writing and reworking that has suffered from its inception in 1256 until its final promulgation in 1348. Traditional criticism has divided the drafting process of the _Primera Partida_ into three phases represented by several manuscript witnesses (Craddock 1974, Pérez Martín 2014), and based in the reordering and reworking of a few laws and sections, the addition and deletion of a few other. We would like to see if stylometric technique can help to unveil the true composition process of this legal code.

A first analysis with stylo package on the text of the 1491 edition shows that within the 7P there are two clear authorial styles, as can be seen in figure 1.

![TEST Cluster Analysis](image)

**Fig. 1.**

And the provisional results of the application of stylo to the _Primera Partida_ (according to 1491 edition) divided in its 24 for sections are interesting and intriguing, as evidenced by the dendrogram in figure 2.
As in the project 7PartidasDigital (FFI2016-75014-P AEI-FEDER, EU) we have transcribed and codified in XML-TEI the 1491 and 1555 editions of the text, and we are working on the transcription and coding of the Primera Partida according to manuscripts LBL (done), HS1, MN0 (done), MN6, ZAB and T11 (for identification of these witnesses see Fradejas Rueda 2016), we will analyze, with the aid of stylo, all these manuscripts to find out how many “hands” intervened in each of those writings phases and to what extent and try to prove or falsify the traditional drafting process.

References

*****
French literature of the 17th c. is usually normalised, i.e. modern French *scripate* are aligned on the contemporary orthography. Among the many reasons invoked to defend this tradition, a new one has recently appeared: it supposedly simplifies stylometric analysis. Indeed, non-systematic spellings introduce noise in the results, two tokens (*e.g.* “étois” *vs* “estois”) of the same type (vb. *être*, ind. impft, P1) being catalogued as two types rather than one. But is it really a problem? Don’t we lose stylometric information by normalising? The question is becoming important, because not only conservative editions are already available (*e.g.* the one of Sorel’s Berger *extravagant* [6]), but their number might increase in a near future future, both for technical reasons, with the development of OCRisation, and for scientific reasons, with the publication of papers advocating for a modification of editorial practices [3, 1].

First tests carried on an original (fig. 1) and a normalised (fig. 2) version of Guez de Balzac’s *Correspondance* (1624 edition) seem to show that clusters are not heavily affected by normalisation:

These dendograms, calculated with Burrow’s Delta, show a cluster analysis of the five most important correspondants of Guez de Balzac. Grouping Boisrobert (the good friend) and Hydaspe (the brother) on the one hand, and the duke of Epernon, the cardinal of La Valette and the bishop of Ayre (the noblemen) on the other hand is the most expected outcome: fig. 1 and 2 therefore provide logical results.

However, if we watch carefully, we see that the regularised text provides slightly stronger results (fig. 2) than non-regularised text (fig. 1): distances are closer to 1 σ in the first case. Interestingly, the Manhattan distance gives opposite results:
Distances based on words frequency between correspondants are ranging between 8 to 11 points for the non-normalised text (fig.3), but remains always above 10 for its regularised version (fig. 4).

In some cases, normalisation seem to create more important problems. It is the case with the euclidean distance:

As we see, only the non-normalised text (fig.5) provides the same clusters than previously, while the normalised version (fig. 6) does not group Hydaspe and Boisrobert together anymore.

Diverging results with the same distance measurement on a normalised and a non-normalised version of the same text implies that not only the lexicon, but the graphical system is assessed. Therefore, it is important to know to what extent ecdotic choices affect our analysis, and to determine which type of calculation provides reliable results.
Guez de Balzac being too small a corpus, we will also use another one based on 17th c. plays. Author clustering for 17th c. French theater has indeed shown excellent results with normalised texts (fig. 7), and permits large-scale analysis that will provide stronger results than Guez’s Correspondance. It will also be an opportunity to reopen the famous Molière-Corneille case [4, 2], which was thought to be closed, but has been recently reopened on another front [5].

Figure 7: Author clustering for 17th c. French

This paper will also be the opportunity to present a workflow specific to Grand Siècle French from the pdf to R, including on the fly normalisation with neural machine translation, and to create a parallel corpus of original and normalised French theatre of the 17th c.

References

*****

Ulrike Henny-Krahmer (Würzburg)

Family Resemblance in Genre Stylistics

The concept of “family resemblance” was introduced into genre theory in the 1960’s as an alternative to the understanding of literary genres as rigid classes whose definition is based on a set of necessary elements (Fowler 1982, Fishelov 1991, Hempfer 2010). Going back to an analogy used...
by Wittgenstein to describe linguistic activities, family resemblance instead proposes loose relationships between works belonging to a certain genre: “we see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail. I can think of no better expression to characterize these similarities than ‘family resemblances’; for the various resemblances between members of a family: build, features, colour of eyes, gait, temperament, etc., etc., overlap and criss-cross in the same way.” (Wittgenstein 1958: 32).

In this contribution, the idea of family resemblance is applied in an analysis of subgenres of the novel. The overall corpus consists of 250 Spanish American novels from the 19th century. Some of them can be clearly associated with a subgenre, for example, the historical novel, while others can in the first place only be understood as representatives of a certain literary school or as “general fiction”. It is examined how the different types of novels relate to each other in a family resemblance setting.

To model family resemblance, similarities between the texts are established on the basis of different sets of features (most frequent words, grammatical categories, topics, and named entities). Following a method proposed by Eder (2017), a ranking of nearest neighbours is used to map the textual similarities to a network structure. Using a multigraph, this approach is extended here to include several levels of connections corresponding to the different kinds of features. Because the similarities only overlap in part, this results in looser clusters of nodes in the network. In this approach, the critical question arises how the various textual signals (authorial, generic, temporal, etc.) can be separated. Two important aspects in this regard are the corpus design and the selection of features. Ultimately, the application of the concept of family resemblance in the digital text analysis setting shows that it can be applied to other textual signals beyond genre, as well.

References

*****

Laura Hernández Lorenzo (Sevilla)
Digital Stylistics applied to Golden Age poetry: is really Fernando de Herrera a transitional poet between Renaissance and Baroque?

Even though approaches to Literature through Digital Stylistics and Stylometry have increased over the years, Digital Stylistic approaches to Spanish Literature are still few if compared with the amount of studies in other languages, specially English. On the other hand, quantitative approaches have been more focused on narrative and long texts (Moretti, 2007), whereas poetry presents more
challenges to computational approaches due to both its shortness and its greater subjectivity component compared to the rest of literary genres. Although there can be found some interesting studies on Golden Age Spanish Poetry—mainly about compilation of corpora and metrical aspects (Navarro Colorado, 2015; Ruiz Fabo, Martínez Cantón, Poibeau, & González-Blanco, 2017)—, there is further research needed on the appliance of computational analysis to the stylistic differences between Renaissance and Baroque style and the place of the authors who acted, according to literary scholars (López Bueno, 2006), as a transition between both styles. One of the most important poets of the Golden Age, Fernando de Herrera, has been claimed to play a crucial role in the evolution of poetic style towards Baroque (López Bueno, 2000). Moreover, his attributed poetic work $P$ has been considered more similar to Baroque’s style and somehow a precedent of Góngora’s style and works (Macrí, 1972). This has been studied to-date using traditional methodologies such as Philology or History of Literature.

Therefore, some questions arise: could digital techniques be useful in researching Herrera’s place in the Golden Age? Are they effective to see if $P$ has a similar place or not? Will Herrera’s undoubted and attributed works appear as more similar to Renaissance or Baroque? Previous research has explored the possibility of characterize Golden Age Spanish Poetry analyzing metrical and semantic aspects, but would plain vocabulary/lexical and NLP information be sufficient to add new insights through, for example, stylometric distance measures?

In order to answer these questions, at this proposal we suggest an approach combining methodologies from Corpus Stylistics (McIntyre & Busse, 2010) and Computational Stylistics or Stylometry (Burrows, 2004; Craig, 2004), which uses the ADSO corpus (Navarro-Colorado, Ribes Lafoz, & Sánchez, 2016) and Herrera’s undoubted and attributed works digitalized by the author (Herrera, 1975) and includes stylometric classification of the literary works of the different periods, contrasting these works using keywords and zeta, as well as seeing where Herrera and $P$ fit in the picture, and the possibility of improving the results and adding more information running the previous tests using POS tags as features.

The keywords will be generated with our software Litcon, whereas the classification and contrasting with zeta are to be run in R through the ‘stylo’ packet (Eder, Rybicki, & Kestemont, 2016). Finally, the POS tagging has been generated through Freeling (Padró, 2012).

The final aim of this study is, firstly, contributing to the characterization of Renaissance against Baroque, secondly, doing a first attempt on assessing the poetic evolution in the Golden Age, and ultimately, analyzing Herrera’s role in this process as well as the one played by his attributed work $P$.

References


**Clémence Jacquot (Montpellier) / Ilaria Vidotto (Grenoble) / Laetitia Gonon (Grenoble)**

*Digital stylistic analyses in “PhraseoRom”: methodological and epistemological issues in a multidisciplinary project*

The ANR-DFG PhraseoRom project (https://phraseorom.univ-grenoble-alpes.fr/?language=en) brings together researchers in linguistics (lexicon, syntax, semantics, linguistics), stylistics and computer sciences to analyze a large annotated corpus of novels (about 2,500 items) from the 20th and 21st centuries in French, English and German.

The tool used to explore this corpus is Lexicoscope (http://phraseotext.u-grenoble3.fr/lexicoscope/, Kraif 2016). It is used to extract recurrent lexicosyntactic constructions (LSC) and allows us to analyze recurrent and significant phraseological motifs. The motif is defined as a “collocational framework” composed of fixed and variable elements which contribute to build the text’s structure and to characterize texts of various genres (Longrée & Mellet, 2013).

Our corpus is divided into 6 sub-corpora which include historical novel, science fiction, fantasy, romance fiction, crime fiction, and ‘general literature’ novel.

Unlike the traditional stylistic approach (author-based analyses or structural approach to describe and question the structure), we base our approach on recent advances in stylometry and the electronically supported extraction of motifs. The aim of the PhraseoRom project is therefore to link the phraseological analysis of a large literary corpus and the stylistic issues concerning its formal and literary implications: does the ‘general literature’ have specific motifs that make it “literary”? If some so-called ‘paraliterary’ productions (Boyer, 2008) share some statistically significant and highly specific LSC, can they be analyzed as a single text?

Our stylistic work therefore needs to redefine the epistemological and methodological assumptions in stylistics, since the methodological approach we used is defined by the large size of the corpus, by the different sub-corpora and the different sub-genres and their comparison and by its multilingual composition.
We assume that the corpus-driven exploration has heuristic potential and we aim to discuss concepts of ‘salience’ (saillance), ‘prominence’ and ‘corpus’ in this specific case. This digital-stylistic approach will find new features, which are revealed by some quantitative criteria.

These criteria must then be included in the continuity of existing stylistic and literary reflections. For example: does the statistical specificity of a motif always imply a stylistic salience? How can we build and use heuristic tools to question literary and stylistic definition of sub-genres and reconsider the controversial distinction between ‘general literature’ and ‘paraliterature’? To what extent do these quantitative features allow us to practically problematize generic hybridity in fiction (a topic that has been already theorized in comparative literature)?

These new kind of heterogeneous corpora change the paradigm of stylistic studies by creating different processes and results (sometimes counterintuitive?). Which interpretative challenges are then created by the evolution from an author-based to a corpus-based stylistic approach? Concrete examples from pilot studies on French corpora, carried out within PhraseoRom, will help us to question more precisely which role digital stylistics plays within multidisciplinary projects in the field of Digital Humanities. Is it a subordinate role? Do digital stylistics and linguistic studies both benefit from their collaboration? Are they involved in a co-creation research work?

References
The complexity of text can be described along many dimensions: Vocabulary, syntax, topics, figure constellations, integrated discourses and much more. The paper will present some such description dimensions of text complexity using examples from the German literature and discuss methods for measuring text complexity using digital texts: How does the measured relate to the respective hermeneutic concept, the dimension of text complexity? So how does our intuition behave, for example, that a 'better' author has a somehow 'larger vocabulary' to what can be grasped with measures of vocabulary richness? A not insignificant problem of this relationship is the existence of a whole wealth of measures to the richness of vocabulary. Some of them correlate strongly, but not all of them.

In a further step we will look at how the measures for these dimensions – vocabulary, syntax, themes, figure constellations, integrated discourses, etc. – relate to stylometrically proven methods. Is complexity a systematic aspect of the author’s style as it is often formally modeled at present, or is it largely independent of it? For this purpose, some of the dimensions of text complexity are to be combined to a multidimensional measure that makes it possible to capture the similarity or distance of the texts from this perspective. How does this measure relate to Delta, whose relatively high robustness for authorship attribution has been proven many times? One may assume that complexity is a more abstract perspective than similarity due to the distribution of words as used by Delta, i.e. texts of two authors can be stylistically very dissimilar, because they prefer different words and sentence forms etc., but can be similar from the point of view of complexity, e.g. because both have only a very limited supply of linguistic patterns. If this is correct, the question arises as to whether the inclusion of text complexity information improves or tends to worsen group attribution due to style.
In a previous study (Salgaro and Rebora 2017), we built a corpus of Italian book reviews. The corpus was divided into three subsets: reviews published on social reading platforms (source: aNobii), in paper magazines (Il Sole 24 Ore), and in scientific journals (Between, Osservatorio critico della germanistica, and OBLIO). All sub-corpora had an approximate size of 650,000 tokens.

With this paper, we use this corpus to answer the following questions: to what extent stylometric methods can be used to classify the texts in the three subsets? How do professional critics, journalists and passionate readers differ in the writing of reviews and what features can be used to identify them?

As demonstrated by (Eder 2015), the first element to influence the quality of a stylometric classification is text length. Preliminary tests – ran with the Stylo R package (Eder et al. 2016) on 5,000-word-long text chunks – showed how Cosine Delta distance, based on just 50 MFW, was able to almost perfectly separate the three subgroups (see Fig. 1). Considered the high variance of text length in our corpus (mean = 259 words; SD = 363 words), we artificially generated a series of sub-corpora composed by text chunks of the same length (varying between 50 and 5,000 words) and we evaluated clustering quality through the adjusted Rand index in the PyDelta Python library. Figure 2 confirms how Cosine Delta distance with 2,000 MFW is the best performing classifier (Evert et al. 2017), but also 200 MFW (i.e., mainly function words) reach a similar – and, in some cases, even better – efficiency. As for text length, clustering quality is quite poor below 1,000 words, while a plateau is reached at about 3,000 words.

To improve the results for shorter chunks, we developed a framework for a machine learning classifier, by operationalizing a series of traditional definitions of literary criticism (e.g. Eco 1979; Gardt 1998; Rodler 2004; Colussi 2017). An extensive lexicon of literary criticism (Beck et al. 2007) was translated into Italian; selections of terms related to mental imagery and emotional aesthetic response were extracted from questionnaires and tools in empirical aesthetics (e.g. Knoop et al. 2016; Fialho et al. in press), translated into Italian, and expanded through the fastText Italian word-embedding model (Joulin et al. 2017). These resources – together with selected features in the LIWC Italian dictionary – were used to measure emotional and cognitive involvement with the reviewed text. The measurements were combined with the results of the stylometric analysis (Cosine Delta, 2,000 MFW) and used to train an SVM classifier.

For a corpus composed by 500-word-long chunks (the length of this abstract), the sole stylometric analysis reached an attribution accuracy of 90.1%, while the SVM classifier scored 93.2%. A slight but promising improvement, if we consider the simplicity of the framework – that can and should be refined further.

With this paper, we hope to have cast the groundwork for a research that might fruitfully combine computational methods and literary theory to study the “style of criticism” of professional and non-professional readers.

References


Note

All scripts and resources used for the analysis are available here:
https://github.com/SimoneRebora/style_of_criticism

Images

Figure 1. Network graph of the corpus (Cosine Delta, 50 MFW, ForceAtlas2 in Gephi). In green: aNobii; in violet: Sole 24 Ore; in red: Between, Osservatorio, and OBLJO.
Cross-language stylometry: Picasso’s writings in Spanish and French.

The example of one of the most known artists who wrote half of his experimental literary 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 the multilingual text-collection is to separate the languages in the collection and compare them to 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; Eder 2011) – very important to know before using stylometry with the same parameters (Deltas) on corpora in different languages, but without treating the problem of stylistics for people writing in two (or more) languages. Juola and Mikros started to think about “cross-linguistic stylistic features” (2016) and analyzed a corpus of tweets by bilingual twitter users (Spanish and English). The results, though, cannot easily be adopted on the question of literary style: length of tweet and word-length suggest here “cross-linguistic similarities” (Juola, Mikros 2016). The similarities in French and Spanish texts by Picasso might be due to different style-markers.

Since Picasso started to publish part of his writings (plays) in 1940ies, people tried to catalogize or to cluster them into literary history. 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 1991; Goddard 2006; Michaël 2008; Rißler-Pipka 2015). Several papers are published about stylometry on James Joyce (O’Sullivan, Eder, Rybicki 2018; O’Sullivan 2014; Clement 2013), but fewer about the authors of Romance Avant-garde (Calvo Tello 2018). Using close-reading as a method of stylistics it is possible to compare texts by examples of phrases, motives, etc., but digital stylistics needs 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 ignore canon and to compare in a quantitative way. Previous work (Rißler-Pipka 2018a+b) showed rough stylistic resemblance between the French writings of Picasso and Raymond Roussel (“Nouvelles impressions d’Afrique” (1901)) and Ramón Gomez de la Serna for the Spanish writings (both authors would not necessarily have been on the list of candidates for influences before the quantitative analysis).

Fig. 5 (in Rißler-Pipka 2018a): Dendrogram for the Picasso-corpus (French), generated with style (Eder, Rybicki, Kestemont 2013)
Fig. 11 in (Rißler-Pipka 2018b): Dendrogram for the Picasso-corpus (Spanish), generated with stylo (Eder, Rybicki, Kestemont 2013)

For deeper cross-language stylometry it is important to ask what are the reasons and what are the style-markers which caused the clustering. Therefore the comparative analysis of both corpora (Spanish and French) will be expanded (for the Spanish collection) and concentrated on the most likely candidates (trying rolling delta for Picasso and Roussel; and for Picasso and Gomez de la Serna). For a more closer look the hypotheses coming from close-reading analysis (Picasso has a stable and small vocabulary which can be captured in topics and he plays a game of word recombination) can be proved wrong or right by using digital stylistic methods (richness of vocabulary, word density, topic modeling, etc.). The results of these analyses for each language should be compared to see if the author is using the same strategy in both languages. Finally, a list of style markers can be tested on the each corpus, keeping in mind that both languages, Spanish and French are related, but completely different in verb phrases (hago = je fais = I do) and both languages give different opportunities to construct ambivalence. Without discussing the differences in languages from a linguistic point of view, cross-language stylistics is a powerful method to improve literary analysis (for bilingual authors, intertextual relations in Romance studies and shifts in style due to new cultural and language backgrounds).
Jan Rohden (Göttingen)

*Digital approaches to poetic style: a quantitative stylistic analysis of Italian Petrarchism*

Petrarch (1304-1374) can be considered one of the most influential poets of European literature. One of the main reasons for this is his well-received collection of Italian love poems known as *Canzoniere*. For centuries Petrarch’s collection was a role model for many European poets, whose works tried to imitate Petrarch’s poetic style.

In several different studies with various approaches, literary research has acknowledged Petrarch’s influence on the literary style of later poetry. Scholars even created a term to describe this phenomenon: Petrarchism. Although stylistic features similar to Petrarch’s poetic style can be found in the works of a large number of authors from different centuries, most of the research literature so far usually focused on fairly small numbers of texts written by either one author or a small group of writers. Thus, our knowledge of Petrarch’s influence on the poetic style of his successors is mainly deduced from different individual analyses of quite few works. Despite the value of these readings, it often remains difficult to distinguish whether specific poetic features associated with Petrarchism are in fact the effect of a general change in poetic style due to Petrarch’s reception, or rather merely the result of individual stylistic choices of certain authors.
Thanks to digital techniques in the past years new methods have been developed, which allow to compare and analyse large collections of literary texts, making it much easier to approach the question of style from a quantitative perspective. Because of that, literary influences, which in the past were mostly deduced from individual interpretations of few texts, can nowadays be analysed on the basis of large text corpora. Such techniques seem helpful to bring out stylistic features common to a large number of texts, regardless of the individual stylistic choices, writing styles or poetic techniques used by single authors.

The proposed presentation poses the question in what way quantitative stylistic approaches can help to analyse Petrarch’s influence on the poetic style of his Italian successors. A contrastive stylometric analysis of a corpus of Italian poetic collections shall examine whether there are stylistic elements common for 16th century Italian poetry, which indicate a change in poetic style, that can be traced back to Petrarch’s reception.

References
Petrarchism

Stylometry

*****
Background

When working with literary texts, one relevant problem for linguistic and literary scholars and in general for machine based text understanding is the classification of text types. The term “text type” refers to a variety of different phenomena reaching from a superordinate view of genre (Chatman 1990: 10) to functionally motivated text types as aggregation of structural or linguistic features (Biber 1988, 1989). While the taxonomy, textual layer and functionality of different theories behind text types may differ widely, a common concept behind text types is the understanding of textual surface structure varying in their respective type of discourse (Fludernik 2000). The text types “description”, “narrative” and “argumentative” emerge very frequently in many theories (Werlich 1975, Adam 1985, Chatman 1990). As we are interested in sentence- or paragraph-wise classification of text types rather than a holistic view on the text and has a strong focus on literary works, this scheme fits our purposes of automatically classifying text types using machine learning. In this work we describe the preparation of a text type dataset and present a bootstrapping based approach for rare classes to obtain a more balanced dataset.

Annotation Scheme

Functionalizing the existing theories to a respective abstraction of surface phenomena, we came to our annotation guidelines:

- descriptive: the description of a physical object (item, landscape, building, person, ...) in its dimensions, parts, and/or properties
- narrative: the representation of a chain of events, actions or activities in its temporal progress driven by persons or other “actors”
- argumentative: the presentation, explanation and justification of an abstract idea (thought, argument, conviction, ...) in its logical context

To consider possible shortcomings of these guidelines or the underlying model and to give the annotators an opportunity to indicate their uncertainty, we introduced additionally the label “unknown”.

In our annotation tool (see figure 1), sentences were presented within a small contextual framing. The annotators were asked to choose one category, if possibly but they had the option to choose more than one category if it was necessary.
Bootstrapping samples for annotation
We chose a subset of the Kallimachos corpus (426 German novels) for our first experiments. In order to bootstrap the presumably most underrepresented text type (descriptive) we annotated a list of frequent adjectives for their potential use in descriptive passages with a score from 0 to 9 and derived an “descriptiveness” score. We divided each text in 100 segments and selected the most and less descriptive two segments by selecting the segments with the maximum resp. minimum score. Figure 2 shows the benefit of bootstrapping descriptive passages: By adjusting the proportion of “min” and “max” when sampling instances, we can balance the labels to an uniform distribution. For the manual gold standard annotation, we choose 30 of these segments randomly. Each sentence was annotated by 3 annotators to judge the complexity of the annotation task, and to find a subset of very reliable annotations when all annotators agree (see figure 3). In general the annotators agreed up to a multi- label Fleiss-Kappa inter-annotator agreement of 0.385 for 1803 annotated sentences (Fleiss 1971). Demanding full agreement the dataset reduces the number of instances to 830 sentences (218 descriptive, 352 narrative, 260 argumentative). Majority vote (i.e. the consent of two of three annotators) leads to 1657 labeled instances. Therefore a consolidating annotation run will reliably increase the number of labels by large margin. For future work we will use this dataset for classifying text types.
Figure 2: Proportion of “descriptive” labels for minimal resp. maximal descriptiveness-scored texts

Figure 3: Number of instances annotated with label choice [D, N, A, U] e.g. for [0,0,2,1], 177 instances were labeled argumentative by two annotators and unknown by one annotator.

References
Julian Schröter (Würzburg)

*The challenge of exploring the style of the German Novelle as a virtually orderless genre.*

Until the 1960s, the German *Novelle* was considered the central genre within the literary field of nineteenth century German literature. Newer investigations have shown by contrast that the traditional idea of the *Novelle* as a coherent and strict genre is mistaken (Pabst (1967), Polheim (1965), Schröder (1970), Meyer (1987)). With regard to regular patterns there is hardly such a thing as a genre of the *Novelle*. Since the category of *Novellen* is not expected to render a class of *Novellen* but a structureless group with exemplars of a certain genre name Delta-based cluster analysis (in the tradition of Burrows (2002)) is not expected to succeed in clustering *Novellen* vs. Non-*Novellen*. Nevertheless authors and readers in the century had pronounced categorical and aesthetic expectations of *Novellen*. According to Walton (1970) I assume, that it is aesthetically relevant to see works of arts as actualizations of certain categories. My aim is to find regionally limited and group-specific textual patterns that generate the relevant categorical expectations of the authors and their audiences.

The line of my argument consists of five parts: First, I shall elucidate my application of Walton’s concept of aesthetic categories. Second, I shall justify my decision to label a text as *Novelle*. Third, I am going to explore the measure of stylistical disorder in the group of Novellen. My explorations are based on unknown journals of the early 19. century (*Taschenbuch Urania*, *Taschenbuch Aglaja*, et al.) which are currently recorded with OCR-recognition by student assistants in the context of a more extensive research project. Thus my results are based on widely parts of the literary field that are widely unknown so far. Fourth, I shall try to customize a very inventive strategy based on supervised learning, which has been developed by Underwood and Sellers (2015) to train a model on texts relative to their genre labels and to narrow time periods (range of 5 years) in order to control the degree of correspondence between the labeling of my model and of the actors within the literary field relative to adjacent periods. This trained model is expected to take decisions which do correspond only in parts to the actual labeling within the literary field. The degree of coincidence of the predictions generated by the model and the actual acts of genre labeling within the literary field may be used as a measure of the relative convergence or divergence of the historical semantics of the genre label. Fifth, this idea of a measure of semantic convergence in historical literary practice shall be evaluated. Sixth and finally, the implications of my results for constructing the concept of genre in Digital Humanities are discussed on a conceptional level.
References

Arjuna Tuzzi (Padova) / George Mikros (Athens) / Michele A. Cortelazzo (Padova)
Applying General Imposters’ method to the Ferrante’s case

Elena Ferrante is the *nome de plume* of an anonymous writer that is highly successful on the international stage and which success far exceeds that of other products of Italian contemporary literature.

From the analyses of a large corpus of 150 novels published in the last 30 years, written by 40 different Italian authors, an international group of scholars (Cortelazzo, Nadalutti, Ondelli, & Tuzzi, 2018; Cortelazzo & Tuzzi, 2017; Eder, 2018; Juola, 2018; Lali, Tria, & Loreto, 2018; Mikros, 2018; Mikros George, 2017; Ratinaud, 2018; Rybicki, 2018; Savoy, 2018a, 2018b; Tuzzi & Cortelazzo, 2018a, 2018b, 2018c) showed that Ferrante reveals traits of originality in both style and content and, amongst the authors included, Domenico Starnone shows the strongest similarities with Ferrante. A further study based on profiling procedure founded on machine learning (ML) and support vector machine (SVM) took into account a new focused and more circumscribed corpus (Cortelazzo, Mikros, & Tuzzi, 2018) that included 86 texts written by a set of candidates who are not strictly novelists and 27 texts of Ferrante’s *La frantumaglia* (Ferrante, 2016). Results mainly pointed to Domenico Starnone, Anita Raja and the members of the staff of the E/O publishing house and revealed the potential existence of different hands.

This study aims to test a relatively new stylometric approach to authorship verification problems, the General Imposters (GI) method (Kestemont, Stover, Koppel, Karsdorp, & Daemelans, 2016; Koppel & Winter, 2014) as it has been implemented in the Stylo package in R (Eder, Rybicki, & Kestemont, 2016). The GI method can be briefly described as 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 “imposter” documents which were not written by the original author. The score calculated represents not only how different is the anonymous text from the other texts of the candidates but also how consistent are the stylistic differences between them.

The Ferrante’s case could be approached as a verification problem since we are not sure whether the real author behind Ferrante’s pseudonym is among the one we have in our corpus. For this
reason, we applied the GI method using the Würzburg distance, originally in the corpus of 150 novels and we verified that Starnone is the most probable author of Ferrante’s books, a result which is in accordance with all stylometric analyses applied to this corpus. We then proceeded to the non-literary corpus and we tested GI method using Ferrante’s interviews, letters and other non-literary texts as the target of our authorship verification process. The results were quite different compared to the corpus of novels since now Starnone was not the only possible author. In many of these texts, both Raja, Martone, Ozzola as well as the E/O staff seem to have authorial contributions.

The GI method not only confirmed previous results but also improved our knowledge on this case since it provides a measure of the attribution strength.

References


Writers in the period of the Enlightenment made frequent use of a repertoire of rhetorical devices, informed by the classical learning and widely-read handbooks. Various research questions suggest themselves relating to the distribution of rhetorical figures among writers in historical periods, and the functions and meanings of these stylistic features, which could usefully be addressed with digital methods. This paper relates to an initial phase of this project, in which the various opportunities are evaluated, and technical barriers are identified and addressed.

In order to be able to identify and categorize rhetorical figures, it is first necessary to have a usable list of such stylistic features. Starting with Lanham (1991), a list was made of rhetorical figures which appeared likely to be susceptible to recognition on the basis of lexis or syntax. An example of a tractable feature is alliteration - a search for repeated initial letters can find examples:

“[..] j’aspire avec ardeur au moment d’être admis à vôtre audience [...]”

Jean Jacques Rousseau to Jean Gabriel La Porte Du Theil, Wednesday, 7 October 1744

For this investigation, texts from the Electronic Enlightenment (EE) in French were used. EE is a major collection of letters focussing on the eighteenth century. The interface for EE does not easily allow complex linguistic queries, so the texts were extracted and explored initially with corpus analysis tools including CQPweb, DiaCollo (an interface allowing visualization of collocations over time), and Antconc. Preliminary results were suggestive that there was likely to be a wealth of relevant material, but that further investigation was hampered by the lack of word-class tagging and lemmatization, further complicated by a high level of orthographic and lexico-grammatical variation in the texts, for example variation in spelling across different authors and time periods.

Experiments with TreeTagger have confirmed that there is, not surprisingly, a need for enhanced lexical resources to deal with the vocabulary found in the Electronic Enlightenment collections. One of the goals of the project is to create, or contribute to the creation of, high-quality, standards-conformant, and re-usable lexical resources, so that other projects can annotate and explore other texts.
Variation in the language of the corpus compared to modern-day French includes differences in lexico-grammar, orthography, tokenization, and capitalization. Other challenges include extended quotations in classical languages without glosses or translations. The texts are hand-written correspondence, differing from the standard language as found in contemporary published works, with elision of accents, switching between languages, spelling mistakes, and unclear passages. Some variation is due to writing by non-native speakers, and phonetic writing by partially educated writers (often women), e.g.:

“Etant da cor je me rendis avecque mon nésessere che luy je presentay mon premier cartieé de pension”

The methodology then of the project then has been to tag all of a corpus of historical French correspondence with a wordclass tagger trained on modern French, then to identify the most commonly ‘unknown’ words and where then added to a custom lexicon, and the corpus was tagged again using the custom lexicon, in order to achieve higher levels of recall and precision in the annotation. Ongoing work will continue to improve the tagging accuracy iteratively, produce a re-usable lexicon, and to investigate further the distribution and use of rhetorical figures in the corpus. Once it is possible to search the corpus for lemmas and morpho-syntactic tags, a much wider range of rhetorical figures can be searched for. A range of results from this ongoing, iterative work are available.

References
Bus time tables

The conference venue is located on the Hubland campus of the University which is about 10-15 mins away from the city center. Therefore, you will find here the bus times tables to get to campus and to get back to the city center. The closest bus stops to the conference hotels are “Barbarossaplatz” (for Hotel Strauss and Würzburger Hof) and “Mainfranken Theather” (for Hotel Amberger). If you present a talk during the Conference, we will provide you with bus tickets during the time of the Conference.

To get to campus ...

**Line 14: Würzburg – Gebrunn**

(get off the bus at “Am Hubland”)

Monday-Friday

![Bus schedule for Line 14 on Monday and Tuesday](Image)

Saturday

![Bus schedule for Line 14 on Saturday](Image)
Line 29: Busbahnhof – Hubland/Campus Nord
(get off the bus at “Philosophisches Institut”)

Monday – Friday

Saturday
Line 114: Busbahnhof – Fachhochschule  
(get off the bus at “Philosophisches Institut”)

Monday – Friday

<table>
<thead>
<tr>
<th>Verkehrsinweis</th>
<th>E</th>
<th>E</th>
<th>Montags-Freitags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Busbahnhof</td>
<td>ab</td>
<td>07.25</td>
<td>07.25</td>
</tr>
<tr>
<td>Rennweg</td>
<td></td>
<td>07.28</td>
<td>07.28</td>
</tr>
<tr>
<td>Letzter Hieb</td>
<td></td>
<td>07.31</td>
<td>07.31</td>
</tr>
<tr>
<td>Letzter Hieb</td>
<td></td>
<td>07.40</td>
<td>07.40</td>
</tr>
<tr>
<td>Pestalozzi-Straße</td>
<td></td>
<td>07.40</td>
<td>07.40</td>
</tr>
<tr>
<td>Pestalozzi-Straße</td>
<td></td>
<td>07.41</td>
<td>07.41</td>
</tr>
<tr>
<td>Universität Sportzentrum</td>
<td>an</td>
<td>07.44</td>
<td>07.46</td>
</tr>
<tr>
<td>FWRG (Sander Ehemalige)</td>
<td>an</td>
<td>07.46</td>
<td>07.49</td>
</tr>
</tbody>
</table>

Saturday

The line 114 does not run on Saturdays.

Line 214: Busbahnhof – Fachhochschule

Monday – Friday

<table>
<thead>
<tr>
<th>Verkehrsinweis</th>
<th>E</th>
<th>E</th>
<th>Montags-Freitags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Busbahnhof</td>
<td>ab</td>
<td>07.08</td>
<td>07.25</td>
</tr>
<tr>
<td>Rennweg</td>
<td></td>
<td>07.08</td>
<td>07.28</td>
</tr>
<tr>
<td>Letzter Hieb</td>
<td></td>
<td>07.12</td>
<td>07.31</td>
</tr>
<tr>
<td>Universität Sportzentrum</td>
<td>an</td>
<td>07.27</td>
<td>07.46</td>
</tr>
</tbody>
</table>

Saturday

The line 214 does not run on Saturdays.
To get to the city center ...

**Line 14: Gebrunn – Würzburg**
(get on the bus at “Am Hubland”)

**Monday – Friday**

<table>
<thead>
<tr>
<th>Time</th>
<th>Gebrunn</th>
<th>Rödigg</th>
<th>Würzburg, Mathem. Institut</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:20</td>
<td>18:00</td>
<td>17:00</td>
<td>14:51</td>
</tr>
<tr>
<td>11:20</td>
<td>18:00</td>
<td>17:00</td>
<td>14:51</td>
</tr>
<tr>
<td>11:20</td>
<td>18:00</td>
<td>17:00</td>
<td>14:51</td>
</tr>
</tbody>
</table>

**Saturday**

<table>
<thead>
<tr>
<th>Time</th>
<th>Gebrunn</th>
<th>Rödigg</th>
<th>Würzburg, Mathem. Institut</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:20</td>
<td>18:00</td>
<td>17:00</td>
<td>14:51</td>
</tr>
<tr>
<td>11:20</td>
<td>18:00</td>
<td>17:00</td>
<td>14:51</td>
</tr>
<tr>
<td>11:20</td>
<td>18:00</td>
<td>17:00</td>
<td>14:51</td>
</tr>
</tbody>
</table>

---

To get to the city center ...

**Line 14: Gebrunn – Würzburg**
(get on the bus at “Am Hubland”)

**Monday – Friday**

<table>
<thead>
<tr>
<th>Time</th>
<th>Gebrunn</th>
<th>Rödigg</th>
<th>Würzburg, Mathem. Institut</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:20</td>
<td>18:00</td>
<td>17:00</td>
<td>14:51</td>
</tr>
<tr>
<td>11:20</td>
<td>18:00</td>
<td>17:00</td>
<td>14:51</td>
</tr>
<tr>
<td>11:20</td>
<td>18:00</td>
<td>17:00</td>
<td>14:51</td>
</tr>
</tbody>
</table>

**Saturday**

<table>
<thead>
<tr>
<th>Time</th>
<th>Gebrunn</th>
<th>Rödigg</th>
<th>Würzburg, Mathem. Institut</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:20</td>
<td>18:00</td>
<td>17:00</td>
<td>14:51</td>
</tr>
<tr>
<td>11:20</td>
<td>18:00</td>
<td>17:00</td>
<td>14:51</td>
</tr>
<tr>
<td>11:20</td>
<td>18:00</td>
<td>17:00</td>
<td>14:51</td>
</tr>
</tbody>
</table>

---

57
**Line 29: Hubland/Campus Nord – Busbahnhof**

*(get on the bus at “Philosophisches Institut”)*

**Monday – Friday**

<table>
<thead>
<tr>
<th>Line: 29</th>
<th>Hubland/Campus Nord – Busbahnhof</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(get on the bus at “Philosophisches Institut”)</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Sat**

<table>
<thead>
<tr>
<th>Line: 29</th>
<th>Hubland/Campus Nord – Busbahnhof</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(get on the bus at “Philosophisches Institut”)</strong></td>
<td></td>
</tr>
</tbody>
</table>
Line 114: Fachhochschule – Busbahnhof
(get on the bus at “Philosophisches Institut”)

Monday – Friday

<table>
<thead>
<tr>
<th>Linie</th>
<th>Montags- freitags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verkehrshinweis</td>
<td></td>
</tr>
<tr>
<td>Universität Sportzentrum</td>
<td>ab 16:34 16:45 17:01 17:16 17:31 17:46 18:01 18:16 18:31 18:46 19:01 19:16</td>
</tr>
<tr>
<td>Wittelsbacherstraße</td>
<td>00:22 00:37 00:52 01:07 01:22 01:37 01:52 02:07 02:22 02:37 02:52 03:07</td>
</tr>
<tr>
<td>Wittelshainplatz</td>
<td>01:22 01:37 01:52 02:07 02:22 02:37 02:52 03:07 03:22 03:37 03:52 04:07</td>
</tr>
<tr>
<td>Großerbruderstraße</td>
<td>02:22 02:37 02:52 03:07 03:22 03:37 03:52 04:07 04:22 04:37 04:52 05:07</td>
</tr>
<tr>
<td>Valentins-Becker- Straße</td>
<td>03:22 03:37 03:52 04:07 04:22 04:37 04:52 05:07 05:22 05:37 05:52 06:07</td>
</tr>
<tr>
<td>Rennweg</td>
<td>04:22 04:37 04:52 05:07 05:22 05:37 05:52 06:07 06:22 06:37 06:52 07:07</td>
</tr>
<tr>
<td>Mandarinien Theater</td>
<td>05:22 05:37 05:52 06:07 06:22 06:37 06:52 07:07 07:22 07:37 07:52 08:07</td>
</tr>
<tr>
<td>Stett Haag</td>
<td>06:22 06:37 06:52 07:07 07:22 07:37 07:52 08:07 08:22 08:37 08:52 09:07</td>
</tr>
<tr>
<td>Busbahnhof</td>
<td>07:22 07:37 07:52 08:07 08:22 08:37 08:52 09:07 09:22 09:37 09:52 10:07</td>
</tr>
</tbody>
</table>

Saturday

The line 114 does not run on Saturdays.

Line 214: Fachhochschule – Busbahnhof
(get on the bus at “Philosophisches Institut”)

Monday – Friday

<table>
<thead>
<tr>
<th>Linie</th>
<th>Montags- freitags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verkehrshinweis</td>
<td></td>
</tr>
<tr>
<td>FHNW/Sandern leitortweg</td>
<td>ab 16:34 16:45 17:01 17:16 17:31 17:46 18:01 18:16 18:31 18:46 19:01 19:16</td>
</tr>
<tr>
<td>Zollhaus Göglingenberg</td>
<td>23:22 23:37 23:52 24:07 24:22 24:37 24:52 00:07 00:22 00:37 00:52 01:07</td>
</tr>
<tr>
<td>Wittelsbacherstraße</td>
<td>00:22 00:37 00:52 01:07 01:22 01:37 01:52 02:07 02:22 02:37 02:52 03:07</td>
</tr>
<tr>
<td>Wittelshainplatz</td>
<td>01:22 01:37 01:52 02:07 02:22 02:37 02:52 03:07 03:22 03:37 03:52 04:07</td>
</tr>
<tr>
<td>Großerbruderstraße</td>
<td>02:22 02:37 02:52 03:07 03:22 03:37 03:52 04:07 04:22 04:37 04:52 05:07</td>
</tr>
<tr>
<td>Valentins-Becker- Straße</td>
<td>03:22 03:37 03:52 04:07 04:22 04:37 04:52 05:07 05:22 05:37 05:52 06:07</td>
</tr>
<tr>
<td>Rennweg</td>
<td>04:22 04:37 04:52 05:07 05:22 05:37 05:52 06:07 06:22 06:37 06:52 07:07</td>
</tr>
<tr>
<td>Mandarinien Theater</td>
<td>05:22 05:37 05:52 06:07 06:22 06:37 06:52 07:07 07:22 07:37 07:52 08:07</td>
</tr>
<tr>
<td>Stett Haag</td>
<td>06:22 06:37 06:52 07:07 07:22 07:37 07:52 08:07 08:22 08:37 08:52 09:07</td>
</tr>
<tr>
<td>Busbahnhof</td>
<td>07:22 07:37 07:52 08:07 08:22 08:37 08:52 09:07 09:22 09:37 09:52 10:07</td>
</tr>
</tbody>
</table>

Saturday

The line 214 does not run on Saturdays.
Additional information

- You can access the wireless internet connection via the networks “Bayern W-Lan” (open network) or “eduroam”.

- Lunch-break: there is a refectory (“Mensateria”) right next to the conference building where you can also pay in cash. A Greek and an Italian restaurant are 10 minutes away by foot. Contact us, if you want to have more detailed information.

- Parking spots are available close to the Conference venue.

- For more information on the city of Würzburg and its history and sights, please have a look at the following websites:

  https://en.wikipedia.org/wiki/W%C3%Bcrzburg

We are grateful for the support of …
Contact:

Organizing committee
“Digital Stylistics in Romance Studies and Beyond”
Lehrstuhl für Computerphilologie
Universität Würzburg
Am Hubland
97074 Würzburg
Germany
digitalstylistics@uni-wuerzburg.de