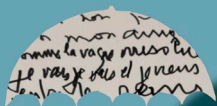


DE GRUYTER

COMPUTATIONAL STYLISTICS IN POETRY, PROSE, AND DRAMA

*Edited by Anne-Sophie Bories, Petr Plechac,
and Pablo Ruiz Fabo*



DE
G

Computational Stylistics in Poetry, Prose, and Drama

Computational Stylistics in Poetry, Prose, and Drama



Edited by
Anne-Sophie Bories, Petr Plecháč,
and Pablo Ruiz Fabo

DE GRUYTER

The open access publication of this book has been published with the support of the Swiss National Science Foundation.

ISBN 978-3-11-078141-0

e-ISBN (PDF) 978-3-11-078150-2

e-ISBN (EPUB) 978-3-11-078156-4

DOI <https://doi.org/10.1515/9783110781502>



This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. For details go to <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

Creative Commons license terms for re-use do not apply to any content (such as graphs, figures, photos, excerpts, etc.) not original to the Open Access publication and further permission may be required from the rights holder. The obligation to research and clear permission lies solely with the party re-using the material.

Library of Congress Control Number: 2022944811

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the internet at <http://dnb.dnb.de>.

© 2022 the author(s), editing © 2022 Anne-Sophie Bories, Petr Plecháč, and Pablo Ruiz Fabo, published by Walter de Gruyter GmbH, Berlin/Boston

The book is published open access at www.degruyter.com

Cover image: Anne-Sophie Bories

Typesetting: Integra Software Services Pvt. Ltd.

Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Contents

About this Volume — VII

Anne-Sophie Bories, Pablo Ruiz Fabo, and Petr Plecháč

The Polite Revolution of Computational Literary Studies — 1

Anne Bandry-Scubbi

Zooming In, Zooming Out: 30 Years of Corpus Stylistics Bricolage — 19

Jan Christoph Meister

Poetry, Phenomenon and Phenomenology — 37

Helena Bermúdez Sabel, Pablo Ruiz Fabo,
and Clara Martínez Cantón

DISCOVERING Spanish Sonnets: A Circular Reading Experience — 67

Chris Mustazza

**In Search of the Sermonic: Machine Listening and Poetic Sonic
Genre — 87**

Éliane Delente

**Can Relationships between Rhythm and Meaning in French Versified
Poetry be Automated? — 99**

Natalie M. Houston

Rhyme Frequency in Nineteenth-Century English Poetry — 117

Jonathan Armoza

Hayford's Duplicates: Cobbling a Model of Melville's *Moby-Dick* — 133

Georgy Vekshin, Egor Maximov, and Marina Lemesheva

**Poeticisms and Common Poetic Discourse in the Digital *Russian Live
Stylistic Dictionary* — 153**

Benjamin Krautter, Janis Pagel, Nils Reiter, and Marcus Willand

**Properties of Dramatic Characters: Automatically Detecting Gender, Age,
and Social Status — 179**

Pablo Gervás

**N-Gram-Driven Word Level Recombination: Exploring a Search Space of
Metrically Valid Verse — 203**

Anne-Sophie Bories, Pablo Ruiz Fabo, and Petr Plecháč

Closing Remarks: What Was This All About? — 223

About the Editors — 233

About this Volume

These collected articles stem from a discussion that began at the third annual international Plotting Poetry conference held in Nancy (France) in September 2019, organized by Anne-Sophie Bories (University of Basel) and Véronique Montémont (ATILF). The group Plotting Poetry brings together literary scholars from all language areas who share an interest in the development of computational and statistical apparatuses to describe and analyze meter, style, and poeticity. For its third instalment, Plotting Poetry focused on questions of poetics, inviting participants to present feedback – positive or not – on the computational and statistical tools they had developed to address issues of poetics, metrics, and stylistics, and to shed light on the fields of literature, linguistics, and literary history. The fruitful exchanges begun during this gathering became the starting point for an ongoing discussion that ultimately formed the material of this collaborative volume.

Anne-Sophie Bories, Pablo Ruiz Fabo, and Petr Plecháč

The Polite Revolution of Computational Literary Studies

Abstract: The progressive digitization of texts, be they literary or not, has had a remarkable impact on the way we access them, making it possible to obtain help from computers towards the analysis of literary works. Treating text as data allows researchers to test existing hypotheses and, sometimes, ask new questions. And yet, what might appear like a scholarly revolution is actually the natural continuation of former efforts. From card files to spreadsheets to deep learning, quantitative approaches are not a disruption of scholarly practices, particularly in the exploration of poetic texts, which typically rely on a highly regulated, and thus readily measurable, material. Whatever the complexity of the method used, viewing texts through this concentrating lens, this restricted gaze, forms a camera obscura within which lines of regularity may appear that we hadn't necessarily thought of beforehand, enriching and furthering the scholarly examination of literary productions.

It is difficult to ignore the growing impact of digitalization on the way we access texts. As we open one device to read an article, look up a word, locate a reference, or share a draft; as our computers overflow with documents; as our online activities are scrutinized and monetized, it is all too obvious that we are living in a world where texts have become data. The emergence of digital corpora and the democratization of information technologies are fundamentally changing and expanding the ways in which we apprehend texts, including literary works.

Simply put, computational literary studies is just what it sounds like: obtaining varying degrees of assistance from computers to analyze literary works. And so, over the last few decades (Jockers, 2013; McCarty, 2005; Moretti, 2000, 2007, 2013; Ramsay, 2011), a trend has developed, an opportunity for literary scholars – alongside other humanities scholars – to adopt quantitative and computational methods, and to explore new and different terrain for analyzing texts, testing

Anne-Sophie Bories, University of Basel, Department of Languages and Literatures, French Seminar, email: a.bories@unibas.ch

Pablo Ruiz Fabo, University of Strasbourg, LiLPa UR 1339, Linguistics, Language, and Speech, email: ruizfabo@unistra.fr

Petr Plecháč, Institute of Czech Literature, Czech Academy of Sciences, email: plechac@ucl.cas.cz

hypotheses, and identifying authors, among other evolutions. Its popularity did not spring up out of thin air; it follows on from a longer history of scholars patiently gathering quantitative evidence for literary and linguistic analysis. Here, one could cite Busa (1951, 1980), Yule (1944), or the Francophone tradition of lexical statistics recently reviewed in Lebart et al. (2019: 8–16).

The level of technicity ranges from compiling remarks in a spreadsheet to advanced machine learning and statistical models. The multiplicity of research goals spans from traditional text interpretation to forensic authorship attribution. Some teams build elaborate devices while others reuse them for different purposes, and most put together a patchwork of tools to fit their particular needs. This transition has received diverse responses from established communities of literary scholars, sometimes enthusiastic, sometimes distrustful. However, in the subgroup of poetry scholars, whose material is quantitative in nature, the overall reaction has been one of welcome. For people preoccupied with counting syllables and classifying rhymes, technology that handles counts and tokens better and faster is a particularly good fit.

This edited volume approaches poetry, narrative, and drama from the perspective of computational stylistics, responding to the current interest in computational and statistical methods to describe and analyze meter, style, and poeticity, particularly to the extent that these methods can open up new research perspectives in literature, linguistics, and literary history. It is not an exhaustive survey of methods and results – which would be a task well beyond our capabilities – but rather presents a balanced selection of salient aspects to draw a picture of both the diversity and the relevance of a highly dynamic field.

Here, and with the objective of giving an overview of some of the significant methods, goals, and pitfalls within the emerging field of computational literary studies, we argue for an inclusive understanding of these novel approaches as a continuation of and a collaboration with more traditional *close reading* techniques.

1 Text representations

The first advantage of digitalization is the possibility it provides to split a corpus, to reduce it into smaller units, be they sentences, phrases, syllables, or even phonological features, for instance. Although the field has largely grown beyond the early task of counting and finding words, the focus on a very fine granularity remains a central, and to a large extent defining, feature of computational literary studies. The many expressions of this principle include straightforward full-text searches, which are useful for navigating texts, locating specific passages, and

mapping lexical rarities and recurrences; analyzing most frequent words (MFW), typically by putting small grammatical words in the spotlight; elaborating on topic models, which group texts based on their shared lexical priorities; and elaborating on word vectors, which embody the coordinates for an expression within a multidimensional semantic space, inferred through word embeddings.

Viewing any text as a boxful of words to be numbered, counted, and systematized narrows our gaze and temporarily diverts it from a literary work's meaning and significance. This restricted gaze, this limited yet overarching view, forms a camera obscura within which lines of regularity appear. These can be studied to reveal an author's unique quirks and the ways in which lexical domains are grouped, or to map kinships between different works. Shifting our attention away from the meaning and significance of texts is but a temporary means to gain new insights, and not a permanent extraction.

2 Objectives

The hermeneutical objectives of computational literary scholars are at least as diverse as those of traditional literary studies scholars. Just as the pool of technological developments is quickly expanding, so too is the list of possible uses. Rather than try to provide an exhaustive inventory of paths and trends, we would like to mention a few prominent approaches and orientations, to catch a glimpse of the extraordinarily wide array of specific processes and results that are branching out in unforeseen directions and combining to form new protocols. Digital humanities conferences and publications make a habit of the unexpected, with scholars finding new, ingenious ways to explore datasets, collect overlooked data, undertake new research endeavours or perform seemingly impossible tasks. As computational approaches gain momentum, they are being applied to a number of very different uses, such as authorship attribution, historical phonetics, literary history and canon definition, textual analysis, and more.

Stylometry, which routinely focuses on aspects of a text to which a typical reader would pay little attention, is often preoccupied with authorship attribution, including as forensic evidence in court. It appears that what makes an individual style unique lies less in the choice of a discourse or genre, or in the associated content words, than it does in a writer's minute habits, such as the way that he or she constructs phrases, distributes short grammatical words, or chooses punctuation. The patterns in the variations of these apparently less

meaningful aspects of language can be used to identify authors, trends, sub-genres, etc., including in non-literary contexts (Jockers, 2014; Karsdorp, Kestemont, Riddell, 2021; Savoy, 2020).

Stylometry focuses so frequently on authorship attribution that the two are sometimes equated with one another, consistent with a definition of *style* that is far removed from that of stylistics and instead tackling a matter that foils habitual literary analysis (see Herrmann, van Dalen-Oskam, Schöch, 2015). However, a more conventional understanding of style, as it is commonly explored by stylistics, can also be handled by means of computational approaches and indeed forms another of the community's recurring hermeneutic goals, sometimes referred to as "computational stylistics." To locate and compile figures and tropes, we can sometimes look for their linguistic imprint, as is done, for instance, to identify direct speech (Brunner et al., 2020; Byszuk et al., 2020; Schöch et al., 2016). And although quantitative analysis, by breaking texts down to sums of measurable elements, at first appears to be blind to the interpretative meaning of literary works, a number of paths do lead to insights about what texts actually mean, including, for instance, sentiment analysis (Kim, 2020; Klinger et al., 2020).

The prospects of advancing literary history and literary theory are improved greatly both by approaching texts as boxfuls of words to create language models and by approaching literary movements as boxfuls of books to produce literary evolution models, thus following in the footsteps of Franco Moretti by looking at more than just the works that have kept being read.

The statistical strategies of computational literary studies are equally wide-ranging, from descriptive statistics, which examines the characteristics of a dataset, through inferential statistics, in which these characteristics are tentatively generalized into a broader, virtual dataset, all the way to advanced machine learning, including topic modeling and deep learning, among others. As one example, topic modeling has been applied to poetry (e.g., in Navarro-Colorado's 2018 analysis of themes in classical Spanish sonnets), drama (Schöch's 2017 study on genre in classical French theater), and narrative (such as Jocker's study of 500 themes in a corpus of approx. 3,346 nineteenth-century novels [2013: 118–153]).

Ultimately, an intimate dialogue between the data-informed *distant reading* and the more traditional but data-guided *close reading* of a corpus can also create a prolific tool for text interpretation and for the study of poetics. Both practices benefit from being closely intertwined, concurrently shedding light on the same object from different angles, for instance, by testing or renewing hypotheses about a literary work.

3 A natural fit

Among scholars studying poetry or poetics, and generally where the examination is focused on forms, the boom in computer-driven research has triggered relatively little controversy and has instead been embraced with overall enthusiasm.

This can be explained by the intrinsic fit between quantitative approaches and stylistics or poetics, making the association a natural step for fields that already relied on gathering, organizing, and analyzing data. With its focus on recurring patterns of style, rhyme, meter, and other regulated or technical aspects of a text, the study of poetry and poetics has always leaned toward quantification. In the production, consumption, and analysis of verse, lines get counted, measured by their numbers of stresses, syllables, and groups, labelled according to their configurations. Rhymes, too, get measured, classified according to their patterns and degree of perfection, with rules for evaluating their originality, difficulty, or semantic relevance. One could somewhat provocatively assert that the textual matter of poetry is quantitative by definition and must necessarily be examined as such, echoing Poe's witty definition of verse: "one tenth of it, possibly, may be called ethical; nine tenths, however, appertain to the mathematics" (Poe, 1984: 26). This, of course, should be taken with a pinch of salt, and we are well aware that many other aspects of poetic texts need to be examined, many other approaches that are relevant to and useful for understanding them and for interpreting literary texts in general (such as psychoanalytic, sociologic, historical, or – more recently – eco-poetic perspectives, to name but a small few). Moreover, many of them can be combined with a quantitative twist. But because poetry, among other literary genres, is such an intrinsically regulated device, we should never ignore its quantifiable aspects when examining it.

It thus comes as no surprise that the field has been quick to accommodate computational approaches with very little friction: they have come as a welcome relief and have facilitated and furthered existing counting practices. Card files, where the researchers kept their records, have been replaced with modern databases, simple paper-and-pencil computations with more advanced methods and models. The goals have remained largely unchanged: identifying patterns within the multiplicity of forms, using these patterns to follow trends throughout literary history, identifying what sets one style apart from another, and testing hypotheses about what might actually constitute the efficacy or the uniqueness of a poetic work.

4 Material

On the pretext of giving an overview of our field of study, we, the editors, would like to address one central issue of computational literary studies – that is, the sourcing, operationalization, mixing, and sharing of our most essential commodity: corpora. The quickly progressing digitization of texts is affording us new, different possibilities to access texts, including literary ones, and almost all computational studies of literature are based on the use of digital texts. This is not a resource that we should take for granted, and obtaining corpora of sufficiently good quality is a crucial step in any computational literary endeavor. This calls to mind Bruno Latour’s very apt wordplay in French, about the *données* (“data”) never being actually *données* (“given”) but rather always having to be *obtenues* (“obtained”) (Latour, 1993: 188).¹

The most desirable approach by far, though not the most frequent, is reusing a corpus and annotations that have already been compiled by others in the event that such a resource exists, is available for reuse, and has been suitably configured for the research questions at hand. One is clearly more likely to find such reusable resources when working on a prominent, widely studied author like Shakespeare (e.g., Arefin et al., 2014; Eisen et al., 2017; Plecháč, 2021) or Molière² after others have already spent a considerable amount of time and energy on creating a very high-quality corpus, sometimes forgoing other hermeneutical goals of their own.

When it cannot be satisfactorily sourced, a corpus must be built by combining OCR technology, fastidious proofing, and further formatting. This process takes time, commitment, and a dizzying number of informed, critical choices. The most common format is some form of – more or less canonical – XML-TEI, that is, the combination of an Extensible Markup Language (XML) tree structure with the terms recommended by the Text Encoding Initiative (TEI).³ Alternatives to XML exist, mostly intended to avoid the cumbersome methods of bypassing its rigid tree structure: the annotations in *standoff markup*, instead of being embedded within the text, have pointers to locations within it, thus eliminating the problem of overlap encountered in XML. Either way, an interoperable format should be chosen so that the precious annotations that have been added to the text can be reused. For corpus preparation does not end with the

1 “Décidément, on ne devrait jamais parler de ‘données’ mais toujours d’‘obtenues.’” (Really, one should never speak of “given” but always of “obtained” [our translation].)

2 Molière has been at the center of a very active debate around whether and why he did (or did not) write some of his plays (cf. Labbé, Labbé [2001]; Cafiero, Camps [2019]; Labbé [2019]).

3 See <https://tei-c.org/release/doc/tei-p5-doc/en/html/AB.html> (accessed May 2, 2022).

full and exact text itself; it also involves adding various annotations, automatically, manually, or both. The annotations transform a literary work into a dataset, operationalize it, and allow for a radically changed view, bringing together very localized phenomena and the scale of an entire corpus.

When building a corpus, one can lean toward an immaculate, fully interoperable, detailed, and exhaustive result, nearing a digital edition of texts, or opt for a more robust, goal-oriented corpus, with just the level of granularity and exactness needed for the project at hand. The first option produces a valuable resource, and one more likely to be reused by other researchers as it will fit the needs of many. It also involves such great time and energy that building it might take over completely, leaving no time for its further use. OCR models are trained on a variety of languages, time periods, physical supports, on manuscripts and printed material, etc. Some are shared publicly using free software licenses (e.g., *Kraken*),⁴ others are proprietary (*ABBY*, *Adobe*, *Transkribus*).⁵ Depending on the quality of the source document, and on the precision needed for the purpose at hand, proofing the OCR results can require a lot of manual validation. Either way, it always takes strong theoretical grounding and much work to address the endless stream of choices to be made with regard to the granularity, robustness, and exhaustivity of the corpus. Endeavors as monumental as meticulously preparing and annotating very large corpora are better suited to either digital edition purposes or long-term projects carried out by teams whose primary mission is to produce a resource for the benefit of a larger community, thus constituting one of the aforementioned trusted sources. A recent undertaking of this type is the ELTeC COST (*European Literary Text Collection*) project for narrative corpora in European literature (Odebrecht et al., 2021).⁶ Two other examples, covering more localized language areas, are the *Deutsches Textarchiv* (2007–2021) for texts in German (1600–1900) and *Frantext* (ATILF, 1998–2021) for French texts since the tenth century.

The second, more economical option, appropriate in cases where texts need to be digitized for the purpose of a specific research question, is building a less perfect but sufficiently robust corpus serving only the hermeneutic goals in question. This can be managed for reasonably sized corpora provided one commits to a spirit of *economy*, only including the precise level of granularity required to fit one's specific needs. These corpora, which are snugly customized to one's own needs, are less likely to be reused by others but still require

⁴ *Kraken* OCR models: https://zenodo.org/communities/ocr_models (accessed May 2, 2022).

⁵ <https://www.abbyy.com/cloud-ocr-sdk/>, <https://acrobat.adobe.com>, <https://readcoop.eu/transkribus/> (accessed May 2, 2022).

⁶ <https://www.distant-reading.net/eltec/> (accessed May 2, 2022).

considerable effort. It is essential to refrain from collecting any more data than actually useful for the hermeneutic question at hand and to take care to only collect it in a way that provides satisfactory interoperability. For this purpose, correct XML files with TEI-compatible annotations, for instance, are a safe choice. They can be reused by others who might happen to share the same annotation needs, whereas a list of remarks in a spreadsheet might fit one's immediate needs perfectly well but be nearly impossible to make interoperable.

The continuum of approaches between these two poles is laden with innumerable decisions to be made, some of them philological or otherwise addressing the specifics of a text, some rather strategic, such as whether to delimit a corpus that can or cannot be borrowed, some with regard to interoperability. FAIR principles (Findability, Accessibility, Interoperability, and Reuse; Wilkinson, 2016) aim to facilitate the circulation of digital goods and tools and prevent identical resources from being gathered by several teams. They also trigger a new series of decisions to be made: Which repository to use? Which format? Should everything be transformed into semantic web triples or should the existing resource simply be made accessible as it is? How to avoid violating copyright laws? Is it necessary to write a manual to accompany the resource, thus improving its accessibility but also adding another layer of careful work and potentially endangering one's research objectives? How we choose to solve these problems should take a number of parameters into account: for instance, the timeframe of the project, the workforce at hand, and the research goals envisioned. Many platforms now exist where resources can be shared, although to our knowledge, it is rare for one smaller team to actually reuse resources made by another smaller team because research tends to be very specialized, and individual projects are not very likely to share the same corpora, even less the same annotation needs. Ultimately, we must decide how much effort to put into unattainable, yet desirable, perfect interoperability. Producing high-quality corpora on a large scale makes a laudable contribution to the scientific community's pooled resources but can only be achieved when it takes over as its own research goal, making it a task better suited to larger institutional bodies than to individual researchers as there is little immediate hermeneutic gain to be made from such a sustained effort. This does not mean that we should not have interoperability in mind when preparing a more modest, isolated corpus, as we should always favor the possibility that other researchers can later reuse our own painstaking collections.

When some but not all of the desired corpus can be reused from other sources, one realistic option is to combine reused and purpose-built corpora, adapting the resources to one another, mixing old and new, owned and borrowed (as is the case, in this volume, of DISCOVer, presented by Bermúdez et al., and of Vekshin et al.'s *Russian Live Stylistic Dictionary*, for instance). When different

projects share an interest in the same types of annotations, it creates a joint niche within which the sharing of resources adds value and meaning to the efforts of each team. The possibility of such mixing is precisely why thought should always be given to the interoperability of a resource and how the FAIR principles can improve the productivity of the entire community.

5 Collecting annotations

Whether they are obtained from a reliable source, gathered automatically through a number of treatments, or need to be collected manually, annotations should be chosen and produced very carefully with respect to two main principles: the previously discussed economy in defining the annotations and rigorous *stability* in the protocol. The latter is particularly true with regard to manual annotation efforts, and it is actually a challenge for the annotator to steer clear of spontaneous common sense. Faced with any number of unexpected occurrences, we inevitably wonder what decision to make and are tempted to deal with issues on a case by case basis, but this brings different annotators – or the same annotator at different times or in different moods – to lean one way sometimes and sometimes the other, tainting the data and potentially invalidating it. Optimal inflexibility is particularly hard to achieve as it contradicts our scholarly practices. As humanities scholars, we have been trained to stay acutely aware of the particular meaning of what we read relative to its context. Here, we need to shift our focus away from what a text says and behave with the reliability of a machine. In this respect, it can be helpful to develop annotation guidelines for manual annotation.

5.1 Machine collection

Fortunately, in this dynamic field where new tools are continuously being developed, machines are indeed getting better and better at identifying the features we wish to have annotated. Today, we can see a progressive move away from applying laws and rules to texts in successive layers and toward the use of artificial intelligence to infer word vectors and topics (in a topic modeling sense), to annotate sentiments, and to generally detect things that would have been thought impossible for a non-human annotator. The main obstacle when it comes to machine learning is, once again, a corpus issue. In order to “learn,” supervised models must be trained on vast amounts of pre-existing, pre-annotated material. In fact, in many cases, even unsupervised models require annotated data

in order to be properly evaluated. An additional issue lies in the decision-making functions of the models and the black-box effect that their opacity creates. Even though many machine learning methods do provide a glimpse of how the decisions were made by the machine by assigning weights to particular features, few researchers actually pay a careful attention to their philological interpretation (Kestemont, 2014).

5.2 Noise

When speaking of data, mention must be made of noise, as what is noise for some can be a signal for others or simply lead to the ruin of one's own data, depending on the quality of the dataset. When noise is made by outliers or by non-representative occurrences, or when the sheer volume of a very frequent phenomena is overpowering and obscures a less frequent but interesting trait, then this noise is actually data. One can choose to keep it or to remove it, one can also choose to extract it and study it specifically. But when noise is the result of too high a proportion of errors in the data, of an inadequate collection protocol, then noise is indeed just pollution. Efforts are then made to remove it or to account for it through error bars, but what it signals is really poor data quality.

5.3 Lost in collection

The question of hermeneutical gain is where computational approaches and distant reading sometimes receive the most criticism. In some cases – for instance, where corpora must be annotated before being fed into an artificial intelligence's "learning" process, or for the impressive digital philological efforts being invested into the development of OCR models – there is so much technical work required, work that takes such considerable effort and needs to be performed on such voluminous corpora, that doing this work becomes the very goal of the research endeavor. The highly valuable contributions made by such development projects serve not the literary researchers' own, immediate hermeneutical benefit but that of future generations. They create a common good and form part of a longer timeframe. Such approaches may be the result of selflessness, a taste for programming and data processing challenges, or simply an emphasis on philological rather than literary issues. Sometimes we also see a form of obliviousness in the quest for ever-more and ever-more exact data, of ever-finer granularity, postponing any hermeneutical

gain to such a faraway future that it does not fit within the timeframe or funding scales of most projects.

These time and effort considerations, crucial to the digital humanities, are why it is so important to reuse and remix materials and devices from other sources when putting together individual projects and to make new contributions as interoperable as possible. Before building a resource, we should be careful not to “reinvent the wheel” every step of the way and to assemble existing pieces before resorting to creating new ones.

6 Heritage and criticisms

One recurring criticism of computational literary studies is that it provides nothing new, but does so in complicated ways. Such is, in a few words, the essence of Nan Z Da’s recent criticism of the digital “debacle” and of those she suspects of merely “counting words” (2019a, 2019b), at best without generating any hermeneutic gain, at worst to the detriment of the pursuit of truth: “What is robust is obvious (in the empirical sense) and what is not obvious is not robust” (Da, 2019a: 601). Her cherry-picking of approaches in her opinion paper (Da, 2019a) sparked quite a debate.⁷ In fact, not all of the paper’s verdicts are undeserved, but it overlooks cases where results happen to disprove the apparently obvious hypothesis at hand and misrepresents the disruption brought about by the digital humanities, citing their confirmation of previous findings as evidence of their imposture, when proving something again by a different method is actually valuable validation.

A more nuanced criticism is put forward by Bode (2017), who believes there are gains to be made in new, digital methods but disagrees with Moretti and Jockey’s departure from more traditional scholarly traditions, thus advocating for the more holistic integration of new and old tools. It is indeed frustrating that, when accessing the power of computational approaches and the spread of large or very large corpora, we usually have to surrender some of the more meticulous precision for which traditional scholarship calls. Bode sees here a gap that must be bridged, but it may also be considered a shift in focus. Less attention is being paid to the exact history of each text and more to a broader trend emerging from many texts. A large corpus is not an exactly faithful representation of a literary

⁷ For a few elements of this debate, see researchers’ responses on the *Critical Inquiry* journal site (Da, 2020; Underwood, 2020; Weatherby, 2020).

period, but systematically analyzing many texts at once provides a different view, a broader focus, one that no close reading could have generated.

The legacy of Moretti's *distant reading* – whether embraced or rejected – can be seen in many of the numerous approaches to digital literary studies, if only due to its huge success and spectacular dissemination. His best-selling works (2007, 2013) have played an important role in encouraging and advertising a general drive towards computational methods. His hermeneutical propositions include a certain Darwinism when considering how genres and subgenres evolve, a view of literary trends based not on the meticulous reading of a few selected works, famed and canonized after the fact despite being essentially exceptional, but rather on vast bodies of published works, huge samples better able to represent literary periods and trends in that they place less importance on which ones later won the struggle for survival.

Jockers favors the term “macroanalysis” to refer to the method of grounding literary analysis in quantitative evidence from large corpora, arguing that this term, by avoiding the confusion with *reading*, better conveys the notion of data analysis (2013: 25–30). Jockers sees in macroanalysis the potential to provide “contextualization on an unprecedented scale,” to assess “the historical place of individual texts, authors, and genres in relation to a larger literary context,” to determine literary patterns and lexica used diachronically or based on social and geographic factors, and to examine how literary themes emerge and decline and how literary tastes evolve.

The iconoclastic dimension of such a methodological shift – in its many forms – has not received universal praise, and Moretti's and Jockers's works are regularly and more or less openly criticized. Is it bad faith to admit large yet non-exhaustive corpora as reliable bodies when their content has not yet been disclosed? Is there imprecision in corpus delineation? Are their results reproducible? The reproducibility of Moretti's findings (2005) about the evolution of British novelistic genres has been questioned (Riddell, 2013: 47–49; Shalizi, 2011: 118, 130–131). Other scholars have tested the intuitions presented by Moretti on the basis of a smaller dataset by trying them out on a larger scale. Picking up on Moretti's (2011) use of character networks as a way to study plots, Algee-Hewitt (2017), for instance, examines plot evolution in English drama since the sixteenth century. But whereas Moretti based his study on single plays, Algee-Hewitt offers a viable, large-scale implementation based on centrality measures in character networks for a corpus of 3,895 plays. Be it valid or excessive, criticism in no way invalidates the pioneering role played by Moretti and Jockers in their impactful and engaging writings, in their founding of the first Literary Lab in Stanford, and in their development, promotion, and establishment of new practices that have since spread widely, going beyond their initial propositions. Around two decades

after the term “distant reading” was introduced,⁸ Underwood draws a positive balance of quantitative approaches to literary studies (2019: xx–xxii). He emphasizes their usefulness for comparing large text samples from different social contexts and for understanding the long-term evolution of literary constructs. Underwood’s distant reading method involves “perspectival modeling” (2019: 36–38): supervised machine learning models are trained on different subcorpora (e.g., according to genre and period). Comparing each model’s predictive behavior on different subsets of data allows us to establish commonalities between those subsets, for instance, to trace similarities between subgenres across time.

It must be kept in mind that, in this globally emerging field, different approaches are stemming from different philological traditions. They serve different goals, follow different paths, and have their own relevance. Given this context, any claim to praise or condemn them as a whole, whatever the reason, is bound to be at best partial, at worst misleading.

7 The polite revolution

The methodological turn toward the digital humanities is commonly referred to as a “digital revolution,” but we would like to stress that it really is quite a polite revolution. An apparent break from traditional methods and the proliferation of quickly evolving tools are bound to trigger some degree of skepticism, bewilderment, or tension, and have indeed been met by some scholarly communities with contrasting responses, ranging from enthusiasm to ambivalence and disapproval, with some questioning the validity or usefulness of the turn, others embracing all or some of it wholeheartedly. While a number of all-out disputes have taken place, this (r)evolution should be seen less as a reductive dispute between “traditionalists” and “innovators,” and more as a moment of acute methodological creativity, with researchers branching out toward novel and often heterogeneous tools, experimenting with a number of possibilities, some of which, in turn, inform and sometimes change what they are striving to grasp. It might appear that the main difference lies in the now seemingly obligatory burden of proof, with hypotheses needing to be tested, proved, or disproved where intuition had until recently sufficed. However, and although one can now indeed test hypotheses, the impact of this novelty needs moderating.

⁸ Moretti discusses the introduction of the term in his paper “Conjectures on World Literature,” which originally dates from 2000 and is also part of his 2013 *Distant Reading* volume (2013: 43–44).

While some habits are being challenged or broken, many obstacles and perils are being maintained. Basing our assumptions on quantifiable, reproducible observations does not necessarily afford our arguments a higher degree of truth. At every step, from corpus preparation to criteria definition, from annotating to extracting, from statistics to interpretations, we have to make choices that are informed but inevitably subjective. This is how Drucker (2011) argued that “data are *capta*,” highlighting their constructed and interpretive nature. Any research effort calls for individual sensibility, all the more so when it examines literature. The necessary economy in data collection requires a hierarchization of equally real phenomena. Literature research datasets, and the analyses drawn from them, are inevitably informed and colored by a team’s assumptions, hypotheses, and way of thinking. An acute awareness of how personal research endeavors are – including computational ones – calls for digital humanists to be humble, to link their efforts to more traditional ones not by means of competition but through mutual enrichment, and encourages multifocal reading, *distant* as well as *close*, the two frames proving or disproving, informing or guiding each other.

Faced with having to choose between easier-to-implement but comparatively limited tools and more advanced, perpetually evolving ones requiring a higher level of technical ability, some researchers have decided to enrich their work by taking a little quantitative detour, while others have become actual programmers and are taking huge initiatives to develop elaborate programs, rarely for their sole benefit and more often for that of the wider community, by making their tools available and appealing to the many (e.g., the CATMA annotation platform discussed in this volume) or by making their approach known and potentially reproducible. Indeed, it is not uncommon for conference submissions in digital humanities and computational humanities journals, particularly in natural language processing, to publish accompanying code and datasets. Many of us are putting together patchworks of borrowed and custom-built devices to do one job, making modest contributions to the pooled resources of computational literary scholars.

In a similar movement of methodological hybridization, computational, *distant* reading methods and more traditional, *close* reading can be intertwined, the hermeneutical gain deriving from the diversification of angles and focal points, a mixing favorable to the emergence of unexpected observations.

Generally, computational approaches to literature face challenges common to all interdisciplinary studies. Scholars belong to two communities with very different expectations, present their work to audiences with different frames of mind, and publish it through channels with utterly different focuses and requirements, from those that tolerate a few figures when the general discourse is

convincing even without them to those with an obligation to publish one's data and calculation methods alongside a demonstration. This dual belonging is really an asset as it provides ample opportunities for thorough discussions in more than one area of study and for traditional and novel methodologies and bodies of knowledge to be intricately linked.

This edited volume addresses a number of topics, methods, and applications in order to build a coherent vision of how computational stylistics can be used to study poetry, prose, and drama. After this opening chapter, which attempts to probe the depth and nature of computational literary studies' potential for an actual renewal of literary studies, the articles presented here by Anne Bandry-Scubbi and by Jan Christoph Meister take very different routes toward in-depth reflections on the question of literature's measurability and how new methods are interacting with more traditional hermeneutic approaches. Both Meister's chapter and the one authored by Helena Bermúdez Sabel, Pablo Ruiz Fabo, and Clara Martínez Cantón, present readers with ready-to-use, easily accessible web interfaces to simplify the use of digital literary resources by a broader audience: the first is *Computer Assisted Text Markup and Analysis*, or CATMA for short; the second is DISCOVer, which facilitates the exploration of the DISCO (*Diachronic Spanish Sonnet Corpus*). The next two chapters, by Chris Mustazza and Éliane Delente, focus on questions of rhythm and its exploration using both distant and close reading and listening methods. Natalie Houston uses her data on English poetry rhymes to address literary history and the Victorian period while Jonathan Armoza explores the heterogeneous strands of writing in *Moby-Dick* to unravel Melville's two successive drafts. Georgy Vekshin, Egor Maximov, and Marina Lemesheva present their method for identifying poeticisms in Russian. Turning away from poetry and toward theater, Benjamin Krautter, Janis Pagel, Nils Reiter, and Marcus Willand put together a method for automatically defining some aspects of character types. Pablo Gervás then presents his own poetry generator, using it to shed light upon what a poetic voice is and what issues are raised by the notion of a non-human poet. The chapters are followed by some brief closing remarks and a broader outlook on the volume as a whole and its relevance within the field. This concluding chapter contains summaries for each article, also paying attention to points that are shared by all of them.

References

- Algee-Hewitt M. Distributed Character: Quantitative Models of the English Stage, 1550–1900. *New Literary History* 2017; 48: 751–782.
- Arefin AS, Vimieiro R, Riveros C, Craig H, Moscato P. An Information Theoretic Clustering Approach for Unveiling Authorship Affinities in Shakespearean Era Plays and Poems. *PLoS ONE* 2014; 9(10): e111445.
- ATILF. Base textuelle Frantext. ATILF-CNRS and Université de Lorraine. 1998–2021. <https://www.frantext.fr/> (accessed April 7, 2022).
- Bode K. The Equivalence of “Close” and “Distant” Reading; or, Toward a New Object for Data-Rich Literary History. *Modern Language Quarterly* 2017; 78: 77–106.
- Brunner A, Tu NDT, Weimer L, Jannidis F. To BERT or not to BERT – Comparing Contextual Embeddings in a Deep Learning Architecture for the Automatic Recognition of Four Types of Speech, Thought and Writing Representation. Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS). Zurich, 2020. <http://ceur-ws.org/Vol-2624/paper5.pdf> (accessed May 4, 2022).
- Busa, R. Rapida e meccanica composizione e pubblicazione di indici e concordanze di parole mediante macchine elettrocontabili. *Aevum* 1951; XXV(6): 479–493.
- Busa R. The Annals of Humanities Computing: The Index Thomisticus. *Computers and the Humanities* 1980; 14(2): 83–90.
- Byszuk J, Woźniak M, Kestemont M, Leśniak A, Łukasik W, Śęła A, Eder, M. Detecting Direct Speech in Multilingual Collection of 19th-Century Novels. In: Sprugnoli R, Passarotti M, editors. Proceedings of the LREC 2020. Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2020). Marseille: ELRA, 2020: 100–104.
- Cafiero F, Camps J-B. Why Molière Most Likely Did Write his Plays. *Science Advances* 2019, 5(11). *Critical Inquiry*. Computational Literary Studies: A Critical Inquiry Online Forum. 2019. <https://critinq.wordpress.com/2019/03/31/computational-literary-studies-a-critical-inquiry-online-forum/> (accessed April 27, 2022).
- Cultural Analytics. “Debates” section of the Journal of Cultural Analytics. 2020 <https://culturalanalytics.org/section/1580-debates> (accessed April 7, 2022).
- Da NZ. The Computational Case against Computational Literary Studies. *Critical Inquiry*. 2019a; 45: 601–639.
- Da NZ. The Digital Humanities Debacle. *The Chronicle of Higher Education*. 2019b. (March 27) <https://web.archive.org/web/20200726175908/https://www.chronicle.com/article/the-digital-humanities-debacle/> (accessed April 27, 2022).
- Da NZ. Critical Response III. On EDA, Complexity and Redundancy: A Response to Underwood and Weatherby. *Critical Inquiry*, 2020; 46: 913–924.
- Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. 2021. Berlin-Brandenburg Akademie der Wissenschaften. <https://www.deutschestextarchiv.de/> (accessed April 27, 2022).
- Drucker J. Humanities Approaches to Graphical Display. *Digital Humanities Quarterly*, 2011; 5(1).
- Eisen M, Riberio A, Segarra S, Egan G. Stylometric Analysis of Early Modern Period English Plays. *Digital Scholarship in the Humanities* 2017; 33(3): 500–528.
- Herrmann JB, van Dalen-Oskam K, Schöch C. Revisiting Style, a Key Concept in Literary Studies. *Journal of Literary Theory* 2015; 9(1): 25–52

- Karsdorp F, Kestemont M, Riddell A. *Humanities Data Analysis: Case Studies with Python*. Princeton: Princeton University Press, 2021.
- Jockers ML. *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press, 2013.
- Jockers ML. *Text Analysis with R for Students of Literature*. Cham: Springer, 2014.
- Kestemont M. Function Words in Authorship Attribution: From Black Magic to Theory? In: *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. Gothenburg: Association for Computational Linguistics, 2014: 59–66.
- Kim E. *Emotions in Literature: Computational Modelling in the Context of Genres and Characters*. PhD thesis. Universität Stuttgart, 2020.
- Klinger R, Kim E, Padó S. Emotion Analysis for Literary Studies: Corpus Creation and Computational Modelling. In: Reiter N, Pichler A, Kuhn J, editors. *Reflektierte algorithmische Textanalyse*. Berlin, Boston: De Gruyter, 2020: 237–268.
- Labbé C, Labbé D. Inter-Textual Distance and Authorship Attribution Corneille and Molière. *Journal of Quantitative Linguistics* 2001; 8: 213–231.
- Labbé D. Réponse à Florian Cafiero and Jean-Baptiste Camps. *Why Molière Most Likely Did Write his Plays*. Institut d'études politiques de Grenoble, 2019. <https://hal.archives-ouvertes.fr/halshs-02383640/> (accessed April 27, 2022).
- Latour B. Le topofil de Boa-Vista. *La référence scientifique: montage photophilosophique*. *Raison Pratique* 1993; 4: 187–216.
- Lebart L, Pincemin B, Poudat C. *Analyse des données textuelles*. Quebec: Presses de l'Université du Québec, 2019.
- McCarty W. *Humanities Computing*. London: Palgrave Macmillan, 2005.
- Moretti F. Conjectures on World Literature. *New Left Review* 2000; 1 (January/February): 54–68.
- Moretti F. *Distant Reading*. London: Verso Books, 2013.
- Moretti F. *Graphs, Maps, Trees: Abstract Models for Literary History*. London, New York: Verso, 2007.
- Moretti F. *Network Theory, Plot Analysis*. Pamphlet 2. Stanford: Stanford Literary Lab, 2011.
- Navarro-Colorado B. On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry. *Frontiers in Digital Humanities* 2018. <https://doi.org/10.3389/fdigh.2018.00015>.
- Odebrecht C, Burnard L, Schöch C. *European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels*. 2021. Zenodo. <https://doi.org/10.5281/zenodo.4662444>
- Plecháč P. Relative Contributions of Shakespeare and Fletcher in Henry VIII: An Analysis Based on Most Frequent Words and Most Frequent Rhythmic Patterns. *Digital Scholarship in the Humanities* 2021; 36(2): 430–438.
- Poe EA. *Essays and Reviews*. New York: The Library of America, 1984.
- Ramsay S. *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press, 2011.
- Riddell AB. *Demography of Literary Form: Probabilistic Models for Literary History*. PhD thesis. Duke University, 2013.
- Savoy J. *Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling*. Cham: Springer, 2020.

- Schöch C, Schlör D, Popp S, Brunner A, Henny U, Calvo Tello J. Straight Talk! Automatic Recognition of Direct Speech in Nineteenth-Century French Novels. *Digital Humanities 2016: Conference Abstracts*. Kraków: Jagiellonian University & Pedagogical University, 2016: 346–353. <https://dh2016.adho.org/abstracts/31> (accessed April 27, 2022).
- Shalizi C. Graphs, Trees, Materialism, Fishing. In: Goodwin J, Holbo J, editors. *Reading Graphs, Maps & Trees: Responses to Franco Moretti*. Anderson, SC: Parlor Press, 2011: 115–139.
- Schöch C. Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. *Digital Humanities Quarterly* 2017; 11:2. <http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html> (accessed April 27, 2022).
- Underwood T. Critical Response II. The theoretical Divide Driving Debates about Computation. *Critical Inquiry*, 2020; 46: 900–912.
- Underwood, T. *Distant Horizons: Digital Evidence and Literary Change*. Chicago, London: The University of Chicago Press, 2019.
- Weatherby L. Critical Response I. Prolegomena to a Theory of Data. *Critical Inquiry*, 2020; 46: 891–899.
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, . . . Mons B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 2016; 3: 160018.
- Yule GU. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 1944.

Anne Bandry-Scubbi

Zooming In, Zooming Out: 30 Years of Corpus Stylistics Bricolage

Abstract: This chapter addresses the relationship between distant and close reading, advocating middle-ground reading with a strong focus on texts per se. By taking an evolutionary approach, it reviews pragmatic use over three decades of the mechanically enhanced reading of eighteenth-century British fiction, viewing texts first as systems then as corpora and later as constituents of embedded and overlapping corpora. In a back-and-forth movement between text and corpus, it explores norm and typicality by looking at canonical and non-canonical fiction.

1 Introduction

In the words of the late John Burrows, I “declare myself, first and last, a student of English literature,” having taken up “computational stylistics” (2010: 13) because the methods and tools it provides make it possible to view texts from a somewhat different perspective, combining distant, middle-ground, and close reading. I am no computer wizard and have almost never created the tools I use, but have rather taken what was at hand, or within my grasp, technically, financially, and intellectually. I therefore consider Lévi-Strauss’s notion of bricolage an apt way to qualify 30 odd years of research on British fiction from the eighteenth and early nineteenth century with the help of a computer, which has entailed the risk of not being taken seriously by specialist scholars. Being French, I was trained in the close-reading tradition (*explication de texte*) by the heirs of structuralism: we saw the text as a “network of signifiers” (Barthes, 1970), and our aim was to lay bare the logic of a text, or rather of a fragment of text, seen as a system, a coherent and dynamic whole. What I like to call “computer-aided textual analysis” extends this approach to complete literary works thanks to the non-linear reading enabled by considering a text a “multi-dimensional space” (Rastier, 2001: 93). Once literary texts became more widely available in digital form, this change of scale and of perspective provided “an enhanced contextualization which changed the conception of textuality and intertextuality,” with the possibility of examining a text within a

Anne Bandry-Scubbi, University of Strasbourg, UR 2325 SEARCH, e-mail: bandry@unistra.fr

set of corpora (Rastier, 2001: 93). Gigantism is a temptation, taking into account the “99.5 per cent” that have not made it into the literary canon along with the 0.5% that have (Moretti, 2013: 66), reading them “distantly” but not as coherent narratives per se. The Stanford Literary Lab and others now provide the (very) big picture, usefully broadening the view of “the rise of the novel” into “one rise” or the gender-shift into several (Moretti, 2005: 5, 27). “My” scale can be called middle-ground reading, where the individual text – say, a novel – is analyzed by zooming out onto a corpus, i.e., an organized set of texts that has been compiled for a meaningful reason (by period, genre, or criteria related to a hypothesis) and zooming in on some of its features, which quantitative and qualitative comparison and contrast show to be relevant to that hypothesis. For technical and financial but mainly intellectual reasons, a coherent literary text forms my usual object of study. This chapter focuses on the trials, errors, and successes of taking an evolutionary approach similar to Martinet’s (2020) and is therefore highly autobiographical, with the embarrassment of a largely self-centered bibliography.

2 Texts as systems

In 1985 I was appointed a teacher of English at the science college of a small French university (Université de Haute-Alsace) and began my PhD in British literature, on the texts written in reaction to Laurence Sterne’s *The Life and Opinions of Tristram Shandy, Gentleman*. The concomitance of these two events has shaped the entirety of my research. It coincided with the implementation of the French national project *Informatique pour tous (Computing for All)*, which made computers available to students and staff but also in libraries, and led to the private acquisition of personal computers at affordable prices. I bought my first machine in 1987. Part of my teaching was English for computer science students. Trying to convince them that “if/then/else” could not be used in *real* English taught me how to communicate with programmers, and I traded proofreading my colleagues’ publications in English for devising a program that produced frequency indexes of the genuine and spurious *Shandy* volumes on university computers, which I could then use on my own.

Software such as the Oxford Concordance Programme was out of my reach, as were digital texts, but networking before the internet, I strove to meet researchers in France and abroad who combined the study of literature, or at least the humanities, and the use of computers. I visited the Oxford University Computing Services (OUCS) and Besançon linguists, and most usefully came into contact with Parisian colleagues who had set up groups in their respective

universities. I became a regular visitor to Liliane Gallet and Marie-Madeleine Martinet's CATI center (*Cultures anglophones et technologies de l'information / Anglophone Cultures and Information Technologies*) at the University of Paris IV-Sorbonne and a regular contributor to Françoise Deconinck-Brossard's RAO group (*Recherche assistée par ordinateur / Computer-aided research*) at the University of Paris X-Nanterre. These annual meetings along with frequent correspondence provided an ideal forum to discuss ideas, methodologies, protocols, and the analysis of data. Simultaneously, I was becoming part of the network of Sterne scholars, which gave me access to a printed concordance and frequency index of Sterne's sermons (thanks to Kenneth Monkman at Shandy Hall) and, even more valuably, a digital copy of Sterne's text. Somewhat like the original readers of *Shandy*, published in installments from 1759 to 1767, I received the 6,250-bpi magnetic tapes volume by volume throughout the year 1989 from an American scholar (Diana Patterson), who was typing them out for a digital facsimile of the first edition. As is well known, Sterne exposes the workings of storytelling and printing conventions, and provoked readers to react to his playful texts. These reactions were the focus of my PhD. By 1990, I had 11 volumes of *Shandy* in ASCII text: the nine genuine ones and the two spurious ones for which I had paid a typist, although I had to correct her work very thoroughly as eighteenth-century English was beyond her skills. The OCR tests I had obtained from Strasbourg and Nice had convinced me that typing could be a better method for texts written prior to 1830.¹ I had also acquired *Text-search*, which ran on my home computer and added to indexing the possibility of concordancing.

I could then work using an approach not yet called corpus stylistics, which, in discussions with Deconinck-Brossard, I named computer-aided text analysis (in French: *Analyse textuelle assistée par ordinateur*, or ATAO for short) after computer-aided design, engineering, publishing, etc. Even though I was very much interested in technology, my aim was not to devise software but to use it to enrich my literary analysis of *Tristram Shandy* by looking at which traits of his writing had been lampooned, parodied, or pastiched. As Milic had written of Sterne, "a clever imitator could perhaps duplicate the imitable features of the vocabulary and thus render worthless the lexical criteria of style description and identification" (Milic, 1967b). Statistics derived from frequency counts of vocabulary helped explain why the spurious third volume of *Shandy* read so badly while the spurious ninth was good enough to fool the first German translator. The most frequent

1 I visited Charles Muller in Strasbourg in 1990 and, at his invitation, met Etienne Brunet at a PhD defense a few months later. My questions were then largely about OCR.

words² gave insights not easily perceived by linear reading (my guides were Burrows on Austen [1987], Milic on Swift [1967a], Kenny [1986], and Farrington [1989]), contrary to the Shandean dashes and fragmented text, which became the craze and were conspicuous on the page, whether genuine, spurious, or simply written under the influence of Sterne. The clumsy, spurious third volume had far less varied vocabulary despite its many narrative false starts; it resorted to the verb *be* far too often and did not manage to catch the correct ratio of *and* and *the*. On the contrary, the author of the spurious ninth volume from 1766 did much better – but admittedly had more to work with. This study, published in 1992, was expanded to include three sequels to *A Sentimental Journey* (Sterne died shortly after having published volumes I and II in 1768, but subscribers were also expecting III and IV, which created a potential market). I examined the most frequent adjectives and nouns to conclude, most notably, that the least successful sequel contained the fewest terms designating males (Bandry, 2000). Not many of the swarm of imitators managed to latch on to Sterne’s ironic and deft sentimentalism. Their attempts to recreate his manner of writing provided me with the opportunity to explore writing taken as a quantitative norm (Sterne’s) and the deviations of imitations. I also gave in to the attribution temptation with *The Clockmaker’s Outcry Against the Author of The Life and Opinions of Tristram Shandy* in collaboration with my Sterne mentor Geoffrey Day. Circumstantial evidence proved far stronger than stylometric comparison.

I was lucky to be able to publish my findings in the journal that had hosted Deconinck-Brossard’s “Confessions d’une dix-huitième branchée” (Confessions of a Wired Eighteenth-Century Scholar; https://www.persee.fr/doc/xvii_0291-3798_1996_num_42_1_1326), now *XVII–XVIII* (of which I was General Editor from 2012 to 2018; <https://journals.openedition.org/1718/>). We collaborated a few years later on *Moll Flanders*.³ “On peut compter sur Moll” (You Can Count on Moll; 1997) relies on frequency lists and keyword concordances, both quantitative and qualitative. By following some of the most frequent substantives throughout the text with concordances, such as *house*, we teased out some of

2 Being no linguist, I have always used a very basic definition of the word: a sequence of letters between two blanks or punctuation marks. For the same reason I have never sought to lemmatize my texts automatically.

3 In the meantime, I had been promoted to senior lecturer in the English department of the Université de Haute-Alsace, and the two of us we were teaching this novel at our respective universities. This computer-aided analysis tied in with work being done by CATI on *Georgian cities* (<http://www.18thc-cities.paris-sorbonne.fr/Space-and-Emotions.html?lang=en>), in which *Moll Flanders* is one of the texts taken into account (for the evolution of this project see Martinet, 2020).

the ways in which Defoe weaves his story, and we challenged critical views of his text while confirming others.⁴ *Gentlewoman* provides a striking example of how rewarding it is to follow a keyword throughout an entire text as, from the start, Defoe gives it a tainted definition that the heroine adopts as a rule of conduct to the very end, through the vagaries of its 64 occurrences. No computer is needed to relish the irony of the final uses, which seriously undermine the respectability of the denomination due to the incongruousness of Moll's newly found adult son calling her a gentlewoman while she gives him a watch without "tell[ing] him [she] had stole it from a Gentlewoman's side" (Defoe, 2011: 281). Yet, following the progression with a concordance from each occurrence to the next is evidence of the snowball effect by which each new use adds a new meaning. Concordances helped us to read the text as a system.

Using the text as a corpus required dividing it into parts. As Defoe's fiction has no chapters or any other marked divisions, we divided it into 22 sections according to textual signals that the narrator uses to indicate a shift from one episode to the next. We did not want the software to decide on partition arbitrarily. We could then compare the parts in terms of vocabulary use. For the first time, we plotted episodes and vocabulary on a map established by correspondence analysis. From the low value of the first extraction and the very crowded graph of the first two factors, we concluded that *Moll Flanders* is a very homogeneous text, but thereby exposed the readers of *XVII–XVIII* to greater complexity than that to which they were accustomed. We were trying out tools that became much easier to use (for me at least) with the program *Hyperbase*.

From 2001, *Hyperbase* became my main tool, which I have used on many corpora since, whereas Deconinck-Brossard applied different tools to her homiletic corpora. *Hyperbase* is the brainchild of Etienne Brunet, a French scholar and developer, who had started with published frequency indexes of an author or a period: *Le vocabulaire de Jean Giraudoux: structure et évolution* (1978), *Le vocabulaire de Proust* (1983), *Le vocabulaire français de 1789 à nos jours* (1981).⁵ I invited him to my master's seminar in 2007, and my students presented computer-aided analyses they had prepared on a corpus of modernist short stories as a basis for discussion. In a few clicks, *Hyperbase* provides user-friendly graphs.

⁴ I had done this with life and opinions in *Shandy* and some of the spinoffs (Bandry, 1993). Very often, the last use of an important term comes with an ironic twist.

⁵ *Hyperbase* uses *Le Trésor de la langue française* by default for external comparisons, but this can be switched to the *BNC* for English; dictionaries for Italian and Portuguese are also available. The software can be used with any Latin alphabet.

The first of my textual analyses that relied on *Hyperbase* focused on *Gulliver's Travels*. In “Gulliver et la machine à compter: une étude de spécificités” (Gulliver and the Counting Machine: A Study of Quantitative Keyness; 2001) I explored quantitative keyness by contrasting the four voyages. The French term “spécificités” is less of a challenge for literary scholars, particularly with the crystal-clear definition provided by Lebart and Salem: “forms ‘abnormally’ frequent in one part of the corpus” (1994: 260; the calculation relies on the hypergeometric model). Results obtained from *Gulliver's Travels* led to a modest proposal for a new way to look at a text that has given rise to a huge amount of critical work. Far less ambitious than Milic’s *A Quantitative Approach to the Style of Jonathan Swift* (1967a), my contribution aimed to offer evidence of features of the text established from non-linear reading. The three conclusions I drew from quantitative keyness were, firstly, that the expression of size contradicts expectations: terms of bigness appear more frequently in Lilliput and of smallness in Brobdingnag; secondly, that the narrowness of the Houyhnhnms’ world and vocabulary is all the more effective as it comes after the richness of the voyage to Laputa, Balnibarbi, Luggnagg, Glubbudubdrib, and Japan. This can be summed up in one fact: “the thing that is not” brings no new word, and certainly no hapax (the third Voyage has 15% more types and 22% more hapax legomena than expected if the other three are taken as a norm). The third main finding is that the combined receding use of first-person plural pronouns and the increasing presence of third-person plural pronouns over the four books convey Gulliver’s gradual alienation from the human race: by the fourth voyage, humans are *they* rather than *we*. This reading did not reveal anything new about Swift’s famous book but showed how the writer achieves the effects created in the text. Going back and forth between book and data made it possible to apply techniques of close reading to a text of slightly over 100,000 words.

3 Corpus-based approach

My next experiment was less convincing but an important step. Thanks to the digital turn taken in most universities and to a paper on my *Sterne* findings presented before the informal, international research English studies group within EUCOR – The European Campus (European Confederation of Upper-Rhine Universities), I was given access to the treasure trove of Chadwick-Healey’s *Eighteenth-Century Fiction Database*, which Basel University had acquired but nobody was using – and which most French university libraries could not afford. After a frenzy of downloading material onto floppy disks, I came home with reliable electronic

editions of nearly 100 texts that I could use on my personal computer.⁶ However, overwhelmed by quantity, I did not at first adopt a corpus approach. At several conferences, I examined a key term or set of terms in a series of fictions from the eighteenth century to explore a theme, but it took me a while to realize why I had lost the satisfaction of providing a view of what I like to call the texture of the text and was occasionally able to illustrate using textual imaging, a name coined after medical imaging (mainly bar charts at that point, produced using a spreadsheet or *Hyperbase*).

I therefore decided to explore in depth how one could take a corpus approach, with *Excel* and *Hyperbase* as my tools. I first concentrated on Defoe's fiction by looking at eight texts,⁷ starting with what I was comfortable with and pushing it somewhat further: lexical richness, the distribution of most frequent words (pronouns, nouns, verbs), a factorial analysis map representing lexical connection. I was able to better explain this as *Hyperbase* provided easier-to-read graphs, and a corpus comprising eight parts was more manageable than the 22 we had devised for *Moll Flanders*. A readily comprehensible paper had shown me the way: it used *Hyperbase* to compare Seneca's tragedies (Mellet, 1998). In the same manner, the eight Defoe novels are positioned according to the vocabulary they share.⁸ Colleagues in mathematics helped me by pointing out that if the first results made sense in terms of what I knew about the texts from literary history and traditional stylistic analysis, I could trust the data the software was producing and use it to develop further analyses. The texts were positioned in ways that made sense both in their groupings (the two *Robinsons*, the two female stories) and in their differentiations (the texts really considered as fiction vs. those with more ambivalent status, with *Journal of the Plague Year* in the middle). I then concentrated on quantitative keyness as I had done with *Gulliver's Travels* but applied this to groups of texts in order to examine whether the congruence between the two *Robinsons* and that between *Moll* and *Roxana* was thematic or came from specific stylistic traits. Other eighteenth-century texts provided external elements of comparison, most notably *Gulliver*, *Joseph*

6 Reliable electronic text had been a challenge so far, as Gutenberg.org editions were not always so, *Shandy* being a case in point. I had carefully proofread the digital *Moll Flanders* and *Gulliver's Travels* before using them.

7 The semi-fictional status of the *Journal* was of particular interest. The debate on Defoe de-attributions had begun some 15 years earlier.

8 A word contributes to drawing two texts together if it belongs to both and to pulling them apart if it only occurs in one of them (Brunet, 2011: 60).

Andrews, *Memoirs of a Woman of Pleasure*, and *Betsy Thoughtless*.⁹ In the background, a more specific research question began to take shape: what differentiates stories of female characters from those of male ones, and what does this have to do with whether they were written by men or by women? Not unexpectedly, what drew the most interest from the very few readers of this unpublished essay was the comparative analysis of how the combined uses of the verbs *see* and *know* structure *Moll Flanders* and *Roxana*, considering the texts as a system rather than as a corpus.¹⁰ I had spotted these two verbs from the list of words whose frequency increases as the corpus unfolds, another finding provided by *Hyperbase*, and therefore made the methodological point that looking at texts as data reveals what we need to focus on: underlying features of style not necessarily perceived in linear reading or traditional literary analyses.

The second essay on Eliza Haywood's fiction aimed to explore the differences and similarities between her racy texts of the 1720s and the ones written in the wake of – and in reaction to – Richardson's *Pamela*, as Haywood, both hailed and derided as the "Great Arbitress of Passion," had been able to adapt to changing tastes and produce fiction that sold in the 1740s after having been a best-selling author 20 years earlier.¹¹ The first stage of the study was to identify Haywood's vocabulary in *Love in Excess* by contrasting it with that of the other 1719 bestseller, *Robinson Crusoe*, and the just as successful *Gulliver's Travels* of 1726. Each text was divided into two narratively logical sections so as to constitute a corpus of six parts roughly equal in length (from 61,000 to 41,000 tokens). Lexical richness established from a set of comparisons (type/token ratio, hapax legomena, successive 1,000-word sections) ranked *Gulliver* first, *Love in Excess* second, and *Robinson* third. It confirmed my interest in studying Haywood, who recent research was pushing as a major forgotten author to be canonized in the feminist rewriting of literary history, in reaction to Ian Watt's founding critical text *The Rise of the Novel* and his infamous dismissal: "The majority of eighteenth-century novels were actually written by women, but this had long remained a purely quantitative assertion of dominance" before Burney and Austen (1957: 298).¹² Some of my conclusions confirmed well-known

⁹ The pornographic content of Cleland's *Memoirs* (1748–1749) does not prevent it from having a very strong formal similarity with contemporary fiction.

¹⁰ This essay and the one on Haywood were part of my *Habilitation*. The analysis mentioned is now available in Bandry, 2018.

¹¹ The qualification comes from the anonymous "To Mrs Eliza Haywood on her Writing" (1732), written after she had become one of Pope's victims in *Dunciad Variorum* (1729).

¹² The seminal series of essays *The Passionate Fictions of Eliza Haywood* was published in 2000 by Kirsten T. Saxton and Rebecca P. Boccicchio after Paula Backscheider, a renowned

facts: Haywood's characteristics were her breathless syntax (the paucity of *and* and *or* is made up for by a superabundance of exclamation marks, parentheses, and dashes, of which the recent printed edition has provided reliable proof) and the repetitive use of specific (and thematic) words: *love, passion, friendship, cry, opportunity, desire*.¹³ It seems a truism to find these features by contrasting the volume with *Gulliver's Travels* and *Robinson Crusoe*. However, the comparison with the quantitative keywords of Defoe's stories of women shows that he did not adopt his rival novelist's vocabulary for his ventures into the feminine, concentrating instead on their survival (*house, money, husband*) and female status (*girl, woman, mother*). Haywood is more interested in the relationships between characters, and particularly, but not only, between women and men. In the rest of her fiction, both from the 1720s and the 1740s–1750s, she explores different combinations. A quantitative keyword approach in a corpus comprising seven texts of varying length by Haywood shows how she varies her recipes with the same ingredients, toning down the raciness in the 1740s but maintaining a strong interest in expressing female sensuality and agency. The difference in size between the parts of the corpus was taken into account as a possible factor for some of the stylistic analyses.¹⁴ Despite this, the brevity of quantitative keywords lists for each part shows the strong homogeneity of the *HAYWOOD* corpus, in contrast with the long lists from *DEFOE*. Haywood varies the distribution of what is clearly *her* vocabulary. In that study (2004) and two later articles I then combined the corpus approach with a focus on one particular text (Bandry-Scubbi, 2010, 2012). Like the view of *Gulliver's Travels* provided by the contrastive study of quantitative keywords, the back-and-forth movements between a text and a corpus to which it belongs, zooming in and zooming out, contextualization and intertextualization, gave me the opportunity to draw out some stylistic features of texts that were becoming part of the literary canon.

Defoe scholar, called for the reinvestigation of her work: “we must [. . .] problematize, complicate, and revise many of the commonly accepted opinions about Haywood's work” (Bakscheider, 1998: 90). David Oakleaf's 2000 edition of *Love in Excess, or the Fatal Enquiry* provided the necessary reliable paper copy against which to compare the electronic Chadwick-Healey version.

13 A strongly marked preference for “words” in Haywood and “word” in Defoe and Swift tipped the scale in favor of not lemmatizing, a point advocated by linguists that I have often found counterproductive for literary analysis. *Hyperbase* makes it very easy to draw up lists and so to gather the forms of a lemma easily when useful.

14 Two long novels were divided into their initial volumes so that the parts of the corpus varied from 12,000 to 85,000 tokens.

4 Embedded and overlapping corpora

This dual level of analysis was set up on a larger scale for the study of another eighteenth-century novel, *Roderick Random* (Bandry-Scubbi, 2009). Smollett's first published work of 1748 (approx. 200,000 words) was written in the context of "the rise of the novel," in rivalry with Fielding, who had positioned his 1742 *Joseph Andrews* as a "species of writing [. . .] hitherto unattempted" (8). The third open contender was Richardson, who had taken the literary market by storm in 1740 with his own "new species of writing," *Pamela*. Both Fielding and Haywood had reacted to Richardson's epistolary novel, the first with the pithy *Shamela* and the second with the somewhat rambling *Anti-Pamela*, both part of a vast movement now dubbed "The Pamela Vogue."¹⁵ Both then went on to write several examples of what came to be called novels. I therefore set up a series of corpora to draw out the stylistic features of *Roderick Random*. The 1740s corpus comprises fictional texts of comparable length from that decade with a balance between male and female authors as well as between stories of male or female protagonists. *Random* being written as a first-person narrative, the *1STPERSON* corpus is composed of fictional autobiographies from the eighteenth century, with some novels also present in the 1740s. The risk of bias caused by the prevalence of Defoe is balanced by the advantage of having several corpora of similar sizes, including the novel under study, as *Random* also belongs to the *SMOLLETT* corpus, which comprises all of Smollett's fiction. I therefore had three corpora of over a million words each, countering the objections of linguist colleagues that I did not have enough data with which to work. Moreover, *Random* taken on its own makes up two different corpora: one in which it is divided into its 69 chapters and another into nine parts, the main stages of the narrative. These overlapping corpora made it possible to reinvestigate the claims made in the 1970s by Smollett scholars who had examined his style without the help of computers but with the logic of samples, which also prevailed in the work of the first digital stylistic studies such as Milic's for reasons of available computer capacity. Indeed, the features of what was described as "writing in the superlative" (Bouc e, 1971) mainly occur in the specific samples they selected (Grant, 1977). Yet these scholars dealt with what one of them called "language as projectile" (Grant, 1982), the very fast pace of this story of an impulsive and lusty young male character depicted in a series of violent adventures, who embodied the name his author gave

¹⁵ All these texts are available in image form in Chadwick Healey's *ECCO*. At DH2016 they outrageously proposed providing all of the material in text form to subscribing libraries who would agree to pay an additional fee.

him: Random. Focusing on this text set in several corpora can be likened to looking through a kaleidoscope: changing the corpus in which *Random* is observed reveals different features by comparison and contrast. Some traits specific to this text recur whatever the corpus; others show how Smollett conforms to the writing conventions of his contemporaries. Quantitative keyness drawn out from the 1740s corpora *1STPERSON* and *SMOLLETT* enabled me to zoom in on the stylistic means that make the reader feel that he or she is being rushed through the text.

Syntax constituted a first set of features: sentence length, the number of parentheses – examined with due caution – the heavy use of WH relative clauses, all of which characterize Smollett’s early fiction as the *SMOLLETT* corpus showed, and as the knowledge that he shortened his sentences when revising his second novel *Peregrine Pickle* confirmed (Bouc , 1971). I then looked at indications of how “time is collapsed” (Stevick) with the accumulation of vocabulary indicating temporality, and at the way Smollett refers to the body (verbs and organs, body parts – only *Woman of Pleasure* ranks higher among the 19 novels taken into consideration). Another quantitative keyword led to an understanding of how Smollett relates this fictional autobiography: *my* is the word most specific to *Random* in *1STPERSON*. It can be deduced from concordances that this fictional world is organized around the speaker who mentions “my” body, “my” situation, and “my” acquaintances more than himself as “I” (Roderick is not the only one who uses the first person, of course). To draw these features out, I experimented with lists, on a larger scale than in the article I wrote with Deconinck-Brossard (2005): WH words, time indications, the vocabulary of the body. One version of this study more focused on its use of *Hyperbase* got me into Brunet’s session at JADT 2010. However, my strong stylistic bent put me in an awkward position among scholars like Jean-Marie Viprey, who had moved from their inspiring analyses of an author’s style (1997) to demonstrating the software they were fine-tuning. Meanwhile my work was deemed too technical to be published in volumes derived from conferences primarily concerned with literature or cultural studies, at which I presented papers on the use of body words by female authors with my corpora as a backdrop (what hands do in *Pamela*, *Evelina*, *Isabella*, or *Pride and Prejudice*, for instance). I nearly gave corpus stylistics up.

However, its use by a doctoral student of mine to study the expression of space in children’s fiction over a century confirmed its potential.¹⁶ In the

¹⁶ Caroline Orbann, “L’Espace imaginaire dans le roman de jeunesse britannique: de *Water-Babies* de Charles Kingsley (1863)   *Charlie and the Great Glass Elevator* (1973),” co-supervised with Professor Monique Chassagnol and defended in 2016.

meantime, Deconinck-Brossard and I had compared the sermons and fiction of Sterne and Swift to test whether genre or author was the strongest criterion of authorship, with the interesting twist that Sterne wove one of his published sermons into *Shandy* (2005). This study relied on our previous work and on Biber's categories. We looked at very frequent words and hapax legomena, quantitative keyness, and verbs of persuasion and of assertion, the latter identified by Biber as indicative of narration. We divided *Shandy* into its chapters to obtain units of a size comparable to sermons, examined Swift and Sterne's sermons within a larger homiletic corpus, and reached the expected conclusion that genre prevailed over authorship. Our interest was in testing methods under the aegis of the French Society of English Stylistics. We had been discussing the notions of stylistic signature (Milic, 1967a; Milic, 1967b) and linguistic fingerprint ever since our first encounter in the early 1990s, talking with computer scientists and forensic linguists, taking into account the computer-assisted analysis of Romain Gary writing under the pseudonym of Emile Ajar, which shows that an author can change the way he writes (Tirvengadam, 1998).

In 2012 I discovered Chawton House Library's Novels Online collection. This "ongoing project [aims at] making freely accessible full-text transcripts of some of the rarest works in the Chawton House library collection," which consists of "works by women, mostly in English, and mostly within the period 1600–1830."¹⁷ I had found my second treasure trove, thanks to which I could combine all my experience, working with a large quantity of reliable electronic texts both canonical and non-canonical. I set up a project entitled "Strategies of Writing: Women's Fiction in the Long Eighteenth-Century, a Corpus-Based Stylistic Analysis" and read up on critical work by scholars linked to Chawton as well as recent advances in what was now being called corpus stylistics (Mahlberg, 2007; Fischer-Starcke, 2010; Jockers, 2013), and the Stanford Literary Lab pamphlets (<https://litlab.stanford.edu/pamphlets/>). My findings were published by *ABO: Interactive Journal for Women in the Arts, 1640–1830*, whose aims matched my objective: computer-aided literary analysis (Bandry-Scubbi, 2015). The challenge I had set myself was to identify the features that constitute typicality within a set of 42 novels by women published between 1752 and 1834, 34 of which come from Chawton Novels Online and eight of which are now part of the canon, by Jane Austen, Frances Burney, Maria Edgeworth, and Eliza Haywood. They come under the heading of "feminine" novels, "domestic comedy,

¹⁷ <https://chawtonhouse.org/the-library/library-collections/womens-writing-in-english/novels-online/> (accessed May 4, 2022).

centring on a heroine, in which the critical action is an inward progress towards judgment” (Butler, 1975: 145), neither gothic nor historical, the other two main categories of fiction that developed in the period. A reference corpus was set up with 34 novels meeting the same space and time criteria, half of them by male authors. The choice was determined in part by which texts were available online. The *CHAWTON34* corpus is embedded within *WOMEN42*, which partly overlaps with *CONTROL34*, both comprising over 5 million words. Seven of the Chawton texts were published by the (in)famous Minerva Press, which had a reputation for saturating the market with formulaic novels: they turned out to be indistinguishable from the rest on lexical connection maps. Correspondence analysis on types distinguishes female texts from male ones, indicating that specific vocabulary is used by each gender. Quite logically, the same process applied to tokens separates texts according to whether their protagonists were male or female. The quantitative keyness of texts by female authors in *CONTROL34* shows that what characterizes these texts is an interest in both genders (whereas fiction by men in the reference corpus takes females into account to a much lesser extent), a strong use of small-group interaction and of dialogue (with a particular liking for *cried* which points to the prevalence of intensity), and a concern for feelings and emotions, family, marriage, and sight. Typicality can reside in the rate at which words are used, as one “eccentric” text becomes unexceptional when tokens are taken into account rather than types. The list of disproportionately frequent terms in *Three Weeks on the Downs* compared to the *CHAWTON34* corpus shows that its originality comes only from its setting (a ship), not the way in which it uses vocabulary. The rest of the title gives away the reason for its presence in the corpus: *or Conjugal Fidelity Rewarded, Exemplified in the Narrative of Helen and Edmund*. *WOMEN42*’s most central text in terms of tokens was chosen as a case study, for it presents the seeming paradox of being “a highly original tale” (Brown, Clements, Grundy, 2006) told in unoriginal terms. *Rachel*, a didactic tale, exemplifies the norm of the corpus by playing with concepts that had become somewhat outmoded by the time it was published, such as sensibility. This quantitative analysis came to the same conclusion as one of the contemporary reviewers (Bandry-Scubbi, 2015: 21–22). The use of tokens clusters most canonical texts at a safe distance from most Chawton novels. From this, it can be inferred that Austen, Burney, Edgeworth, and Haywood (unsurprisingly) drew on the same stock of words as their less famous female contemporaries, but they used them at different rates. The difference between these canonical texts and the rest of the reference corpus provides a basis on which to identify common vocabulary traits. This approach makes the less famous fiction a benchmark rather than providing a normative judgment of texts that left a small footprint in literary history.

Although the Chawton staff did not know of any other use made of these electronic texts, I was surprised to receive an article while I was giving a paper on “Women’s Novels 1750s–1830s and the Company They Keep: A Computational Stylistic Approach” at DH2016 in Krakow. Using far more sophisticated techniques, Jan Rybicki had compared the Chawton texts to those by “famous men” and “famous women” authors of eighteenth- and early nineteenth-century fiction to identify the gender of authors using a multivariate analysis of the 100 to 1,000 most frequent words. We reached the same main conclusion: women who became famous (Burney, Edgeworth, Austen) were the ones who wrote more like their male counterparts. This validated my choice to stick to a number of novels I can actually read and focus on, navigating between close and middle-ground rather than distant reading, because my interest resides in how stories are told, how unfolding a text into a network of words (the French *explication de texte*) enables one to relish its texture.

I have since relied on this study as a basis to examine different themes, sometimes specific (e.g., the narrative use of miniatures [Bandry-Scubbi, Friant-Kessler, 2018]), sometimes broad (the modes and narrative use of leisure in a forthcoming article).¹⁸

5 Conclusion

What Corpus Stylistics can do beyond the obvious provision of quantitative data is help with the analysis of an individual text by providing various options for the comparison of one text with groups of other texts to identify tendencies, intertextual relationships, or reflections of social and cultural contexts. (Mahlberg, 2007: 221)

When I came upon Mahlberg’s work, I related strongly to the new name given to the kind of research I had been doing all along: it asserts equal status with corpus linguistics and shows the evolution of the work carried out since the 1950s and 1960s by the pioneers who called it lexicometry or stylostatistics. The recent definition of style by Herrmann et al. takes corpus stylistics into account, but not exclusively: “Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively” (Herrmann et al., 2015: 44). This shows, I think, the maturity of the discipline. All

¹⁸ I am also supervising two doctoral theses on similar corpora, one with a quantitative approach and one without: Juliette Misset, “Reading the Didactic Mode: British Novels, 1778–1814”; Lucy-Anne Katgely, “Entre obscurité et renommée: trajectoires et chemins de traverse des romans britanniques féminins de 1789 à 1830.”

scales are needed: close reading, distant reading, and, as I claim, middle-ground reading. In the interplay between a specific text and a set of carefully constructed corpora, words “knock on the door” (Rastier, 2001: 96) thanks to results such as quantitative keyness and the systematic analysis of concordances. Stylistic hypotheses can thereby be formulated, checked, proven wrong, lead to others, and participate in the rewriting of literary history (Moretti, 2013: 64) or, more modestly, counter critical views of a text. Hopefully, such tinkering often proves more fruitful than that of the Lagado professor: “he had emptied [sic] the whole Vocabulary into his frame, and made the strictest Computation of the general Proportion there is in Books between the Number of Particles, Nouns, and Verbs, and other Parts of Speech” (Swift, 2008: 172). The result derided by Gulliver should coalesce into major works “improving speculative Knowledge” but has not gone beyond the stage of “several Volumes in large Folio [. . .] of broken sentences” (ibid). With or without an Oulipian touch, mechanically enhanced writing is now on its way. Mechanically enhanced *reading* should not lose sight of the literary text as an entity with its own logic, aiming to give the reader – whether or not a scholar – pleasure.

References

- Backscheider PR. The Shadow of an Author: Eliza Haywood. *Eighteenth-Century Fiction* 1998; 11(1): 79–102.
- Bandry A. Gulliver et la machine à compter: une étude de spécificités. XVII–XVIII 2001; 53(1): 145–157. <https://doi.org/10.3406/xvii.2001.1601>.
- Bandry A. Les livres de Sterne: Suites et fins. XVII–XVIII 2000; 50(1): 115–136. <https://doi.org/10.3406/xvii.2000.1481>.
- Bandry A, Deconinck-Brossard F. On peut compter sur Moll [Flanders]. XVII–XVIII 1997; 45: 171–190. <https://doi.org/10.3406/xvii.1997.2078>.
- Bandry-Scubbi A. Chawton Novels Online, Women’s Writing 1751–1834 and Computer-Aided Textual Analysis. *ABO: Interactive Journal for Women in the Arts, 1640–1830* 2018; 5(2): Article 1. <http://dx.doi.org/10.5038/2157-7129.5.2.1>.
- Bandry-Scubbi A. La difficile acceptation du foyer dans *The History of Jemmy and Jenny Jessamy* d’Eliza Haywood. In: Lysøe E., editor. *Signes du feu*. Paris: L’Harmattan, 2010: 87–97.
- Bandry-Scubbi A. Les mots de Haywood. In: Deconinck-Brossard F, editor. *Recherche Assistée par Ordinateur (RAO)*. Paris: Université Paris-Nanterre: 2004. <https://crea.parisnanterre.fr/archives/archives-des-anciens-groupes-de-recherche/recherche-assistee-par-ordinateur-rao> (accessed April 4, 2022).
- Bandry-Scubbi A. Renaissance de/chez Eliza Haywood. In: Bazin C, Leduc G, editors. *Littérature anglo-saxonne au féminin: (re)naissance(s) et horizons, XVIIIe-XXe siècles*. Paris: L’Harmattan, 2012: 17–39.

- Bandry-Scubbi A. Roderick Random Amidst Eighteenth-Century Fiction: A Computer-aided Textual Analysis. *XVII-XVIII* 2009; 66: 205–225. <https://doi.org/10.3406/xvii.2009.2399>.
- Bandry-Scubbi A. Roxana, réseaux de mots, prisons des corps. Journée d'étude spéciale agrégation d'anglais, 24 novembre 2018, Université de Lorraine. <https://videos.univ-lorraine.fr/index.php?act=view&id=7240>.
- Bandry-Scubbi A, Deconinck-Brossard F. De la lexicométrie à la stylostatistique? Sterne et Swift: textes croisés. *Bulletin de Stylistique anglaise* 2005; 26: 67–85.
- Bandry-Scubbi A, Friant-Kessler B. Peindre en corpus: Miniatures et roman anglais féminin (1751–1834). In: Deconinck-Brossard F, Gallet-Blanchard L, editors. *Palette pour Marie-Madeleine Martinet*, 2017. <http://www.csti.paris-sorbonne.fr/centre/palette/index.html> (accessed April 4, 2022).
- Barthes R. *S/Z*. Paris: Seuil, 1970.
- Biber D. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press, 1988.
- Boucé P-G. *Les romans de Smollett*. Paris: Didier Erudition, 1971.
- Brown S, Clements P, Grundy I, editors. *Orlando: Women's Writing in the British Isles from the Beginnings to the Present*. Cambridge: Cambridge University Press Online, 2006. <http://orlando.cambridge.org/> (accessed March 4, 2015).
- Brunet E. *Manuel de référence pour Hyperbase 9.0*. Nice: Université de Nice Sophia-Antipolis, 2011.
- Brunet E. *Le vocabulaire de Jean Giraudoux: structure et évolution: statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la langue française*. Geneva: Slatkine, 1978.
- Brunet E. *Le vocabulaire de Proust*. Geneva: Slatkine, 1983.
- Brunet E. *Le vocabulaire français de 1789 à nos jours: d'après les données du Trésor de la langue française*. Geneva: Slatkine, 1981.
- Burrows JF. *Computation into Criticism: A Study of Jane Austen's Novels and an Experimental Method*. Oxford: Clarendon, 1987.
- Burrows JF. Never Say Always Again: Reflections on the Number Game. In: McCarty Willard, editor. *Text and Genre in Reconstruction*. Cambridge: Open Book, 2010: 13–36.
- Butler M. *Jane Austen and the War of Ideas*. Oxford: Clarendon, 1975.
- The Clockmaker's Outcry Against the Author of The Life and Opinions of Tristram Shandy*. London: Burd, 1760.
- Deconinck-Brossard F. Confessions d'une dix-huitième branchée. In: *XVII-XVIII* 1996; 42: 111–133. <https://doi.org/10.3406/xvii.1996.1326>.
- Defoe D. *Moll Flanders*. Oxford: Oxford University Press, 2011 [1722].
- Farrington MG, Farrington J. A Stylometric Analysis. In: Batestin M, editor. *New Essays by Henry Fielding: His Contributions to The Craftsman (1734–1739) and Other Early Journalism*. Charlottesville: University Press of Virginia, 1989: 549–591.
- Fielding, H. *The History of the Adventures of Joseph Andrews*. Oxford: Oxford University Press, 1986 [1742].
- Fischer-Starcke B. *Corpus Linguistics in Literary Analysis: Jane Austen and Her Contemporaries*. London: Continuum, 2010.
- Grant D. Roderick Random: Language as Projectile. In: Bold A, editor. *Smollett: Author of the First Distinction*. London: Vision, 1982: 129–147.
- Grant D. *Tobias Smollett: A Study in Style*. Manchester: Manchester University Press, 1977.
- Haywood EF. *Love in Excess, or the Fatal Enquiry*, ed. by Oakleaf D. Peterborough, ON: Broadview, 2000.

- Herrmann B, van Dalen-Oskam K, Schöch C. Revisiting Style, a Key Concept in Literary Studies. *Journal of Literary Theory* 2015; 9(1): 25–52.
- Hyperbase: Logiciel hypertexte pour le traitement documentaire et statistique des corpus textuels. Version 9. 2011. Université de Nice.
- Jockers M. *Macroanalysis: Digital Methods & Literary History*. Urbana, Chicago: University of Chicago Press, 2013.
- Kenny A. *A Stylometric Study of The New Testament*. Oxford: Clarendon, 1986.
- Lebart L, Salem A. *Statistique textuelle*. Paris: Dunod, 1994.
- Lévi Strauss C. *La pensée sauvage*. Paris: Plon, 1962.
- Mahlberg M. *Corpus Stylistics: Bridging the Gap between Linguistic and Literary Studies*. In: Hoey M, Mahlberg M, Stubbs M, Teubert W, editors. *Text, Discourse and Corpora: Theory and Analysis*. London: Continuum, 2007.
- Martinet M-M. *Digital Representation of City Cultural History: Feedback on the Twenty-year Long Interdisciplinary Experiment*. *ILCEA* 2020; 39. <https://doi.org/10.4000/ilcea.8645>.
- Mellet S. *Les tragédies de Sénèque vues à travers Hyperbase*. In: Mellet S, Vuillaume M, editors. *Mots chiffrés, mots déchiffrés: mélanges offerts à Etienne Brunet*. Paris: Champion, 1998: 255–271.
- Milic LT. *A Quantitative Approach to the Style of Jonathan Swift*. The Hague: Mouton: 1967a.
- Milic LT. *Information Theory and the Style of Tristram Shandy*. In: Cash A, Stedmond J, editors. *The Winged Skull: Papers from the Laurence Sterne Bicentenary Conference*. London: Methuen, 1967b.
- Moretti F. *Distant Reading*. London: Verso, 2013.
- Moretti F. *Graphs, Maps, Trees*. London: Verso, 2005.
- Rastier F. *Arts et Sciences du texte*. Paris: PUF, 2001.
- Rybicki J. *Vive la différence: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies*. *Digital Scholarship in the Humanities* 2016; 31(4): 746–761. <https://doi.org/10.1093/llc/fqv023>.
- Saxton KT, Boccicchio RP, editors. *The Passionate Fictions of Eliza Haywood: Essays on Her Life and Work*. Lexington: University Press of Kentucky, 2000.
- Sterne L. *A Sentimental Journey*. London: Becket & De Hondt, 1768.
- Sterne L. *The Life and Opinions of Tristram Shandy, Gentleman*. London: Dodsley, 1759–1767.
- Swift J. *Gulliver's Travels*. Oxford: Oxford University Press, 2008 [1726].
- Textsearch: A Full-Text Retrieval System for Humanities Research. Version 3.1. *LinguaTECH*, 1987.
- Tirvengadam V. *Linguistic Fingerprints and Literary Fraud*. *CHWP* 1998; A.9. <https://www.digitalstudies.org/article/id/7097/> (accessed April 4, 2022).
- Viprey J-M. *Dynamique du vocabulaire des Fleurs du Mal*. Paris: Champion, 1997.
- Watt I. *The Rise of the Novel*. Harmondsworth, Middlesex: Penguin, 1957.

Jan Christoph Meister

Poetry, Phenomenon and Phenomenology

Abstract: The phenomenology of the aesthetic artefact on the one hand and the methodology of digitally analyzing and modeling empirically observed phenomena on the other appear to be epistemological opposites. Each belongs to one or the other of the opposing “Two Worlds” (CP Snow) – the former to the subjectively experienced “world” that is the subject of the humanities and arts, the latter to the quantifiable and objective domain that is of relevance to the sciences.

I argue against this dualistic distinction. I do so as a digital humanist, and for three reasons: firstly, because from a philosophical-historical perspective, the so-called methodological divide turns out to be a discursive trope rather than a logical necessity; secondly, because attempts to bridge the gap between hermeneutic and quantitative, formalistic methods can be traced back to well before the “digital turn,” namely to the late eighteenth century; thirdly, because it is indeed possible to conceptualize and build tools for the digital analysis of aesthetic artefacts – notably for that of literary texts – in which hermeneutic and formal-analytical methods can be productively combined. One such tool is the annotation software CATMA, which I discuss briefly as “proof of concept” for a new practice of digital text analysis termed “scalable reading.” This epistemic practice explores the continuum between “close” and “distant” reading approaches in a methodical and controlled fashion that is designed to bring into fruitful contact the phenomenological and the phenomenal, i.e., the empirical perspective on poetic texts and poetry and their conceptualization. Setting such an ambitious objective has significant consequences for the design and development of software like CATMA, for it requires us to approach our task with a data and workflow model in mind that can map the fuzzy yet highly productive logic of the discursive hermeneutic circle onto the rigid base layer of non-ambiguous binary code.

1 Toward a digital hermeneutics of literature

What is poetry? Essentialist questions are tricky. To begin with, they imply that what we are trying to define is indeed a “thing,” a distinct phenomenon that exists and that can be objectively registered, described, and categorized. Moreover,

Jan Christoph Meister, University of Hamburg, e-mail: jan-c-meister@uni-hamburg.de

an essentialist approach makes it difficult to conceptualize “things” that, rather than being (real or ideal) objects that exist on their own, manifest themselves by way of processes in which we, the observers, are cognitively, emotionally, or indeed existentially invested. Poetry falls into the latter category, so let me rephrase my opening question in phenomenological terms: what is it that we *do* with poetry, and what does it do for us?

One answer is that in dealing with poetry – in reading texts, writing texts, in sharing and discussing them – we perform a symbolic practice using the code system of natural language.¹ However, in poetic communication, the familiar linguistic codes can be manipulated in surprising ways, some of them merely unusual, others outright ungrammatical or even mind-blowing. By way of convention, we take these deviations and irregularities to be intentional and thus meaningful rather than accidental. This intentional deviation from the norm utilizes what Jakobson (1960) termed the “poetic function” of language, one of six functional modes, which provides its users with the unique option to turn any of the other five functions of everyday signifying practice (or indeed of any combination thereof) into a self-referential reflection on language use “at the flick of a syllable,” so to speak.²

Considered from a philosophical rather than a semiotic perspective, this super-encoding of language aims at more than mere meta-representational navel gazing. Indeed, the poetic function invites us to transcend language as a given and to problematize that which comes before code, verbalization, and representation: our modes of experience. As a symbolic practice, poetry can thus point beyond language itself, to epistemic alternatives and consequences. The purpose of what the Russian Formalists termed *ostranenie* – the poetic estrangement of and from familiar forms of language use – is therefore not merely *l’art pour l’art*. Rather, it is of an epistemological order, namely that of *aesthesis*: the facilitation and exploration of “seeing” again through language. As Šklovskij postulated in his futurist-inspired manifesto “Resurrecting the Word,” words that are used in everyday speech become “symbols devoid of imagery” and thus “familiar,” but

1 Our use of natural language is but one of many domains of symbolic practice, and so it follows that poetry, from a logical perspective, is not *necessarily* defined as a linguistic utterance. In principle we might just as well “speak” poetically in, say, symbolic gestures, or in computer code. This begs the question of why exactly natural language has become the predominant cultural choice of code for the manifestation of poetic practice. One possible answer is that the linguistic code, because of its high level of abstraction from the iconic, comes with the most liberal license to deviate from pragmatic and substance-centered modes of reference. In other words, we are accustomed to using it in a referential, and in a non- or auto-referential manner (a functional variety not an established practice for, e.g., the musical code system.)

2 On Jakobson’s concept of the *poetic function* of language, see Waugh, 1980.

these familiar words have lost their experiential quality: “[. . .] neither their internal forms (images), nor the external one (sounds) are experienced anymore. We do not experience the familiar; we do not see but only recognize it” (Šklovskij, 1914: 64).

Here is an example of how a poem – that is, an example taken from the subclass of texts whose poeticity tends to be most prominently marked – can invite us to transcend the referential, the “recognizing” mode, and explore the epistemological dimension, that of “seeing” via the linguistic code:³

the great advantage of being alive
 (instead of undying)is not so much
 that mind no more can disprove than prove
 what heart may feel and soul may touch
 – the great(my darling)happens to be
 that love are in we, that love are in we

and here is a secret they never will share
 for whom create is less than have
 or one time one than when times where-
 that we are in love, that we are in love:
 with us they've nothing times nothing to do
 (for love are in we am in i are in you)

this world (as timorous itsters all
 to call their cowardice quite agree)
 shall never discover our touch and feel
 – for love are in we are in love are in we;
 for you are and i am and we are (above
 and under all possible worlds) in love

a billion brains may coax undeath
 from fancied fact and spaceful time –
 no heart can leap,no soul can breathe
 but by the sizeless truth of a dream
 whose sleep is the sky and the earth and the sea.
 For love are in you am in i are in we.

EE Cummings (2015)

³ In the following I will focus on this subclass, i.e., on what we traditionally refer to as “poems” in the emphatic sense. I am consciously trying to avoid the pitfalls of a definition in terms of traditional genre concepts here (dramatic, lyric, epic) because the distinction is in fact irrelevant to my argument.

In Cummings's poem, at least three types of irregular language encoding challenge the reader: deviation from standard English syntax and word order ("love are in we"), deviation from orthography (lower-case "I"), and irregular use of punctuation and diacritics (no space before and after brackets). These three markers of poetic language use could, of course, be easily detected by taking a "mechanically enhanced reading" approach. Indeed, based on the Formalists' definition of poetry, one could sketch out the system architecture for a powerful computational "poetry detector" that would (a) diachronically harvest and mine all digitized primary texts and their interpretations in world literature for documented correlations between patterns of non-standard language use in the former and related interpretive attributions (*innovative, trivial, meaningful*, etc.) in the latter; then (b) measure the level of standard deviation of a given text string in terms of a defined set of code registers (lexical, grammatical, rhetorical, prosodic, thematic, genre-typological, etc., etc.) at a defined point in time; and in the background (c) continuously update the normative standard distributions across all those registers after every detection run. Moreover, this dynamic detection system could be implemented in an unsupervised machine learning architecture and could compute, for every candidate primary text, a "poeticity coefficient" that would be, on the one hand, contingent on how a user decided to set the norm values for various parameters (epoch, genre, reception time, scope of comparative cultural and literary context to be considered, etc.) and, on the other, on how much the system has already learned about the development of "poeticity" across a defined period of time.

This hypothetical digital humanities (DH) approach to poeticity presupposes what Moretti (2013) has described as "operationalization," i.e., making the object we wish to investigate – in our case: poetry – measurable and manipulable with our tools. Two requirements must be met for this: firstly, we have to gain a clear understanding of what it is that we are looking at and for in poetic texts, and, secondly, we need to be able to define these *phenomena* in the stringent and exhaustive logical form required in order to model and analyze them by way of computer algorithms. However, as difficult as this might seem, it is still merely a pragmatic constraint, and so for the sake of the argument, let us simply assume that it was already met. We would now face a second problem, this time a conceptual one: does our formal, computationally "actionable" representation of the object come close to the poetic text as we experience it *phenomenologically*? Would our approach enable us to capture the "sensation of an object," which according to Šklovskij is what the poetic text actually aims to mediate?

It is doubtful. For the sensation of an object is not merely defined in terms of how something presents itself objectively to our senses so that they may then register it as a phenomenon, as an "observable fact." Rather, we have to proceed

from the phenomenon to its phenomenology, to the experience of a poem in terms of an encounter with a noumenon, with a “mind entity” that takes on the form of a “structure(s) of consciousness as experienced from the first-person point of view” (Smith, 2018). This experientiality comes over and above the perception and recognition of an object per se: we are now in the territory of the subjective phenomenological encounter. And if we take seriously the DH claim to further humanistic research methods, then it is this encounter that we need to operationalize as well, and not just the recognition of an external object.

This is a tall order: can a computational approach bring us in contact with the poem’s “secret they never will share / for whom create is less than have / or one times one than when times where”? Is not the digital approach to artistic phenomena one that more or less epitomizes how “a billion brains may coax undeath / from fancied fact and spaceful time”? It is this methodological dilemma upon which I would like to deliberate in the following.

2 The trope of the methodological divide

The DH attempt to *machiner la poésie* is indeed an undertaking that is ambitious on a conceptual level, for we are trying to bring into fruitful contact the phenomenological and the partially empirical, partially abstract, and theoretical perspective on symbolic artefacts such as poems, novels, paintings, plays, films, architecture, music, etc. In other words, we aim to combine qualitative and quantitative conceptualizations and models from the same object domain. This venture can only be successful if we step away from deliberating technological and pragmatic questions for a moment and focus at first on the more fundamental methodological challenge that we face.

Probably the biggest hindrance to be overcome at the outset is what I would like to term the *trope of the methodological divide*. It postulates a principled methodological and epistemological distinction that separates the humanities and the sciences as two opposing domains of knowledge, experience, and practice – indeed, as distinct *cultures*, as CP Snow (1993) so provocatively claimed in his famous 1959 *Reed Lecture*. However, upon closer inspection it turns out that Snow’s polemic itself was in fact an attempt to provoke his contemporaries and, in particular, the “men of letters” – the humanists and artists – to reengage with the other side, the modern sciences and mathematics. The intention of the *two cultures* polemic was thus by no means to deepen the divide but rather to call it into question. And, in doing so, Snow made it quite clear that, as far as he was

concerned, the onus was primarily on the humanists, in other (my) words, on the advocates of the phenomenological to reengage in dialogue with the other side.

However, before we take up this challenge and deliberate upon how to cross our present divide – the one between qualitative and quantitative methods of textual study and, in particular, of poetry – let us recall how the concept, the imagery, and indeed the ideology of a divided methodological terrain, of which Snow's *two cultures* polemic is but one example, has come about and shaped our thought and our disciplines. Why is it so prevalent?

For Western cultures, the idea of conceptualizing epistemological domains in terms of distinct “worlds” originated in classical antiquity. According to Plato, our mind has two principal faculties: that of mental *anamnesis*, intellectually *remembering* that which is true and eternal, that which exists in the realm of *ideas*; and the faculty of *empeiria*, the ability to sensually perceive that which is fleeting and impermanent, that which exists in the world of *things* (Hager, 2017). From Plato's philosophical perspective, the *ontological* distinction between things and ideas thus plays out as an *epistemological* distinction between two ways of acquiring knowledge about things and ideas (Manuwald, 2009).

Of course, not everyone will subscribe to Plato's idealist premise, to his axiom that ideas, which are not bound to matter and its constant transformation, are *more* real than things. But whether you are a materialist or an idealist, the postulate of conceptual mapping between ontology and epistemology itself holds true: it is indeed a corner stone of all rational philosophy and science. For Plato goes beyond merely positing distinct “worlds” or realms of reality, such as the secular vs. the holy, or the permanent vs. the impermanent. Rather, he points to the fundamental *methodological* consequence of any ontological dualism. As soon as we decide to conceptualize the world in terms of distinct domains – such as real vs. ideal, or logical vs. phenomenological – we must also be prepared to accept that the mode of acquiring rational knowledge about any of these domains will necessarily be domain-specific.

Time and again philosophers have made recourse to this finding and pointed out that it simply does not make sense to try to acquire knowledge – that is: to find out systematically what the fact is and what it is not with regard to the domain under investigation – by applying a conceptually inadequate method of enquiry. And this methodological constraint is the ultimate limitation on rationality itself. The most prominent example: it is logically impossible to prove or disprove the existence of God via rational discourse, as Immanuel Kant demonstrated in his 1781 *Critique of Pure Reason*. God is a phenomenological reality, not a measurable phenomenon.

Plato's appeal to methodological rigor, however, did not go uncontested – there is also a tradition of argumentation to the contrary. Kant, for one, provoked

this response himself. As soon as eighteenth-century Romantics had absorbed his philosophy, the vision of bringing the distinct methodologies into a mutually beneficial and productive relationship emerged again. In 1805, the German writer Heinrich von Kleist thus remarked in a letter to his friend Ernst von Pfuel: “I can solve differential equations and I can write verses: are these not the extreme limits of the human potential?” (Kleist, 1982: 160). Against the backdrop of this all but modest self-advertisement Kleist claimed to be part of an avant-garde that possessed aesthetic as well as mathematical, abstractive competencies, a combination of intellectual abilities which he declared to be highly desirable. This was a vision certainly more fundamental than that of *machiner la poésie* – for the goal was no less than that of methodologically reintegrating the two conceptual paradigms that had been outlined as distinct during the latter half of the eighteenth century.

3 Hermeneutics as a methodological definiens of the humanities

However, the nineteenth century would frustrate those ambitions. As the modern natural sciences began their ascent, the domains of metaphor and formula drifted further and further apart, and the Renaissance ideal of a coherent *universitas* of disciplines and methods lost its appeal. Disciplines now became increasingly defined not only by their respective object domains but even more so by their repertoires and types of methods, and more fundamentally by how they conceptualized their domains. One such method that defined the humanities in particular was hermeneutics. Discipline-specific practices of methodical text interpretation had of course already emerged, in particular in biblical and juridical exegesis. But it was only in 1808 that the classicist Friedrich Ast coined the term *Hermeneutik* to describe the method as such. The term was in part descriptive and in part programmatic; it designated and called for the precise codification of textual interpretation as a scholarly method practiced by humanist scholars. One such codification that became particularly important was presented by Friedrich Schleiermacher in his 1838 *Hermeneutik und Kritik (Hermeneutics and Criticism)*, which combined theological and philosophical reasoning.

But outside theological and classicist discourse, the impact of hermeneutics on contemporary philological practice remained marginal: at the beginning of the nineteenth century, most of the evolving modern philologies were more concerned with collecting and editing language-specific bodies of texts than with methodically interpreting individual works or oeuvres. The explication of the

meaning of a literary text remained the prerogative of commentators and critics, who were able to draw on personal intuition and inspiration rather than making recourse to systematized knowledge and methods. Against this backdrop, the meteoric rise of the modern natural sciences demonstrated the exact opposite: for the new sciences were defined not only by their respective subject domains but more fundamentally by their repertoire of experimental and empirical methods and their mathematical approaches toward building and testing domain-specific theories. The “old” sciences, and the liberal arts in particular, of course, did have their own methods and terminologies, but they lacked a comparable and distinctive methodological self-awareness.

It was only in 1900, when Wilhelm Dilthey published his study *Die Entstehung der Hermeneutik* (*The Rise of Hermeneutics*), that this lacuna was resolved. Dilthey tried to demonstrate how the historical development of one particular scholarly method of text interpretation – *hermeneutics* – had in fact brought about the cluster of historical disciplines that we now refer to as the humanities. And, as in Plato, the ontological and the epistemological argument went hand in hand once more. Turning a weakness into a strength, Dilthey wrote:

Human sciences have indeed the advantage over the natural sciences that their object is not sensory appearance as such, no mere reflection of reality within consciousness, but is rather first and foremost an inner reality, a nexus experienced from within. (Dilthey, 2018: 317–318)

Dilthey was acutely aware that the immediacy of this inner reality, this experience that need not be mediated by the senses, also poses a particular epistemic problem. How is it possible to gain *objective* knowledge about that reality if my point of view is unavoidably subjective? Dilthey’s solution to the problem was a detour via intersubjectivity: in order to truly understand myself, my own inner reality, I must by necessity compare my self-experience with that of other individuals, and vice versa. The task, therefore, is to define a procedure by which I can objectively understand your inner reality. On this point Dilthey states the following:

But the existence of other people is given us at first only from the outside, in facts available to sense, that is, in gestures, sounds, and actions. Only through a process of re-creation of that which is available to the senses do we complete this inner experience. Everything – material, structure, the most individual traits of such a completion – must be carried over from our own sense of life. Thus the problem is: How can one quite individually structured consciousness bring an alien individuality of a completely different type to objective knowledge through such re-creation? What kind of process is this, in appearance so different from the other modes of conceptual knowledge?

Understanding is what we call this process by which an inside is conferred on a complex of external sensory signs. [. . .] Such understanding ranges from grasping the babblings

of children to *Hamlet* or the *Critique of Pure Reason*. Through stone and marble, musical notes, gestures, words, and texts, actions, economic regulations and constitutions, the same human spirit addresses us and demands interpretation. (Dilthey, 2018: 318–319)

What are the methodological consequences? Let us assume that my goal is to understand your “inner reality.” In order to do so in the procedurally correct hermeneutic fashion, I have to combine two operations, *exegesis* and *synthesis*:

- *Exegesis* – I will interpret as a manifestation of your inner reality every individual “sensory sign” by which that reality objectifies itself and becomes visible to me on the outside. Dilthey suggests that for pragmatic purposes these “signs” should ideally be “fixed and relatively permanent objectifications of life” (Dilthey, 2018: 319). The paradigmatic form of such an expression is the fixated linguistic utterance, i.e., the text.
- *Synthesis* – I will then proceed to combine all these individual “signs,” which I have interpreted as individual expressions of your inner reality, in order to synthetically reconstruct your complex inner experience of reality as a phenomenal whole. However, the construction of that phenomenal whole is not merely additive: it takes place under certain constraints, some of which are logical (such as: non-contradiction) and others contextual (such as: historical plausibility).

From the perspective of the philosophy of science, one can thus read Dilthey’s *The Rise of Hermeneutics* as an attempt to provide the humanities, and in particular the philologies, with a methodological foundation *ex post*. At the same time, Dilthey’s appeal to a practice of methodological rigor in textual and historical understanding also counters the claim for epistemological hegemony made by late nineteenth-century positivism by ascertaining the legitimacy of the historical disciplines and of their focus on the subject-centered interpretation of texts as “signs” of a lived experience of reality.

What we refer to as the *qualitative* approach today across various disciplines in the humanities and social sciences was indeed conceptually prefigured in Schleiermacher’s, Dilthey’s, and Gadamer’s notions of *Hermeneutik*, and in an epistemology based on the premise that the phenomenon under investigation – the text, the image, the gesture – is merely the outer sign of an inner phenomenological experience in somebody’s mind. The nexus between that mind, its subjective inner experience, and the objective outer sign is, then, what we try to explore in a qualitative study.

4 Methodological trajectories: The example of folklore studies

And so, the methodological dividing line seems to be clearly defined once again: where the domain of the inner experience of reality is concerned, hermeneutics is the appropriate method; where the outer world is at stake, empiricism will be the candidate of choice. But this conceptual division between terrain and method had in reality already become obsolete when Dilthey wrote *The Rise of Hermeneutics*. For by the end of the nineteenth century, positivism's reach extended not only into the field of artistic creation, to which, e.g., the Realist and Naturalist literary movements testified; positivism and empirical methods had also begun to affect humanities scholarship, and philology in particular.

The development of folklore studies between 1810 and 1927 illustrates this development. When the Grimms compiled their famous collection of *Kinder- und Hausmärchen* from 1812 to 1858, their project aimed for an anthology inspired by the Romantics' mythological interest in sagas and folk tales as an expression of what Dilthey would later term the "inner reality" of a culture or people. However, from the 1880s onward, the focus in folklore studies became more precisely defined in methodological terms by what was referred to as the Finnish School and its historical-geographical method of study (Frog, 2013). Rather than merely collecting narratives, the Finnish School foregrounded the genetic aspect and employed analytical and empirical approaches in order to retrace how and where narratives had originated, and how they had travelled across cultures. At the beginning of the twentieth century these genetic studies eventually culminated in Antti Aarne's catalogue *Verzeichnis der Märchentypen mit Hilfe von Fachgenossen* (1910), which was derived from various collections of folk tales, including that of the Grimms. Taking a similar approach, the American folklorist Stith Thompson collected and published his six-volume *Motif-Index of Folk-Literature* (1932–1936), which was then cross-indexed with the second, expanded English edition of Aarne's catalogue and jointly published by the two folklorists as *The Types of the Folktale: A Classification and Bibliography* (Aarne and Thompson, 1961). Hans-Jörg Uther (2004) eventually contributed yet another set of tale types, resulting in the ATU motif register, short for *The Types of International Folktales: A Classification and Bibliography; Based on the System of Antti Aarne and Stith Thompson*.

The ATU now lists close to 5,000 (!) types of folktales and is thus, on the one hand, an impressive example of how fruitful a rigorous, long-term, systematic collection endeavor in the study of oral and literary history can be. Yet at the same time, it also demonstrates its limitations: for the ATU is indeed a *catalogue* – not a *typology*. Like an encyclopedia, it is restricted to listing all findings and

putting them into their historical and genetic context. But it does not aim to provide any substantial theoretical insights into the logic of the phenomena sampled; indeed, one might even argue that it consciously avoids stating a clear-cut definition of what a folk narrative or a motif is from the outset. For example, in his 1946 book *The Folktale*, Thompson himself merely observes that a “*motif* is the smallest element in a tale having a power to persist in tradition. In order to have this power it must have something unusual and striking about it” (Thompson, 1977: 415). Such reliance on a broad definition of the phenomenon sought is characteristic of explorative, bottom-up collection efforts, and it makes perfect sense here: for theoretical over-specification is counter-productive during the first empirical phase of enquiry. During that phase, we are more likely to settle for a pragmatic trade-off between how exact our finds are and how many of them we can make – in other words, between *precision* and *recall*, as we would put it in today’s data mining parlance.

The Russian Formalist Vladimir Propp took the exact opposite approach. In his 1928 *Morphology of the Folktale* (1968), Propp presented a 31-function formula of one particular type of Russian fairy-tale. This formula, he claimed, had been abstracted from a sample of 100 such narratives that he had analyzed in terms of a functional narrative model. Unlike his predecessors, who had focused on collecting folk tales and on comparing them on a content level, Propp’s functional approach did not describe the phenomenon of the fairy-tale from an experiential (phenomenological) perspective. Rather, Propp analyzed every tale against the backdrop of a theoretical model that postulated that a narrative should be considered top-down as a complex functional whole of agents, patients, and transformative events, and not bottom-up as a mimetic representation of individual character-centered, psychologically motivated actions that eventually converged into a “story.” This, as Claude Levy-Strauss, Claude Bremond, Thomas Pavel, and others later observed, was the actual birth of the formalist and structuralist approach to narratives and literature.

The examples of the Grimms and the development of both the ATU catalogue and Propp’s functional approach are all drawn from a subdiscipline that played a central role in the development of the modern humanities at large. In the late eighteenth and early nineteenth centuries, folklore studies were instrumental in the Romantic ideology of language-based national identity, of which it was deemed to provide empirical historical evidence. In the course of some 80 years, this field then progressed from the encyclopedic to the taxonomic to the formal approach, following a parallel methodological trajectory to that taken by empiricism to positivism, which reshaped contemporary natural and the social sciences. And one might easily retrace comparable progressions from, say, 18th-century *Sprachgeschichte* (historical linguistics) up to early-20th Saussurean linguistics, or

from the early eighteenth-century *Erfahrungsseelenkunde* (the nascent field of empirical psychology) to Freud's late nineteenth-century psychoanalytical theory. Indeed, by the beginning of the twentieth century at the latest, systematization, abstraction, and formalization had impacted many humanities disciplines. In short: the methodological shift from phenomenological to empirical to theory-based approaches and conceptualizations of the domain under investigation was by no means a discerning feature of the other, of scientific culture.

5 Mining the gap: Transgression vs. transformation

By retracing the methodological trajectory of these humanities disciplines from the early nineteenth to the beginning of the twentieth century, we learn two things: firstly, that the “methodological divide” between the sciences and the humanities is indeed a cultural construct and a theorem that may have served a useful purpose in its original historical context – but at the beginning of the twenty-first century, we should be able to appreciate it for what it has become: a mere discursive trope. Secondly and more importantly, adopting this historical perspective enables us to gain a better understanding of the constraints but also of the benefits of attempts to engage in inter-disciplinary and inter-methodological exchange between the phenomenon-based and mathematical-modeling-based disciplines on the one hand and, on the other, those disciplines whose primary focus is the phenomenological impact of issues, ideas, and discourses on the human mind. This methodological distinction is similar to, but not exactly the same as, the traditional sciences vs. humanities dichotomy, for appreciating a specific individual discipline's methodological profile in relation to another can no longer be a matter of drawing clear lines. Rather, we should think of our undertaking as an attempt to register similarities and differences in terms of a Wittgensteinian feature matrix, which can then help us to identify possible family resemblances between disciplines. Instead of neurotically “minding” the gap, we will then be able to “mine” it and look out for interesting points of contact that have been relegated to the abyss.

In this regard, Propp's 31-function formula of the Russian fairy-tale presents an interesting case once again. In 1975, when his book had finally been translated into English and published in the US, his formula was discovered by the AI researcher David Rumelhart. In an article titled “Notes on a Schema for Stories,” Rumelhart drew extensively on Propp and sketched out a story grammar that he

claimed would in principle enable artificial intelligence to *generate* stories (Rumelhart, 1975). His article was visionary, and it served to make Propp a household name in AI. To this day, AI researchers interested in story generation continue to refer to Propp's formula as a conceptual blueprint (see, e.g., Gervas, Peinado, 2006).

Yet the bold AI adaptation of Propp's model is based on a shortcoming which, from an orthodox hermeneutic perspective, boils down to a fundamental procedural mistake: namely of ignoring the historical and methodological context of an original text in the course of its appropriation. As for historical context, Propp himself had been at pains to mark it not only by his choice of title – "*Morphology of the Folk Tale*" – but also by way of including a motto taken from Goethe's 1817 publication titled "On Morphology."⁴ In the passage that Propp quotes, Goethe discusses the pros and cons of morphology as a science, and eventually concludes: "Its arrangement of phenomena calls upon activities of the mind so in harmony with human nature, and so pleasant, that even failures may prove both useful and charming" (Goethe, 1988: 60). If nothing else, Propp's use of the motto thus served the rhetorical purpose of an *exculpation* – his formalist undertaking was self-critically advertised as an approach that by necessity is of a speculative and self-reflective nature rather than an ideal mathematical abstraction from reality.⁵ Rumelhart's seminal article, however, was perfectly agnostic toward this historical and philosophical context – a fault that we can only partially hold against him, for the publishers of the first American translation of Propp's book had in fact omitted the Goethean motto. And thus, in the absence of this contextual information, Rumelhart and many other researchers with an interest in AI story generation had *adapted* Propp's formula – but they had not *understood* it in an emphatic, hermeneutic sense.

Be this as it may, the adaptation proved extremely fruitful for the (then) emerging field of AI. If nothing else, its example demonstrates one thing: postulating the dos and don'ts of inter-disciplinary and inter-methodological exchange in a normative fashion from the perspective of either side will lead to nothing. As

⁴ One should note that the full German title of Goethe's publication "Zur Naturwissenschaft überhaupt, besonders zur Morphologie: Erfahrung, Betrachtung, Folgerung, durch Lebensereignisse verbunden," emphasizes the experiential and subjective dimension of the empirical process. Literally translated, the title announces deliberations "On the Natural Sciences in General, in Particular on Morphology: Experience, Observation, Conclusion, as Connected Through Life Events" (my translation; JCM). Propp quotes, in slightly redacted form, from the section "On Morphology" published by Goethe in 1817 (Goethe, 1994: 127).

⁵ Propp was probably also concerned about the emerging Stalinist critique of Russian Formalism and hoped that this caveat might help to defend him against the potential accusation of furthering "revisionism."

long as we think of disciplines in terms of strictly delineated conceptual and methodological terrains separated by a “gap,” any adaptation of one discipline’s concepts, theorems, methods by another will always be deemed metaphorical appropriation at best and sacrilege at worst. However, if we imagine the diversity of the scientific endeavor at large in terms of a methodological continuum and posit individual disciplines in relation to its axes, the picture changes. It is not the *transgression* across domains that is foregrounded; what matters now is the necessary *transformation* of the individual concept, theorem, model, etc. taken from one discipline and moved along the methodological continuum to another discipline.

To realize, reflect upon, and perform this operation as one that is necessarily transformative is the essential task and ethos of truly interdisciplinary ventures like the digital humanities. And this is the philosophical backdrop to one such transformative methodological encounter that I will now discuss. In this particular case, the project concerns the development of a richer concept of text annotation, which aims to combine a taxonomic, phenomenon-centered declarative method of linguistic provenance with a phenomenologically oriented approach rooted in the philological practice of textual commentary.

6 The paradox of top-down markup: From “structured annotation” to structuring annotations

As recent corpus-based digital studies have amply demonstrated, the “distant reading” of large text corpora using DH methods can now produce highly interesting insights into the development of historical forms, genres, styles, themes, etc. over time. But it is a different type of knowledge altogether: the traditional historical and hermeneutic approach in literary studies was characterized by the “close reading”-based, in-depth analysis, and interpretation of a small set of texts often taken from an established, normative literary canon. Most of the new digital methods trade the hermeneutic precision and ingenuity of this “closeness” to the exemplary text for the statistical generalizability of results obtained from the perspective of methodological “distance” between observer and object domain.

The new methods’ disinterest in the qualitative, singular, and highly original text is, however, only in part a conscious methodological choice. The analysis of very large text corpora is equally subject to pragmatic constraints. We simply do

not have the time or resources to process large numbers of texts “manually,” or rather: intellectually – that is, by way of processes of analysis, categorization, interpretation, critique, and validation, etc. controlled by human intelligence. In response to these pragmatic constraints, digital corpus analysis therefore relies heavily on data mining techniques, natural language processing technology, and statistical pattern analysis. The more these methods can ignore the contextually induced specificity of the phenomena under investigation – such as the historicity of a particular language variant, the genre specificity of a topic, etc. – the more powerful they become.

On the other hand, the exploration of context and its impact on meaning is and remains the forte of hermeneutic engagement with objects: here, it is precisely the specificity of an individual constellation that counts and not its generic traits. Rather than differences in the scopes of object domains – i.e., corpus size – it is the fundamentally different conceptualization of the role played by context-dependency that is at the bottom of the methodological “gap” between close and distant reading approaches. While context-dependency is epistemically productive in close reading, distant reading is all about abstracting from the individual and its singular contextual conditioning.

One way to resolve what at first sight may look like a principled methodological impasse between the “close” and hermeneutic, and the “distant” and statistical methods of text studies is to enrich digital text analysis with one of the oldest and probably the most powerful method of traditional text studies: annotation. As Burnard already stated, “Text markup is currently the best tool at our disposal for ensuring that the hermeneutic circle continues to turn, that our cultural tradition endures” (Burnard, 1998).

Digital text markup is now practiced in many different forms. It spans the entire range of methods, from declarative operations that are strictly taxonomy- and rule-driven – like, for example, POS tagging – and thus lend themselves to automation, to extremely context-dependent and often rather “fuzzy” interpretive operations that resist formalization, as Bauer and Zirker (2013) demonstrate. Because of this functional variance, we can, on the one hand, conceptualize markup as a traditional philological and linguistic practice that has been successfully transformed into a DH practice in corpus-based, computer-assisted, and ideally fully automated research routines. But on the other hand, the more it becomes functionally equivalent to the commentary and *exegesis* of a given digital text, the more markup can also be considered a prime candidate for demonstrating the demands and constraints to be met if we want to support qualitative research into texts and corpora by digital means. And this becomes much easier if we consider the distinction between quantitative and qualitative approaches not as one between opposing paradigms but rather as one between two ideal types

that, in reality, is mapped onto a continuum of different methods used for different purposes. These methods may tend to blend on occasion, for example, when linguistic markup encounters grammatical polyvalence in a term that can only be resolved by making recourse to its semantic classification – an operation that is essentially interpretive.

As more and more digital text corpora have become available over the past 20 years, so too have digital annotation tools and platforms. In a recent survey titled “An Extensive Review of Tools for Manual Annotation of Documents,” Neves and Ševa (2019) evaluated 78 such tools and then selected 15 web-based annotation systems (referred to as *web services*) for a more detailed evaluation based on 26 criteria. Similar to Bauer and Zirker (2013), the authors note that annotation as such

[. . .] can vary from an unstructured short piece of text, such as a comment on a text passage, as supported by the Hypothesis tool, to structured annotations by means of highlighting text spans or drawing relations between them. (Neves, Ševa, 2021: 147)

Neves and Ševa’s survey and evaluation of annotation tools is the most comprehensive to date. Although it is impressive in scope and methodologically sound, the study has been somewhat calibrated to the authors’ own disciplinary background and research interest in bioinformatics. The prospects of applying machine-learning-based natural language processing (NLP) routines for, e.g., the automated extraction of terms from research publications plays a particular role in this regard. This also explains why the evaluation metric is aligned with the “structured annotation” method, i.e., an approach toward applying markup to digital texts focused on rules and taxonomy / classification ontology.⁶ The authors justify their preference as follows:

The advantage of structured annotations are [sic!] various, such as their straightforward use for machine learning purposes, the possibility of computing statistics, as well as direct comparison between the various annotators and their agreement. Further, annotations can be either enforced by (strict) predefined guidelines or performed in a relaxed way, without predefined guidelines. The former is always preferable since it enforces homogeneity across the various documents and annotators. (Neves and Ševa 2021: 147)

⁶ This conceptual bias does not necessarily invalidate the survey results, but it does of course impact on the results in that at least two of the criteria will automatically favor annotation tools that were built to suit the NLP/ML paradigm. These criteria are: *F4 – Support for ontologies and terminologies*; *F11 – Support for inter-annotator agreement (IAA)*. In addition, criterion *F6 – Integration with PubMed* is obviously intended to help identify particularly useful applications for bioinformatics research.

The concept epitomizing this methodological preference for “structured” annotation is *inter-annotator agreement* (IAA). And it is indeed a *concept* and not merely a technical term for the statistical measurement of annotation variance across annotator subjects, as we will see in a moment.

Arguments in favor of enforcing or optimizing IAA generally tend to be of a pragmatic nature: for example, a high level of inter-annotator disagreement will normally render a manually annotated text corpus unsuitable as a training corpus for the purposes of taking a supervised machine learning approach where the performance of an algorithm is measured against the human-generated annotation norm, and where algorithms are iteratively optimized with a view to (re-)producing human “gold standard” annotation results as well as possible. Considered conceptually rather than merely from a pragmatic point of view, this reasoning is, however, already based on a fundamental choice of research paradigm. Central to this paradigm is the imperative of the replicability of experimental results, provided that such experiments were conducted under sufficiently controlled conditions. This is, of course, the normative ideal of the experimental sciences, which originated in the nineteenth century and would be adopted by most modern disciplines devoted to the study of nature and human social behavior. The relevance of that approach and ideal to contemporary studies in the humanities should not be denied as a matter of principle. But this has only recently begun to have an impact on the practices of the core humanistic disciplines – i.e., philosophy, philology, history, theology, and the liberal arts in general – due to the advent of data-analytics-based methods of exploration, hypothesis generation, and theory verification. And so, while these empirical, quasi-experimental approaches may be strong candidates for testing, augmenting, and objectifying the findings that have been produced by this set of disciplines that, for the past 200 years, has relied heavily on the hermeneutic method, empirical approaches cannot for the time being (if ever) replace hermeneutics’ *speculative* reasoning power.

And here, too, the limitation is a matter of pragmatics inasmuch as it is a consequence of the epistemic and philosophical orientation fundamental to the (traditional) humanities. The pragmatic constraint is one of complexity and dynamics: we encounter culture and symbolic practices such as literature, the visual arts, the performing arts, etc. phenomenologically – that is, as ongoing historical processes from which it is difficult to abstract. To “freeze” this complex of processes and model it, part by part and sector by sector, in a set of well-defined, experimental settings would be an extremely ambitious interdisciplinary venture, something akin to a “cultural genome” project. At the same time, pretenses held by some schools of thought about having achieved the dispassionate stance of a neutral “scientific” observer, at least for a particular subdomain (e.g., structuralism’s claim to the scientification of literary studies) have regularly been unhinged as

soon as the next such school (say, deconstruction) has staked its claim. In the end, studying culture, whether in part or as a whole, is not like studying rock formations in geomorphology or the properties of fluids in physics. We are too close to culture as observers, both temporally and spatially, and too deeply embedded within it existentially to be able to fully practice an ideologically disinvested, distant reading of it – yet. For as more and more cultural artefacts and practices are being mediated digitally, and as the collections and corpora grow in size and number, new opportunities for exploring the power of calculated mathematical alienation from the phenomenal may well materialize.

Indeed, the epistemic and philosophical aspect of this problematic relates to a more fundamental issue, namely the particular notion of knowledge that informs the humanities. While there are, of course, indisputable, conventionalized, domain-specific facts that we might state, exchange, learn, and argue about, our main objective as humanists is not producing, disseminating, or acquiring this base-level factual knowledge. Such knowledge, when formulated and codified in terms of propositional statements, is essentially referential – it concerns statements about “facts” claimed to hold true for a specific object domain. We prove and disprove them by checking whether they are empirically verifiable, i.e., in terms of a quantifier – and where they are, this is then deemed “knowledge” worth attaining. However, in the humanities this referential knowledge is just the base layer, which must be complemented by a second type of knowledge that is defined in terms of an agential qualifier. Instead of asking whether or not unicorns exist in a given world (factual knowledge), the question turns into what kind of relevance someone’s assumption that unicorns exist in a particular world may have for his or her worldview. This, then, is speculative knowledge production. It takes place in a type of conditional “thought experiment” that is essentially self-reflective – for it is not the referential truth value of the proposition itself that counts but its epistemic relevance for whomever is conducting that experiment that matters.

This then raises the question of how to prevent speculative knowledge production from turning into idiosyncratic intellectual narcissism. The answer is simple: we test speculative knowledge, but not by way of a controlled experiment that tries to model a historical real-world context. Humanists use a different verification mechanism – critical discourse – in other words: the exchange of rational, transparent pro and con arguments with the shared goal of furthering knowledge and understanding.

To date, annotation tools built on this philosophical basis and with these methodological considerations in mind have been few and far between: for how can one operationalize the discursive validation of competing “statements of fact” about a text that are encoded as digital markup? To begin with, we would

need an annotation tool that allows us to capture conflicting markup, rather than one that enforces IAA by way of an annotation protocol or by running some sort of “syntax checker” that will automatically eliminate any potential markup ambiguity during the annotation process. Rather than taking recourse to a normative, hierarchical classification ontology, one would probably apply a more flexible, partially non-hierarchical classification system.⁷ On that basis, we could then proceed to add functionalities that would assist us in the explorative, discursive, bottom-up structuration and successive formalization of this markup. Overall, such a tool would methodologically complement those that are by design restricted to supporting through markup what is essentially the top-down application of a predefined, “structured” annotation schema. From a hermeneutic perspective, such complementation is indeed essential: for limiting the enrichment of a document with annotated information, through digital markup, to the mere *declaration* and *classification* of words or sentences in terms of a predefined taxonomy – a structured annotation schema – goes against the grain of truly *interpretive* procedures. To interpret a text hermeneutically means to expose it to a context. Conversely, if markup is restricted to the practice of top-down declaration in terms of a defined epistemic context, then no new “knowledge” (in the emphatic humanist sense elaborated upon above) can be created. Rather, we will end up making declarative statements that circle back onto the epistemic and theoretical paradigm whose offspring they are. And so, for interpretive practices, the normative combination of “markup” and “top-down” results in a paradox that is not just terminological but epistemological too. For philological, hermeneutic text annotation we need a tool that can handle both: declaration and interpretation, structured and schema-compliant, top-down classification, as well as unstructured, “bottom-up” discursive exploration via annotations that might at some point result in a new structure and schema – but always with the goal of handing this structure over to further critique and refinement.

7 Concept ontologies, however, are not necessarily hierarchical – one of the most promising options for a more flexible ontology concept is the “lattice structure” proposed by McQueen and Huitfeldt at the 3rd Expert Workshop of the forTEXT project, January, 24–25, 2020, on “Non-hierarchical concept ontologies and markup schemas” (Jacke, 2020).

7 Building CATMA as a hermeneutic annotation platform

Developing such an annotation tool has been the goal of the CATMA project that my team and I started in 2008. CATMA – short for *Computer Assisted Text Markup and Analysis* – builds on John Bradley’s DOS-program suite TACT (*Textual Analysis Computing Tools*) from the mid-1980s, one of the first desktop applications for digital text annotation. TACT soon became popular in the new field of humanities computing that had just begun to evolve at the time. Today, its successor CATMA is an open-source, collaborative digital text annotation and analysis platform made freely available as a web service (at <https://catma.de>), working with digital texts in any (UTF-8-encoded) language and used by individual researchers and in research projects worldwide. The technology and architecture of the CATMA web service programmed in JAVA might now be considered cutting edge.⁸ However, CATMA’s conceptual design is already some 212 years old: it is essentially that of the hermeneutic circle.

Hermeneutic procedures are considered “circular” in two respects. Friedrich Ast, who not only coined the term “hermeneutics” in 1808 but was also the first to draw attention to the hermeneutic circle, already observed that the “foundational law of all understanding and knowledge is to find the spirit of the whole through the individual, and through the whole to grasp the individual” (Ast, 1808: 75).⁹ This is the circularity of the part-whole-relationship, which emphasizes the compositional aspect of “understanding” and knowledge as dynamic relational constructs rather than atomic entities. The second aspect is that of the inescapable processual circularity of acts of judgment, which has been discussed by thinkers such as Gadamer: there is no *understanding* of a phenomenon that does not depend on a (however vague) *pre-understanding* and, conversely, no *pre-understanding* that does not already point back to an implicit, anterior *understanding*.

Taken together, the compositional and processual circularity of the hermeneutic approach have concrete implications for the design of a tool that aims to support hermeneutic practice. Such a tool must enable its user to practice, produce, and leverage annotation in such a way that the three methodological

⁸ For a more detailed description of CATMA’s conceptual logic and functionalities, see Meister, 2020.

⁹ My translation of the original German: “Das Grundgesetz alles Verstehens und Erkennens ist, aus dem Einzelnen den Geist des Ganzen zu finden und durch das Ganze das Einzelne zu begreifen.”

axioms of hermeneutics can be met, namely conceiving of the interpretive pursuit of an object's meaning as a process of discovery characterized by iteration, (inter)subjectivity, and reflexivity. For the actual development process, these requirements were translated into three goals:

- to set out with a high-level, hermeneutically functional text / meta-text data model;
- in terms of the markup technology being applied, to make sure to (a) support both modes of annotation, the declarative “top-down” and the non-deterministic, proto-interpretive, “bottom-up” variant, and (b) to integrate both annotation variants seamlessly with text analytical functions so that annotation and analysis can be performed in iteration;
- from a logical perspective, to provide processes and functions that make it possible to tolerate, document, and leverage inter-annotator variance as a potential hermeneutic indicator rather than premise such the categorization of such variance as a “disagreement” indicative of a procedural fault.

Figure 1 illustrates CATMA's text/meta-text data model:

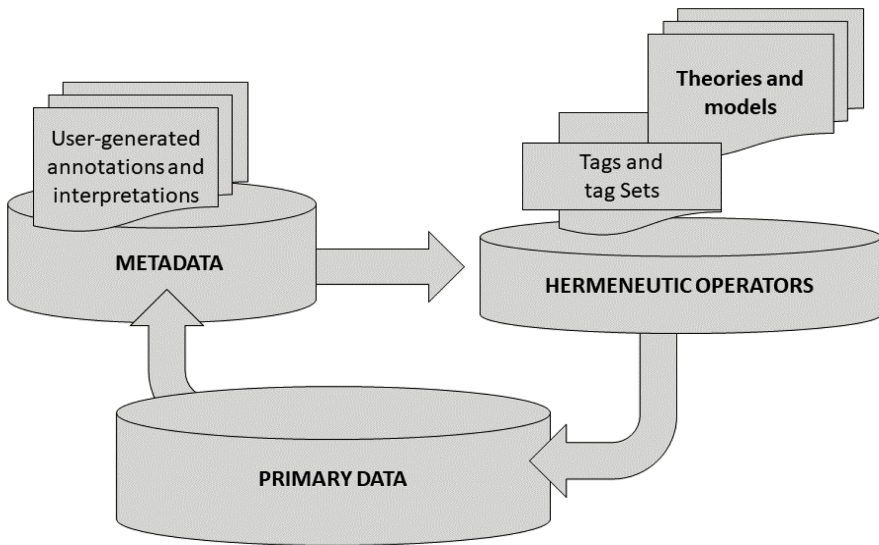


Fig. 1: A hermeneutic text/meta-text data model.

In a text annotation and analysis environment like CATMA, all *primary data* is – text. *Metadata*, then, is really any type of meta-text referring back to the primary text – from structured, machine-readable metadata in the sense of an

XML-TEI-header declaration right up to the verbalized, written-down interpretation of, say, a chapter in a novel.

Machine-generated metadata is usually deterministic – but user-generated, i.e., human-generated, metadata can and normally will vary more the less formal and the more interpretive a description is. This is why our model “stacks” annotations that relate to the same primary text without stipulating validation criteria or schemata upfront, i.e., as an integral part of the data model itself. Based on a schema-agnostic data model, it was thus possible to design and develop CATMA as an annotation platform that differs from others in four respects:

- CATMA is “undogmatic” in that it does not force you to make yes/no decisions when annotating a text. You can, of course, also use a predefined taxonomy and decide to stick to that – but you can also spontaneously add or redefine annotation tags. Moreover, you can even generate and analyze conflicting annotations (that is, the *markup*) of your text or corpus.
- CATMA supports collaborative annotation: a group of users can collaboratively annotate and analyze a text or a corpus.
- CATMA’s data model extends from text to annotation to meta-annotation: you can annotate an annotation or comment on it, thus treating annotations as discursive arguments rather than as declarative statements of fact.
- CATMA seamlessly integrates functions for annotation and analysis, which can thus be used iteratively and recursively. In doing so, CATMA models the hermeneutic circle in a “digital” fashion.

These four key features characterize CATMA as a unique digital tool that allows us to “mine the gap” between the hermeneutic (qualitative) and the statistical (quantitative) approach to texts, and in particular to semantic phenomena. Or to put it differently: CATMA’s underlying model conceptualizes the context dependency of literary phenomena as a type of “parameterization” – and if you change this parameter, the resulting interpretations, in all likelihood, will change as well. For example, if you want to annotate the poem by Edward Easton Cummings for grammaticality, chances are that your digital markup will declare a two-line phrase such as

– the great(my darling)happens to be
that love are in we, that love are in we

to be “un-grammatical” on a number of accounts from the perspective of a normative grammar of contemporary American English. In CATMA we might want to annotate this and similar phrases by assigning a tag for “grammaticality” and qualifying its value as “false” (see Fig. 2):

the great advantage of being alive
 (instead of undying) is not so much
 that mind no more can disprove than prove
 what heart may feel and soul may touch
 -the great (my darling) happens to be
 that love are in we, that love are in we

and here is a secret they never will share
 for whom create is less than have
 or one time one than when times where-
 that we are in love, that we are in love:

with us they've nothing times nothing to do
 (for love are in we am in I are in you)

this world (as timorous ilsters all
 to call their cowardice quite agree)
 shall never discover our touch and feel

The screenshot shows the CATMA interface with two main panels: 'Active Tagsets' and 'Active Annotations'. The 'Active Tagsets' panel has a table with columns 'Tagsets' and 'Tag Color'. It lists three tagsets: 'poetic language' (green arrow icon), 'word-order' (red arrow icon), and 'grammaticality' (yellow arrow icon). The 'grammaticality' tagset is highlighted in blue. The 'Active Annotations' panel has a table with columns 'Annotation' and 'Colc'. It lists three annotations: 'word-order' (red arrow icon), 'grammaticality' (yellow arrow icon), and 'false'. The 'word-order' annotation is highlighted in blue. The 'Writable Annotation Collection' is identified as 'cumings-jcm'.

Fig. 2: Assigning a “grammaticality” tag in CATMA.

And if we were to do this throughout the poem, we could then search not only for all instances where this specific combination of tag and value occur but also for, say, repeating words in these combinations or other tags that occur in proximity to this pattern, etc.¹⁰

Obviously, if we change the context parameter to, say, that of a Russian Formalist, the result will be different; our markup might record a case of *ostranenie*, i.e., of an intentional poetic deviation from the linguistic norm practiced by a given cultural community. And both annotation variants are of course “correct” in terms of their respective frame of reference. But let us go even one step further and assume that we have never heard of Jakobson’s concept of the poetic function nor of the Russian Formalists, nor of Brecht. We would simply read those two lines and annotate them in CATMA: some of us would be happy to draw on our knowledge of English grammar, thus sticking to the first setting of the context parameter, and to mark up the phrase as “ungrammatical.” The rest of us would come to various conclusions. For example, some might attribute the ungrammaticality to a printing error; others would argue that “love are in we” is

¹⁰ Annotation is only one of two procedures supported by CATMA: in fact, its true potential lies in its seamless combination of annotation and the analysis of the text itself, of any text/annotation-combination, and even of the triple combination of text/annotation/meta-annotation. For more detailed information on CATMA’s analytical function, see Meister, 2020; on the CATMA query language, see <https://catma.de/how-to/query-language/> (accessed April 24, 2020).

merely a rhetorical trope (inversion) that, at the time the poem was written, was so overtraded among American authors that no contemporary reader would have bothered to bat an eyelid. And so forth. Let us collect all of this; moreover, let us ask annotators to provide, record, and discuss the rationale for their individual annotation decisions – in other words: get annotators to annotate and comment on their annotations (which is supported by CATMA – see Fig. 3) and enable them to make transparent the individual frames of reference that informed their choices.¹¹ The result of such a collaborative effort, then, would be the philological equivalent of a jointly authored *dense description*, which in itself has already become the subject of discursive critique and reflection.

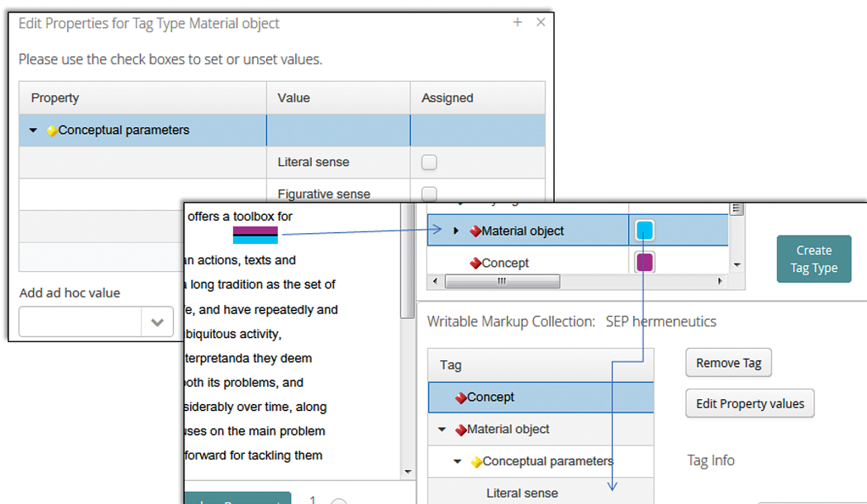


Fig. 3: Meta-annotation of conceptual context in CATMA.

And at this point, statistics and the quantitative approach might then become relevant again, for if we collect sufficient amounts of such “dense” but comparatively “unstructured” annotation data, correlations between annotation decisions and contextual parameters in and across phrases, texts, corpora, etc. might emerge that would escape the human eye but that can be captured by an algorithm. However, being able to handle complexity is, of course, only the first and more trivial of two advantages of taking this computational approach: for unlike human beings, algorithms are per se also agnostic to context in that it does not

¹¹ For a detailed description of CATMA’s system architecture, see the Appendix by Marco Petris in Meister, 2020.

matter to them existentially. While algorithms can be designed to factor in context parameters in their mathematical domain model (in this instance: a model of the text/meta-text/context correlations) – they will not *respond* to these parameters in phenomenological terms. Also, algorithms do not have favorites either – context is merely one of a number of variables, in this instance a conceptual one.

Obviously, if there is more than one unstructured annotation on the same text string, chances are that there is also more than one context at play. Provided we model context in terms of a value, this will automatically result in a case of quantitative variance, which is then an indicator of a likely qualitative disagreement among annotators. And so, using a tool that allows us to generate, document, trace, and analyze these cases of quantitative variance in user annotations can, in the end, lead to qualitative progress: to clarification, to discussion, to discursive resolution, to a reflective and conscious branching out of routines and approaches, and not least to a critique of underlying theorems and concepts.¹² In short, a tool based on a context-sensitive model of understanding enables its users to regard interpretive contexts not merely as distortive “noise” but as conceptual parameters that can either be preset (as in the case of a taxonomy- or ontology-guided, declarative markup approach) or be left undefined upfront in order to then study the range of values with which the variable will be instantiated in the recorded annotation instances.

8 Scalable reading: Exploring the methodological continuum

Broadening one’s interest in text annotation and text analysis to include this qualitative, user-centered aspect is more than simply making a pragmatic decision on the level of which markup practice to apply: it is indicative of a more fundamental enrichment of the research paradigm. Instead of merely practicing the “scholarly primitives” (Unsworth, 2000) of discovery and annotation in a digital environment, we are now simultaneously adopting a critical perspective on the logic of the markup process itself, and on its premises and constraints, as a type of user-generated interpretive annotation. By consciously broadening its scope to include this higher-level reflexive dimension, CATMA’s design does indeed go beyond the immediate pragmatic concern of reconciling the quantitative

¹² On using a CATMA-based approach to identify underspecified narratological concepts, see Gius, Jacke, 2016.

and the qualitative approach toward our primary object, i.e., the text or text corpus. For the conceptual benefit of enabling users to explore the full methodological continuum is ultimately of an epistemological order: the philosophy that informs the design of CATMA is that of adopting a holistic perspective on the hitherto divided modes of acquiring knowledge about a given domain in terms of a structured or an unstructured, a top-down or a bottom-up approach. This philosophy, then, is at its foundation a genuinely humanistic one in that it readdresses the challenge faced by the late eighteenth-century Romantics, who aimed to integrate the abilities to “solve differential equations, and write a verse,” as Heinrich von Kleist put it (Kleist, 1982: 160). In contemporary twenty-first-century DH, this challenge is taken up by combining mathematical and computational modeling with our traditional hermeneutic research agenda – in short: by mapping the ideal types of a formalist and a phenomenological epistemology onto one another in a controlled fashion.

Reconceptualizing the distinction between quantitative and qualitative approaches as a continuum of methodologies is also having an impact on the more recently proclaimed distinction between methods of textual study based on “close” and “distant” reading (Moretti, 2000). Rather than forcing its users to decide upfront which of the two methods to follow, a tool like CATMA invites us to explore the benefits of what Weitin (2017) has termed a “scalable reading” approach. This approach allows us to calibrate and adapt our research methods in relation to the object domain at hand and to the research question under investigation. Of course, for the scientific mode of text enquiry it is paramount that these decisions are made transparent and that they can stand up to critical scrutiny – in the same way as experimental settings in the natural sciences have to be made explicit, justified, and made available for critical examination and validation. A scalable approach should therefore not be misunderstood by digital humanists as a license to shift the goal posts as they see fit; rather, it entails a commitment to reflect upon and communicate how the various epistemic parameters shown in Fig. 4 below have been set in one’s approach:

9 Conclusion

As we have seen, the call to bring qualitative and quantitative approaches and methods of scientific inquiry into contact again is certainly no new idea – but to do so by drawing inspiration from the digital humanities and to combine computational and hermeneutic approaches and procedures is nevertheless one of the most exciting methodological experiments of our times. Moreover,

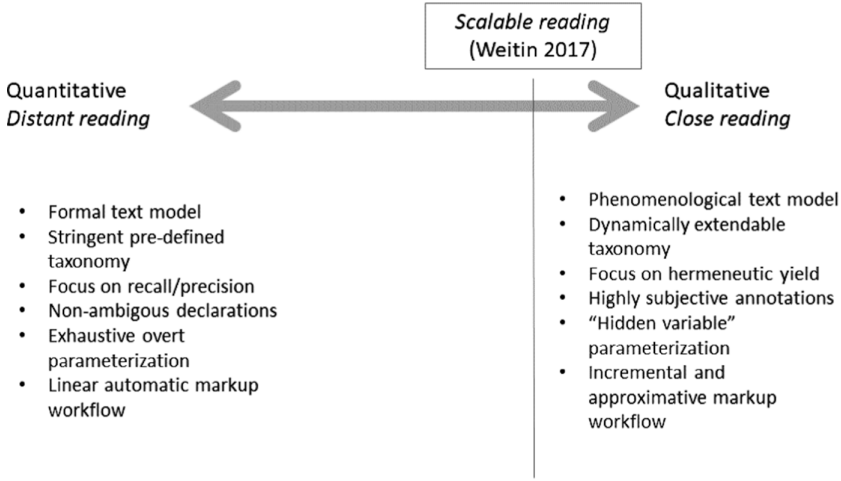


Fig. 4: Epistemic parameters of the “scalable reading” continuum.

from a philosophical perspective, this amounts to taking up a “grand challenge,” for in order to model the full phenomenological-empirical continuum of human understanding, which ranges from processes of elementary perception to those of theory-guided cognition and ultimately to our reflexive, human sense-making activities, we have to learn how to:

- conceptualize symbolic artefacts (text, language, objects, and practices) in both a holistic (phenomenological) and an analytic (formal) mode;
- support our research practices as both systematic and intuitive routines;
- combine the linear input/output transformation of data with iterative and discursive data processing models; and
- mediate between the “distant” and the “close” perspective on symbolic artefacts by developing scalable models and methods for capturing, annotating, interpreting, and analyzing data and metadata.

Will this, then, enable us to *machiner la poésie*? The answer depends on what exactly we mean by *machiner* – produce, manipulate, or process. As a literary scholar, I am quite happy to restrict my ambition to the latter. As an affectionate reader of poetry, on the other hand, I am prone to echo Cummings’s verdict: while “a billion brains may coax undeath / from fancied fact and spaceful time,” any intelligence – including an artificial or computational one – unable to experience human emotion directly and on a phenomenological level will fail to “discover our touch and feel.” And as a digital humanist, what really interests me is by definition: what happens in between.

References

- Aarne A. Verzeichnis der Märchentypen mit Hilfe von Fachgenossen ausgearbeitet. Helsinki: Suomalainen Tiedeakatemia, 1910.
- Aarne A, Thompson S. The Types of the Folktale. A Classification and Bibliography. Helsinki: Suomalainen Tiedeakatemia, 1961.
- Ast F. Grundlinien der Grammatik, Hermeneutik und Kritik. Landshut: Jos. Thomann, 1808. Digitized facsimile available from the Bayerische Staatsbibliothek digital/MDZ: <http://mdz-nbn-resolving.de/urn:nbn:de:bvb:12-bsb10582792-2> (accessed April 24, 2020).
- Bauer M, Zirker A. Whipping Boys Explained: Literary Annotation and Digital Humanities. In: Price K, Siemens R, editors. *Literary Studies in the Digital Age*. MLA, 2013. <https://dlsanthology.mla.hcommons.org/whipping-boys-explained-literary-annotation-and-digital-humanities/> (accessed April 21, 2022).
- Burnard L. On the Hermeneutic Implications of Text Encoding. 1998. <http://users.ox.ac.uk/~lou/wip/herman.htm> (accessed April 22, 2020).
- Cummings EE. The Great Advantage. In: Firmage GC, editor. *Edward Estlin Cummings: Complete Poems, 1904–1962*. New York, London: Liveright Publishing Company, 2015: 705.
- Dilthey, W. 3. The Rise of Hermeneutics. In: *Wilhelm Dilthey: Selected Works*, vol. IV: *Hermeneutics and the Study of History*, ed. by Makkreel RA, Rodi F. Princeton: Princeton University Press, 2018 [1900]: 235–258. <https://doi.org/10.1515/9780691188706-006> (accessed April 21, 2022).
- Frog SK. Revisiting the Historical-Geographic Method(s). In: Lukin K, Frog SK, editors. *Limited Sources, Boundless Possibilities: Textual Scholarship and the Challenges of Oral and Written Texts*. RMN Newsletter, 7, Special Issue. Helsinki: University of Helsinki, 2013: 18–34.
- Gervas P, Peinado F. Evaluation of Automatic Generation of Basic Stories. *New Generation Computing* 2006; 24: 289–302. <https://doi.org/10.1007/BF03037336>.
- Gius E, Jacke J. Zur Annotation narratologischer Kategorien der Zeit: Guidelines zur Nutzung des CATMA-Tagsets. 2016. <http://heureclea.de/wp-content/uploads/2016/11/guidelinesV2.pdf> (accessed May 6, 2020).
- Goethe, JW. *The Collected Works*, vol. 12: *Scientific Studies*, ed. and transl. by Miller D. New York: Suhrkamp, 1988.
- Goethe JW. *Werke*, Hamburger Ausgabe in 14 Bänden, vol. 13: *Naturwissenschaftliche Schriften I*, ed. by Erich Trunz. Munich: Verlag CH Beck, 1994.
- Hager FP. *Empeiria*. In: Ritter J, Gründer K, Gabriel G, editors. *Historisches Wörterbuch der Philosophie online*. Basel: Schwabe Verlag, 2017. <https://doi.org/10.24894/HWPh.838> (accessed April 24, 2020).
- Jacke J. Workshop Report: Non-hierarchical Concept Ontologies and Markup Schemas. In: *forTEXT: Literatur digital erforschen. Report on the 3rd Expert Workshop*. 2020. <https://fortext.net/news/2020/workshop-report-non-hierarchical-concept-ontologies-and-markup-schemas> (accessed April 24, 2020).
- Jakobson R. *Linguistics and Poetics*. In: Sebeok TA, editor. *Style in Language*. Cambridge: MIT Press, 1960: 350–377.
- Kleist H. Letter to Ernst von Pfuel (1805). In: *An Abyss Deep Enough: Letters of Heinrich Von Kleist, with a Selection of Essays and Anecdotes*, ed. and transl. by Miller PB. New York: Dutton, 1982: 160.

- Manuwald B. *Wiedererinnerung/Anamnesis*. In: Horn C, Müller J, Söder JR, editors. *Platon-Handbuch: Leben, Werk, Wirkung*. Stuttgart: Metzler, 2009: 324–328.
- Meister JC, Horstmann J, Petris M, Jacke J, Bruck C, Schumacher M, Flüh M. CATMA 6.0.0. 2019. <https://doi.org/10.5281/zenodo.3523228> (accessed April 24, 2020).
- Meister JC. From TACT to CATMA or a Mindful Approach to Text Annotation and Analysis. In: Nyhan J, Rockwell G, Sinclair S, editors. *On Making in the Digital Humanities: Essays on the Scholarship of Digital Humanities Development in Honour of John Bradley*. Forthcoming 2022. Preprint, submitted in 2020. http://jcmeister.de/downloads/texts/Meister_2020-TACT-to-CATMA.pdf (accessed April 24, 2020).
- Moretti F. Conjectures on World Literature. *New Left Review* 2000; 1. <https://newleftreview.org/issues/111/articles/franco-moretti-conjectures-on-world-literature> (accessed April 24, 2020).
- Moretti F. Operationalizing. *New Left Review* 2013; 84: 103–119. <https://newleftreview.org/11/84/franco-moretti-operationalizing> (accessed April 24, 2020).
- Neves M, Jurica Š: An Extensive Review of Tools for Manual Annotation of Documents. *Briefings in Bioinformatics* 2021; 22(1): 146–163. <https://doi.org/10.1093/bib/bbz130> (advance access publication 2019; accessed April 22, 2022).
- Propp V. *Morphology of the Folktale*. Austin: University of Texas Press, 1968.
- Rumelhart D. Notes on a Schema for Stories. In: Bobrow DG, editor. *Representation and Understanding: Studies in Cognitive Science*. New York: Academic Press, 1975: 211–236.
- Schleiermacher F. *Hermeneutik und Kritik mit besonderer Beziehung auf das Neue Testament*. Berlin: G. Reimer, 1838. *Deutsches Textarchiv*, http://www.deutschestextarchiv.de/schleiermacher_hermeneutik_1838/ (accessed April 24, 2020).
- Šklovskij V. Resurrecting the Word. In: Viktor Shklovsky: A Reader, ed. and transl. by Berlina A. New York, London: Bloomsbury Publishing, 2017 [1914]: 63–72.
- Smith DW. Phenomenology. In: Zalta EN, editor. *The Stanford Encyclopedia of Philosophy*. 2018. <https://plato.stanford.edu/archives/sum2018/entries/phenomenology/> (accessed April 22, 2020).
- Snow CP. *The Two Cultures*. London: Cambridge University Press, 1993 [1959].
- Thompson S. *The Folktale*. Berkeley: University of California Press, 1977 [1946].
- Unsworth J. Scholarly Primitives: What Methods do Humanities Researchers have in Common, and How Might our Tools Reflect This? In: *Symposium on Humanities Computing: Formal Methods, Experimental Practice*. King's College, London, May 13, 2000. <http://www.people.virginia.edu/~jmu2m/Kings.5-00/primitives.html> (accessed April 24, 2020).
- Uther HJ. *The Types of International Folktales: A Classification and Bibliography. Based on the System of Antti Aarne and Stith Thompson*. Helsinki: Suomalainen Tiedeakatemia, 2004.
- Waugh LR. The Poetic Function in the Theory of Roman Jakobson. *Poetics Today* 1980; 2(1):57–82. <https://doi.org/10.2307/1772352> (accessed April 24, 2020).
- Weitin T. Scalable Reading. *Zeitschrift für Literaturwissenschaft und Linguistik* 2017; 41(1):1–6. <https://doi.org/10.1007/s41244-017-0048-4> (accessed April 24, 2020).

Helena Bermúdez Sabel, Pablo Ruiz Fabo,
and Clara Martínez Cantón

DISCOVering Spanish Sonnets: A Circular Reading Experience

Abstract: With DISCO, the *DI*achronic Spanish Sonnet *COR*pus, we collected 4,085 sonnets written from the fifteenth to nineteenth centuries, including canonical and lesser-studied authors from both Spain and Latin-America, with detailed author metadata, metrics, rhyme scheme, and enjambment annotations. The dataset is available in public repositories, offered in plain text and TEI, and enriched with RDFa, a linked data format. The corpus was intended for research and teaching, as well as for non-specialist use. Some questions naturally emerge: How can we easily navigate this corpus and its rich metadata? How can we identify trends using its annotations? How can users not proficient in XML query languages and linked data benefit from such a dataset? To address these issues, we have created DISCOVer, a user-friendly web interface that allows users to explore the DISCO corpus. It was conceived as a means to help students, researchers, and general readers, making the corpus accessible to a wider audience. The interface displays literary annotations in individual texts as well as quantitative data on user-defined subcorpora. In turn, from the aggregated data, we can go back to the texts on which the data are based. Thus, the interface helps us to assess the features of individual texts in the context of quantitative data about larger parts of the corpus and vice-versa. It also helps us to nuance hypotheses based on aggregated data by consulting the texts from which they are derived.

Acknowledgments: The interface was developed thanks to a Josef Dobroský Fellowship funded by the Akademie věd České republiky (Czech Academy of Sciences; 2018) and hosted at the Institute of Czech Literature AS CR in March 2019, as well as a Zdeněk Pešat Fellowship during its Spring 2019 program, also hosted at the Institute of Czech Literature AS CR during the same period.

Helena Bermúdez Sabel, University of Neuchâtel, Institute of Language Studies,
e-mail: helena.bermudez@unine.ch

Pablo Ruiz Fabo, University of Strasbourg, LiLPa UR 1339, Linguistics, Language, and Speech,
e-mail: ruizfabo@unistra.fr

Clara Martínez Cantón, UNED, Department of Spanish Literature and Literary Theory,
e-mail: cimartinez@flog.uned.es

1 Introduction

In this chapter, we would like to present DISCOVer (<https://prfl.org/discover/>), an interface that provides both descriptive and analytical tools that allow users to explore the *Diachronic Spanish Sonnet Corpus* (DISCO). This is a dataset we started creating in 2017 that is available on GitHub¹ and Zenodo,² offered in plain text and XML-TEI, and enriched with RDFa, a linked data format (Ruiz Fabo et al., 2021).

The sonnet is of prime importance in European poetry; an interface that allows users to explore a comprehensive corpus of sonnets is thus relevant for literary scholarship. One goal of this interface is to offer a set of functions that help users to explore the corpus in an intuitive way. The target users are not just literary scholars but also students and non-specialists. With this interface we hope to make our diachronic sonnet corpus accessible to a wider audience. The conceptualization behind DISCOVer is not about simply querying the database to read the texts contained in the corpus. We intended to create a tool that can help us to gain new knowledge about the sonnet genre in Spanish and that can be used in educational settings.

The chapter is structured as follows: Section 2 contains an overview of the state of the art, followed in Section 3 by a description of the corpus. Section 4 presents the DISCOVer interface itself. Besides exploring the main functionalities of the interface, we discuss how it can mediate an effective *circular reading* of the sonnet in Spanish and its metrical and rhyme features: the interface allows us to easily access individual texts in the collection as well as aggregated quantitative data on prosody and rhyme for the corpus overall, or for metadata-based subcorpora (e.g., poems by a given author, from a given period, or sharing a common origin).

2 State of the art

One fundamental difficulty for digital humanities studies of Spanish literature is the scarcity of digital resources (Agenjo, 2015). This relates, firstly, to digital corpora for Spanish poetry but also to tools for automatically annotating Spanish poems with literary information and for utilizing the annotations. We will first

1 <https://github.com/pruizf/disco> (accessed April 20, 2022).

2 <https://doi.org/10.5281/zenodo.1012567> (accessed April 20, 2022).

cover existing work on digital resources (Section 2.1) before addressing automatic annotation (Section 2.2) and visualization (Section 2.3).

2.1 Digital corpora for poetry in Spanish

While resources are not abundant, there are some examples, especially for certain periods. For medieval literature we have generic collections like BiDTEA (Gago Jover, 2015) and ADMYTE (Marcos Marín, Faulhaber, 1992), as well as digital resources pertaining to poetry, like the multimedia editions of *Cancionero de Palacio* or the *Cancionero de Upsala* available at the CPDL (Choral Public Domain Library, 1998–). For later periods, Navarro-Colorado et al. (2016) have presented the *Corpus of Spanish Golden-Age Sonnets*, which covers major authors from the fifteenth to seventeenth centuries, with automatic metrical annotation (stress patterns).

As for the sonnet corpora available, besides the aforementioned *Corpus of Spanish Golden-Age Sonnets*, there is the *Sonnet-Archiv* (Elf Edition), which is organized as a forum and has less coverage than our DISCO project. The *Biblioteca del Soneto* (Sonnet Library) (Biblioteca Virtual Miguel de Cervantes, 2007) is organized alphabetically, rather than by meaningful criteria for literary scholarship such as periods. Both the *Sonnet-Archiv* and the *Biblioteca del Soneto* are traditional websites that only present the poems' text. Author metadata in these corpora are very limited and unavailable in a machine-readable format (for a discussion of related issues, see Calvo Tello, 2017).

DISCO complements this growing ecosystem with the addition of a meaningful representation of sonnets from the fifteenth to the nineteenth centuries with literary annotations and descriptive information about the authors.

2.2 Automatic poetry annotation in Spanish

The sonnet is a form that is “manageable” to treat computationally: it obeys clear restrictions; precise literary annotations can thus be automatically implemented. Variability stays within bounds, making meaningful comparison across poems easier in terms of scansion or rhyme types.

The sonnet has received attention from the computational linguistics community (Navarro Colorado, 2015; Navarro Colorado, 2016; Navarro Colorado,

Ribes, Sánchez, 2016; Agirrezabal, 2017) including in the ADSO project (Navarro Colorado, 2017).³

In their classical form, sonnets in Spanish comprise 14 hendecasyllable lines. Early work on the metrical analysis of the Spanish hendecasyllable was carried out by Gervás (2000). The ADSO scansion tool by Navarro Colorado (2017) is a more recent system and is the tool used to annotate the stress patterns in the DISCO corpus. Automatic scansion in Spanish has also been carried out by Agirrezabal (2017), de la Rosa et al. (2020), Marco Remón and Gonzalo (2021), and Mittmann (2016).⁴

Multilingual rhyme detection has been performed by Plecháč (2018); we used his Rhyme Tagger tool to annotate our corpus. We have proposed automatic enjambment detection in our earlier work (Ruiz Fabo et al., 2017; Martínez Cantón et al., 2021), using the enjambment typology by Quilis (1964). We used this system to annotate our corpus.

2.3 Poem visualization and corpus exploration interfaces

Regarding interfaces for poetry exploration, we can distinguish between two main types of work, most of which has been carried out on English poetry. On the one hand, there are interfaces that help perform close readings of a single poem. On the other, we have tools intended to help explore a complete corpus and provide aggregated quantitative analyses on the text collection, along the lines of distant reading.⁵ As we discuss, our DISCOVER interface allows users to explore the corpus using aggregated data but does show the same annotations at the poem level, thereby helping users to move back and forth between the individual and the aggregated level.

Regarding tools for enhancing the close reading of poetry in English, *Myopia* focuses on comparative analyses of different (manually) TEI-encoded versions of a poem, highlighting areas of divergence across encodings (Chaturvedi, 2011, Chaturvedi et al., 2012). It utilizes annotations for sound, metrics, and content (e.g., images or emotion-related language). Continuing with resources for English poetry, *Poem Viewer* provides detailed phonetic and rhyme analyses, presented in visually appealing ways to help detect patterns (Abdul Rahman et al., 2013).

³ For the ADSO project, see <http://adso.gplsi.es/index.php/en/adso-project/> (accessed April 20, 2022).

⁴ <https://aoidos.ufsc.br/> (accessed April 20, 2022).

⁵ A different type of tool has been proposed by Meneses, Furuta, and Mandell (2013). Their *Ambiances* platform is geared toward helping authors to visualize their work in progress while writing, thereby allowing the visualization to influence the creative process.

Poemage offers a similar functionality to *Poem Viewer*, emphasizing the dynamics of poetic features and their interactions as they evolve within the poem (McCurdy et al., 2016). *SPARSAR* allows us to visualize different automatically identified features (Delmonte, 2015): sound devices, metrical schemes, and semantic categories such as negative vs. positive polarity or abstract vs. concrete expressions. This resource thus aids in the exploration of sound/meaning correlations. The *Eighteenth-Century Poetry Archive* displays rhyme analyses and rhetorical figures (Huber, 2020). In addition, it offers morphological and syntactic annotations among other linguistic information. It also integrates *Poem Viewer* for sound analysis. *For Better For Verse* is an interface for the visualization and manual annotation of metrical patterns in English poetry, which can be used for pedagogical purposes (Tucker, 2011). Regarding resources for languages other than English, the *Buscador de sonetos del Siglo de Oro*⁶ offers an interface for exploring Spanish Golden Age sonnets.⁷ Its search functionalities allow us to query the corpus with different levels of granularity, including the metrical pattern of a line. The search results provide detailed references to sonnet lines from the corpus matching the search criteria. The result list includes a link to the complete works of the *Biblioteca Cervantes* in which that particular sonnet is contained (Candela et al., 2017).

Moving to interfaces that focus on corpus exploration via quantitative analyses, we would like to mention the examples for Czech poetry that inspired our work. Plecháč and Kolár (2015) have created a visualization interface called the *Database of Czech Metres*.⁸ This offers metrical and rhyme scheme analysis, also identifying fixed forms (e.g., sonnet or rondel). It provides different quantitative analyses for the poems' features, using metadata like periods or authors to aggregate feature values. *Gunstick – Database of Czech Rhymes* focuses on rhyme pair analysis (Plecháč, Ibrahim, 2013).

The first thing that differentiates our DISCOVER interface is, of course, that it is designed for Spanish sonnets.⁹ It offers similar functions to the *Database of Czech Metres* with the exception that the metrical information in DISCOVER is always accessed through the text. However, one original feature of our work is the addition of enjambment annotations. Our rhyme visualization utilizes code from the *Database of Czech Rhymes* but improves navigation between the level

⁶ This translates as “Golden Age Sonnet Search Tool.”

⁷ <https://data.cervantesvirtual.com/goldenage/> (accessed April 20, 2022).

⁸ http://versologie.cz/v2/tool_dcm/ (accessed April 20, 2022).

⁹ Adding other Spanish forms would be feasible. It would involve annotating the metrical patterns with tools designed (unlike Navarro Colorado's) for non-fixed meter, such as *Rantanplan* (de la Rosa et al., 2020).

of aggregated rhyme analysis and the individual poem level. In short, given the set of metadata that can be used to aggregate results, the literary annotations available, and the focus on moving across the single text and aggregated data level, the DISCO interface provides an original way to explore a poetry corpus.

3 Source description: The DISCO corpus

Electronic corpora are an effective way to create resources with wide coverage specific to a poetic form or genre. Sonnet-specific anthologies for Spanish are not abundant; the same is true of other poetic forms of major importance in Spanish, like the *décima*. The lower costs involved in electronic corpora compared to print editions make it easier to create electronic resources of this type.

With DISCO, we collected 4,085 sonnets: 2,676 from the nineteenth and early twentieth centuries, 321 from the eighteenth century, and 1,088 from the Spanish Renaissance and Golden Age (fifteenth to seventeenth centuries). There are a total of 1,204 authors, including canonical and lesser-studied authors from both Spain and Latin America, with detailed author metadata, meter, rhyme scheme, and enjambment annotations. Details about the different methods used to annotate the corpus are available elsewhere (Ruiz Fabo et al., 2021). The corpus intends to provide a wide sample inspired by distant reading approaches (Moretti, 2005; Jockers, 2013). Most of the raw texts were extracted from the Biblioteca Virtual Miguel de Cervantes (1999), with some eighteenth-century texts coming from Wikisource.

The corpus is available in plain-text and in TEI formats. XML-TEI P5 was used given this standard's advantages in terms of reuse, storage, and retrieval. Author metadata were extracted or inferred from unstructured content in the sources (year, places of birth and death, and gender), and placed in the TEI header or in a metadata table in the case of the plain-text version. To prepare the corpus, we closely followed the TEI guidelines and RIDE's criteria for Digital Text Collections (Henny-Krahmer, Neuber, 2017).

Additionally, authors have been assigned VIAF identifiers and have been described using RDFa attributes.¹⁰ This gives the corpus an entry point to the Linked Open Data Cloud, enhancing its findability. The corpus is available as a GitHub repository¹¹ and has been saved in Zenodo¹² in line with good practice for data use, reuse, and conservation.

10 VIAF is the Virtual International Authority File: <https://viaf.org/> (accessed April 20, 2022).

11 <https://github.com/pruizf/disco> (accessed April 20, 2022).

12 <https://doi.org/10.5281/zenodo.1012567> (accessed April 20, 2022).

4 The DISCOVer interface

The DISCO corpus was intended for research, teaching, and also for non-specialist use. Although it would be possible to extract information from DISCO using XML query languages, not all target users have such skills. Therefore, with DISCOVer we created a user-friendly interface that would allow a wider audience to benefit from the corpus. In 4.1 we introduce the interface functions and intended uses. In 4.2 we describe the circular reading process enabled by the interface, highlighting how iterative access to single texts and aggregated data on different text sets can help us to test and improve our hypotheses regarding meter, rhyme, and enjambment. We also discuss an example of rhyme analysis. In 4.3 we provide some technical details regarding the interface architecture.

4.1 Main functions

We would first like to describe the interface's search workflows and how users can select a subcorpus with which to work or multiple subcorpora to explore simultaneously. We will then present the prosodic features for which the interface provides quantitative analyses and plots, and the way individual texts are presented. Finally, we will discuss rhyme analysis.

4.1.1 Search and subcorpus creation

DISCOVer offers corpus query and navigation via two main search functions: author search and faceted search based on author metadata. Facets are available for periods, author gender (female or male), and author origin (Latin America or Europe). Our intention was to highlight the work of female poets and Latin American authors by including specific facet values for them (Fig. 1a). For author search, we implemented a free text predictive search box (Fig. 1b).

The set of authors matching a query is displayed on a results pane (Fig. 1c). From there, one can select either individual poems or all poems for a single author. The user can provide a name for the selected subcorpus; this name will be used to identify results for that subcorpus in the charts. For the selected poems, users can access several charts providing quantitative aggregated data on prosodic and rhyme features for the selection. Users can also access the texts for the selected poems. Details are given in the following subsections.

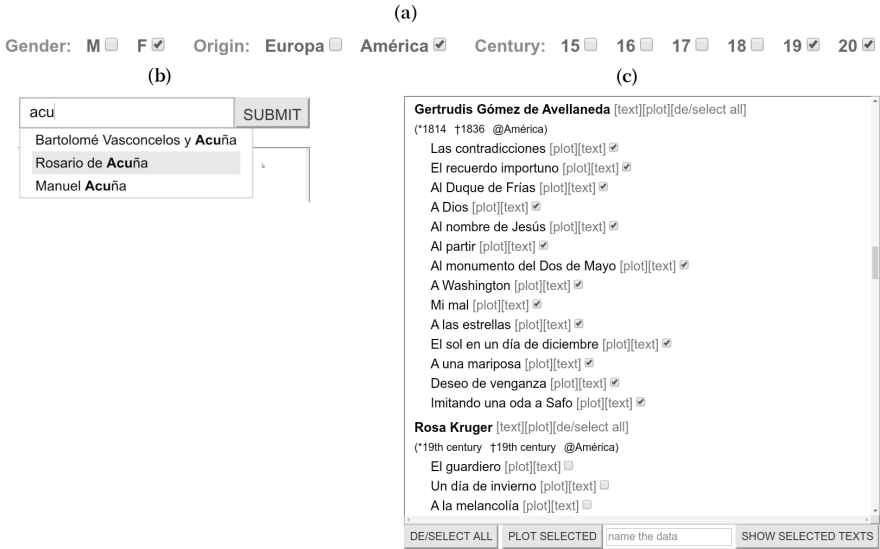


Fig. 1: Corpus query and subcorpus creation functions. In (a), we see the faceted search options. In (b), we see the predictive free-text search box. In (c), we see the query results for the faceted search from (a): Latin American women writers from the nineteenth and twentieth centuries. The result list in (c) acts as a text selection pane: it can be used to create the named subcorpora (e.g., all texts by an author or a combination of individual texts by different authors). The interface can then plot aggregated results for prosodic features (see Fig. 2) and display the subcorpus texts (Fig. 3).

4.1.2 Prosodic feature plots

Charts display features related to both metrical and rhyme schemes as well as the presence of enjambment in the poems selected by the user (more fine-grained rhyme information is treated in a separate view; see 4.1.4 below). Multiple selections can be plotted to make comparisons (Fig. 2). The charts thus allow us to compare the different metrical features of each poem with other poem and/or author selections and with the complete corpus. This helps determine how a poem is representative of or diverges from the typical values for an author, period, or group defined according to social factors (origin and gender). Fig. 2 gives an example: on the left, we see that alternate rhyme in the quatrains is overrepresented in Latin American nineteenth- and early twentieth-century female writers compared to European female writers from the same period. However, by selecting different author groups and individual authors using the search and filtering functions, we obtain a more nuanced picture, and we can

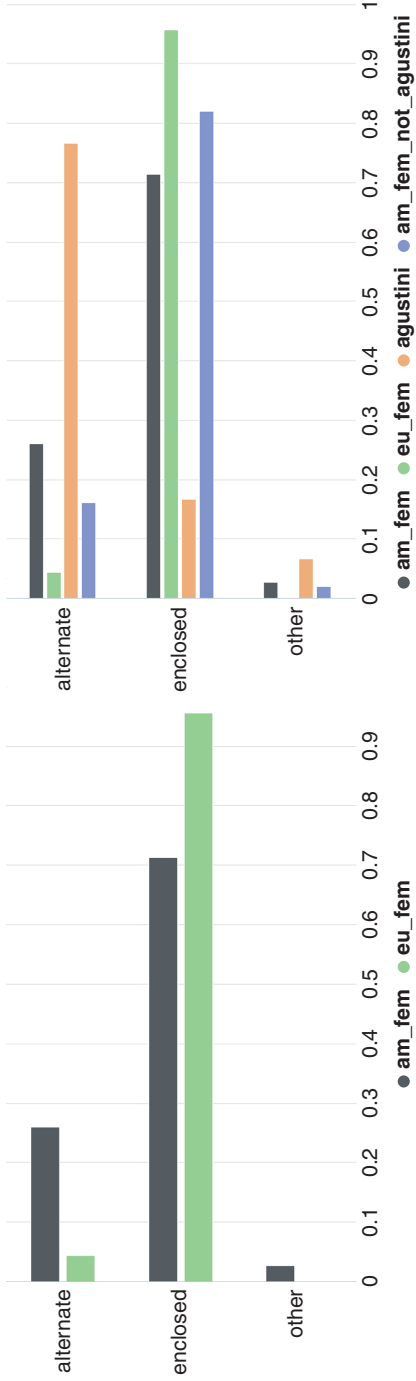


Fig. 2: Aggregated results based on the corpus metadata and prosodic annotations. The charts display quatrain rhyme scheme percentages in different subcorpora. Other annotations available are stress pattern, tercet rhyme scheme, and enjambment type. On the left, we see quatrain rhyme scheme use in works by Latin American women (*am_fem*) vs. European women (*eu_fem*) in the nineteenth and early twentieth centuries. On the right, two additional subcorpora have been plotted: Delmira Agustini (*agustini*) and all Latin American women except Delmira Agustini (*am_fem_not_agustini*).

verify whether the preference for alternate rhyme is homogeneous across authors in the Latin American group or heavily influenced by certain individual authors. The data on the right suggest that this overrepresentation is influenced by the almost exclusive use of alternate rhyme in authors like Delmira Agustini (75% of her quatrains), rather than being a clear preference in the group as a whole.

4.1.3 Text-level view

From the search results panel (Fig. 1c), a user can select a single poem or several poems and access their text. From the text, a simple click brings up the metrical scheme of each line, the rhyme scheme of the sonnet or the presence of enjambments, including the description of their typology (see Fig. 3).

Rhyme scheme Metrical scheme (stressed syllables) Enjambement

Juana Borrero

*1877 †1896 @América

Rêve

Su voz debe ser dulce y persuasiva	A	2, 3, 5, 6, 10	↵
y soñadora y triste su mirada	B	4, 6, 10	
Debe tener la frente pensativa	A	1, 4, 6, 10	
por un halo de ensueños circundada	B	2, 3, 6, 10	
Su alma genial, cual pálida cautiva	A	1, 4, 6, 10	↵
de un astro esplendoroso desterrada,	B	1, 2, 6, 10	
sueña con una nube fugitiva	A	1, 4, 6, 10	
y con el traje de crespón de un hada	B	4, 8, 9, 10	
Cuando la sonda azul de los delirios	C	4, 6, 10	↵
disipa sus nostálgicos martirios	C	2, 6, 10	
borrando del pesar la oscura huella,	D	2, 6, 8, 10	
él se acuerda en la noche silenciosa,	E	1, 3, 6, 10	
de aquella virgencita misteriosa	E	2, 6, 10	↵
que dejó abandonada en una estrella	D	3, 6, 8, 10	

Cross-clause enjambment
(encabalgamiento oracional)

Fig. 3: Individual text annotation display: rhyme scheme, metrical scheme, and enjambment. A tooltip specifies enjambment type, e.g., cross-clause enjambment on line 13. Clicking on the rhyme scheme provides access to rhyme word analysis (Fig. 4).

The presentation of the rhyme scheme in individual poems is, moreover, interactive: if the user clicks on one of the letters of the rhyme scheme, a new results

window is presented that describes the overall use of the corresponding rhyme word in the corpus (see 4.1.4 below and Fig. 4).

4.1.4 Rhyme view

The *rhyme view* (Fig. 4) displays statistics on rhyme pair distribution in the corpus. The user can choose which rhyme pairs to analyze by selecting a rhyme word in an individual poem (see Fig. 3) or from a rhyme search, as will be detailed below.

The rhyme view is highly interactive. As can be seen in Fig. 4, on the top-left pane, the distribution of rhyme pairs is presented as a pie chart (alternatively in table format). We see that *sombrío* (shaded, somber) rhymes more frequently with *frío* (cold). Clicking on pie chart sectors (or table rows when the table format is active) will limit results to the rhyme words selected.

The top-right pane shows the distribution of the selected rhyme pairs by author, century, and continent; clicking on the relevant button will update the chart with data aggregated according to each of those criteria. Rhyme pairs can be plotted individually and simultaneously. Therefore, while the pie chart offers a quick overview of their frequencies, the stacked bar charts provide a means for comparing their distribution according to the criteria mentioned.

The bottom pane consists of a sortable table with line pairs in which the selected rhyme words appear. To ensure access to individual texts from the aggregated data view, users can access the complete text of a poem from this table by clicking on the title. In turn, as stated above, clicking on the rhyme scheme column for each line of an individual poem provides access to the overall rhyme view for that line's rhyme word. This allows the user to shift between the level of individual text analysis and the level of quantitative data based on text collections. The ability to switch easily between these two levels is one of the purposes of our interface.

Rhyme-based corpus exploration can be accessed by selecting rhyme words while reading individual poems, but it can also be carried out by searching for rhyme words using the *rhyme* tab on the interface's menu, where we have made it possible to search rhyme words using author and century filters.

4.2 Circular reading

The interface was designed to help users test their hypotheses in two ways. First, hypotheses based on a small number of texts can be tested by accessing relevant aggregated quantitative data. Second, hypotheses based on aggregated

This resource replicates [Gunstick](#), the rhyme database and related tools developed by the [Versologie](#) research group.

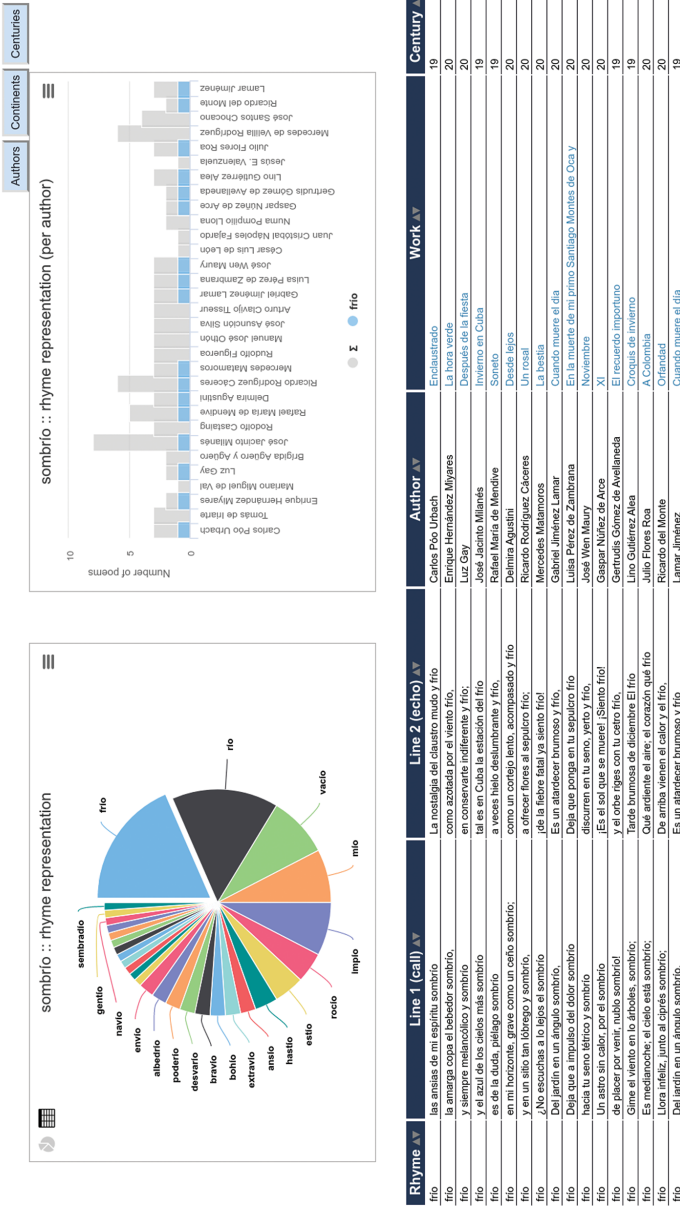


Fig. 4: Rhyme view. Corpus-level results for rhyme word *sombrio* (shaded, somber). The interactive top-left pane shows rhyme pair distribution and allows us to select rhyme words (see also Fig. 5). The top-right pane shows the distribution of selected rhyme pairs by author, region, and period. The lower pane puts quantitative results in context by showing rhyming line pairs and providing links to the complete text of the poems.

data can be assessed by accessing the individual texts on which the data rely. Combining both procedures to iteratively check one's hypothesis can be referred to as *circular reading*. This can be seen as a type of radial reading geared toward data examination (McGann, 1991).

For instance, consider the poem in example (1), available on DISCOVER, by Brígida Agüero, a nineteenth-century Cuban poet. A literal English gloss for each expression is given beneath the lines.

- (1) Resignación
 Resignation
- 1 ¡Soberano Señor Omnipotente,
 Sovereign Lord Almighty
- 2 por quien el Sol espléndido fulgura,
 by whom the Sun splendid shines,
- 3 el ave canta, el céfiro murmura,
 the bird sings, the zephyr whispers
- 4 y vierte sus raudales el torrente!,
 and pours its flow the torrent!,
- 5 oye mi voz: el alma reverente
 listen to my voice: the soul reverent
- 6 implora tu piedad en su amargura;
 implores your pity in its bitterness;
- 7 mitiga un tanto mi letal **tristura**,
 *mitigates somewhat my lethal **sadness**,*
- 8 mi cruel angustia, mía **ansiedad creciente**
 *my cruel anguish, my **anxiety growing***
- 9 Al través de una triste perspectiva,
 Through a sad perspective
- 10 miro tan sólo un porvenir **sombrío**,
 *I look at only a future **somber**,*
- 11 y más mi pena sin cesar se aviva
 and more my sadness without cease livens
- 12 Un mal terrible me atormenta **impío**
 *A pain terrible torments me **cruelly***
- 13 mas si te place que muriendo viva,
 but if it pleases you that dying I shall live,
- 14 “cúmplase en mí tu voluntad, Dios mío”
 “let be done upon me thy will, Lord mine”

In this poem we notice that the rhyme word on line 10, *sombrío* (somber), has a negative connotation and rhymes with another negative word, *impío* (cruelly), on line 12. Lines 6 and 8 also feature a negative rhyme pair, *amargura-tristura* (bitterness-sadness). Negatively connoted rhyme words are in bold. Conversely, the rhyme pair on lines 2 to 3 can be seen as more positively connoted, evocative of powerful natural phenomena such as sunlight or a torrent's stream (neither of which is presented as destructive in the poem): *fulgura-murmura* (shines-whispers). Positively connoted rhyme words are underlined. A third possibility appears in line pair 5 and 8: *reverente* (reverent, worshipping), which can be seen as positive, rhymes with *ansiedad creciente* (growing anxiety), a negative emotion.

In the same poem, we see that rhyme pairs can share the same (negative or positive) polarity or show a contrasting polarity. We may want to explore the proportion of agreement vs. contrast in positively/negatively connoted words within a rhyme pair in the author's period. The rhyme analysis tool in DISCOVER helps us carry out this analysis. Let us start with the word *sombrío*, which can suggest positive emotions (the relief from the scorching sun provided by a shaded area) as well as negative ones (gloomy, somber). In the rhyme word overview (Fig. 5), we find some clearly negative rhyme words like *vacío* (emptiness) or *hastío* (ennui) besides *impío* (impious, cruel), which is already available in "Resignación," our departure poem. On the same chart, we can access rhyme words that are likely to have positive connotations in this corpus, such as *pío* (devout) or *rocío* (dew). The case of *frío* (cold), the most frequent rhyme for *sombrío*, is ambiguous: it could be positively connoted if seen as relief from extreme heat but also negatively connoted as an unpleasant sensation. As stated above, the rhyme view shows, besides the aggregated data, a table with both lines where the rhyme pair occurs, and we can access the complete poem text from there (Fig. 4). In the case of words whose polarity is ambiguous, full-text access helps to determine which polarity the rhyme words have in each case and whether the rhyme pair agrees or contrasts in terms of positive/negative connotation. In the specific case of *sombrío*, we can see that it is used mostly negatively and reinforces a negatively connoted rhyme word, although some examples with positive connotations or contrast within the rhyme pair can be quickly found by inspecting the results on the interface.¹³ Besides the

13 In "Abril y Amor" (April and Love) by Cuban writer José Jacinto Milanés, most rhyme words convey positive emotion (<http://prf1.org/discover/showpoem.php?id=1905> [accessed April 20, 2022]). In "¡Ave, César!" (Hail, César!), the quatrains contrast positively and negatively connoted words rhyming with *sombrío* (<http://prf1.org/discover/showpoem.php?id=1905> [accessed April 20, 2022]).

sombrío example, other rhyme words could be analyzed in order to systematically collect data about positive/negative connotation in rhyme pairs.

Our main point is that a navigation interface that allows users to access aggregated data on rhyme or other features while studying individual texts, and to go back to individual texts from the quantitative data view, can help them to iteratively refine hypotheses about those features' behavior in a text or in the corpus at large. The point of departure for analysis can be either the aggregated level or the single-text level, and DISCOVER is a tool facilitating this procedure.

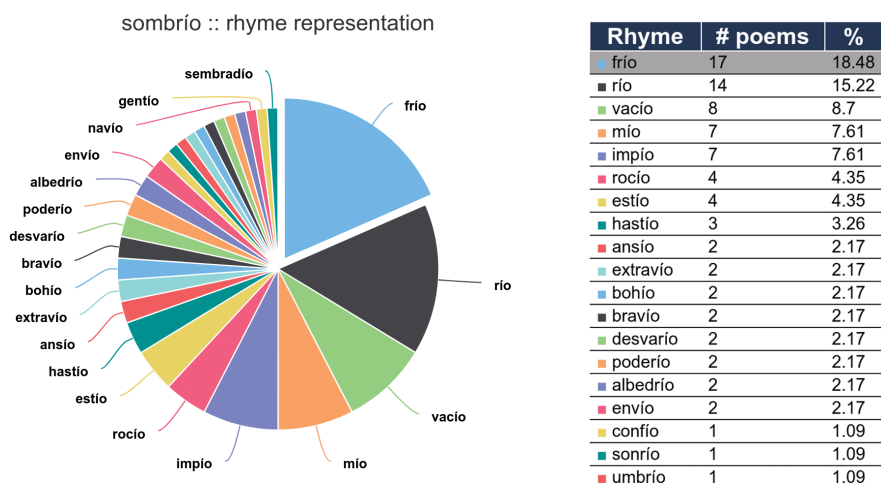


Fig. 5: Corpus-level overview of rhyme pairs for the word *sombrío* (shaded, somber) in pie chart and table format. This is the top-left pane of the interface's rhyme view (Fig. 4), and it allows us to select rhyme pairs in order to display their context (bottom panel in Fig. 4) and subcorpus distribution (top-right in Fig. 4).

4.3 Technical remarks

Two SQLite databases were used: one for rhymes and the other for text and author metadata. The back end has been programmed in PHP. The front end is mainly coded in HTML5/JavaScript.¹⁴

¹⁴ The interface repository with no safety-sensitive materials is available at <https://github.com/HelenaSabel/DISCOVER> (accessed April 20, 2022).

Interactive charts rely on the Highcharts JavaScript framework.¹⁵ Thanks to this framework, all charts produced on the interface can be exported in vector or bitmap format. The ability to download the data itself would be a useful future addition.

There has been minimal use of RDFa, and it will be extended in the future by utilizing the richly encoded source dataset (see Section 3).

We are committed to developing a fully accessible interface. Continuous efforts are being made to become fully compliant with the *Web Content Accessibility Guidelines* (Caldwell et al., 2008).

5 Conclusion

DISCOVer was built with a tool in mind that would allow researchers to perform and combine complex queries to address their specific research questions. The tool is intended to help enable a circular reading process: from distant reading through prosodic plots to the close reading of texts with their detailed metrical analyses, and back again to distant reading based on another feature like rhyme. DISCOVer's functions thus help users to identify trends and outliers, contributing to our knowledge of the sonnet in Spanish. Similar functionality could be implemented with the same benefits in interfaces for other poetry corpora.

Future work will involve two areas: on the one hand, enriching the annotations of the DISCO corpus, thus opening the door to new functions in DISCOVer; on the other hand, expanding the search engine. Our plans include adding sentiment analysis to the rhyme words and a classification of imperfect rhymes. We also intend to lemmatize at least the rhyme words, thus enabling queries to be made by both form and lemma. With regard to the utilization of features that have already been encoded, we will add a metrical pattern search: the goal is to enable the circular reading of metrical schemes along the lines of what we have implemented for rhyme analysis.

In addition, and following up on our studies on what impact exposing students to automatic prosodic annotation has on their comprehension of those prosodic features and their grasp of digital humanities methods (Martínez Cantón, Ruiz Fabo, 2019), we would like to test DISCOVer in a pedagogical setting. We would be interested in assessing whether using the interface improves students' understanding of the sonnet form in Spanish and its evolution.

¹⁵ <https://github.com/highcharts/highcharts> (accessed April 20, 2022).

References

- Abdul-Rahman A, Lein J, Coles K, Maguire E, Meyer M, Wynne M, Johnson CR, Trefethen A, Chen M. Rule-based Visual Mappings – With a Case Study on Poetry Visualization. *Computer Graphics Forum* 2013; 32(3pt4): 381–390. <https://doi.org/10.1111/cgf.12125>.
- Agenjo X. Las bibliotecas virtuales españolas y el tratamiento textual de los recursos bibliográficos. *Ínsula: revista de letras y ciencias humanas* 2015; 822: 12–15.
- Agirrezabal M. Automatic Scansion of Poetry. PhD thesis. University of the Basque Country, 2017.
- Biblioteca Virtual Miguel de Cervantes. Biblioteca Virtual Miguel de Cervantes. 1999. <http://www.cervantesvirtual.com> (accessed April 20, 2022).
- Biblioteca Virtual Miguel de Cervantes. Biblioteca del Soneto. 2007. https://www.cervantesvirtual.com/portales/biblioteca_del_soneto/ (accessed April 20, 2022).
- Caldwell B, Cooper M, Guarino Reid L, Vanderheiden G, editors. Web Content Accessibility Guidelines (WCAG) 2.0. W3C, 2008. <https://www.w3.org/TR/WCAG20/> (accessed April 20, 2022).
- Calvo Tello J. Review of Corpus of Spanish Golden Age Sonnets by Borja Navarro Colorado, María Ribes Lafoz and Noelia Sánchez (ed.). *RIDE* 2017; 6: <http://ride.i-d-e.de/issue-6/corpus-of-spanish-golden-age-sonnets/> (accessed April 20, 2022).
- Candela G, Escobar P, Navarro-Colorado B. In Search of Poetic Rhythm: Poetry Retrieval Through Text and Metre. In: *DATECH2017: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*. 2017: 53–57. <https://doi.org/10.1145/3078081.3078085>.
- Choral Public Domain Library (CPDL), 1998–. <http://www2.cpd.org> (accessed April 20, 2022).
- Chaturvedi M. Visualization of TEI Encoded Texts in Support of Close Reading. MSc thesis. Miami University. Oxford, OH, 2011.
- Chaturvedi M, Gannod G, Mandell L, Armstrong H, Hodgson E. Myopia: A Visualization Tool in Support of Close Reading. In: *Digital Humanities 2012*. Hamburg: Hamburg University Press, 2012: 148–150.
- de la Rosa J, Pérez Á, Hernández L, Ros S, González-Blanco E. Rantanplan, Fast and Accurate Syllabification and Scansion of Spanish Poetry. *Procesamiento del Lenguaje Natural* 2020; 65: 83–90.
- Delmonte R. Visualizing Poetry with SPARSAR – Visual Maps from Poetic Content. In: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver, CO: Association for Computational Linguistics, 2015: 68–78.
- Elf Edition. Sonett-Archiv. <http://sonett-archiv.com> (accessed September 22, 2018).
- Gago Jover F. La biblioteca digital de textos del español antiguo (BIDTEA). *Scriptum Digital* 2015; 4: 5–36.
- Gervás, P. A Logic Programming Application for the Analysis of Spanish Verse. In: *Computational Logic – CL 2000*. Berlin: Springer Berlin Heidelberg, 2000: 1330–1344.
- Henny-Krahmer U, Neuber F. Criteria for Reviewing Digital Text Collections, Version 1.0. A Review Journal for Digital Editions and Resources 2017; 6: <https://www.i-d-e.de/publikationen/weitereschritten/criteria-text-collections-version-1-0> (accessed April 20, 2022).
- Huber A, editor. Eighteenth-Century Poetry Archive. Eighteenth-Century Poetry Archive (version 1.2). 2020. <https://www.eighteenthcenturypoetry.org/index.shtml> (accessed September 29, 2020).

- Jockers ML. *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press, 2013.
- Marco Remón G, Gonzalo, J. Escansión automática de poesía española sin silabación. *Procesamiento del Lenguaje Natural*, 2021; 66: 77–87. <https://doi.org/10.26342/2021-66-6>
- Marcos Marín F, Faulhaber CB (coord.). *ADMYTE. Archivo Digital de Manuscritos y Textos Españoles*, 1992. <http://www.admyte.com/admyteonline/contenido.htm> (accessed April 20, 2022).
- Martínez Cantón C, Ruiz Fabo P. La evaluación de herramientas como modo de aprendizaje e introducción de las Humanidades Digitales en el aula universitaria: La experiencia docente “Poesía distante.” *Didáctica (Lengua y Literatura)* 2019; 31: 171–190. <https://hal.archives-ouvertes.fr/hal-02421983v2> (accessed April 20, 2022).
- Martínez Cantón C, Ruiz Fabo P, González-Blanco E, Poibeau T. Enjambment Detection as a New Source of Evidence in Spanish Versification. In: Bories AS, Purnelle G, Marchal H. *Plotting Poetry: On Mechanically Enhanced Reading*. Liège: Presses Universitaires de Liège, 2021: 93–112.
- McCurdy N, Lein J, Coles K, Meyer M. Poemage: Visualizing the Sonic Topology of a Poem. *IEEE Transactions on Visualization and Computer Graphics* 2016; 22(1): 439–448. <https://doi.org/10.1109/TVCG.2015.2467811>.
- McGann JJ. *How to Read a Book*. In: *The Textual Condition*. Princeton, NJ: Princeton University Press, 1991: 101–128.
- Meneses L, Furuta R, Mandell L. *Ambiances: A Framework to Write and Visualize Poetry*. In: *Digital Humanities 2013*. Lincoln, NE: University of Nebraska–Lincoln, 2013: 307–309. <http://dh2013.unl.edu/abstracts/ab-365.html> (accessed April 20, 2022).
- Mittmann A, Wangenheim A, Luiz dos Santos A. Aoidos: A System for the Automatic Scansion of Poetry Written in Portuguese. In: *CICLing, 17th International Conference on Intelligent Text Processing and Computational Linguistics*. Cham: Springer, 2016: 611–628.
- Moretti F. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London, New York: Verso, 2005.
- Navarro-Colorado B. A Computational Linguistic Approach to Spanish Golden Age Sonnets: Metrical and Semantic Aspects. In: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver, Colorado: Association for Computational Linguistics, 2015: 115–113.
- Navarro-Colorado B. Hacia un análisis distante del endecasílabo áureo: Patrones métricos, frecuencias y evolución histórica. *Rhythmica: Revista española de métrica comparada* 2016; 14: 89–118.
- Navarro-Colorado B. A Metrical Scansion System for Fixed-Metre Spanish Poetry. *Digital Scholarship in the Humanities* 2017; 33(1): 112–127. <https://doi.org/10.1093/lhc/fqx009>.
- Navarro-Colorado B, Ribes Lafoz M, Sánchez N. Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation. In: *Proceedings of the Language Resources and Evaluation Conference*. Paris: ELRA, 2016: 4360–4364. http://www.lrec-conf.org/proceedings/lrec2016/pdf/453_Paper.pdf (accessed April 20, 2022).
- Plecháč P. A Collocation-driven Method of Discovering Rhymes (in Czech, English, and French Poetry). In: Fidler M, Cvrček V, editors. *Taming the Corpus: From Inflection and Lexis to Interpretation*. Cham: Springer, 2018: 79–95.
- Plecháč P, Ibrahim R. *Gunstick – Database of Czech Rhymes*. Prague: Institute of Czech Literature AS CR, 2013: <http://www.versologie.cz> (accessed April 20, 2022).
- Plecháč P, Kolár R. The Corpus of Czech Verse. *Studia Metrica et Poetica* 2015; 2(1): 107–118.

- Quilis A. Estructura del encabalgamiento en la métrica española. Madrid: Consejo Superior de Investigaciones Científicas, Patronato Menéndez y Pelayo, Instituto Miguel de Cervantes, 1964.
- Ruiz Fabo P, Bermúdez Sabel H, Martínez Cantón C, González-Blanco E. The Diachronic Spanish Sonnet Corpus (DISCO): TEI and Linked Open Data Encoding, Data Distribution and Metrical Findings. *Digital Scholarship in the Humanities* 2021; 36(Supplement_1): i68–i80. <https://doi.org/10.1093/llc/fqab067>.
- Ruiz Fabo P, Martínez Cantón C, Poibeau T, González-Blanco E. Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets. In: *LaTeCH-CLFL 2017, Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Vancouver: Association for Computational Linguistics, 2017: 27–32.
- Tucker HF. Poetic Data and the News from Poems: A “For Better for Verse” Memoir. *Victorian Poetry* 2011; 49(2): 267–281.

Chris Mustazza

In Search of the Sermonic: Machine Listening and Poetic Sonic Genre

Abstract: This chapter argues that poets' performances of their work are often shaped by sonic genres apposite to, or altogether outside, reading styles recognized as "literary." Poetic performances include instances of political radio speeches, sermons, vaudeville monologues – all voices that poets "do" as a dimension of sonic form. Departing from this premise, the chapter introduces a method of machine listening based on hypothesis testing to identify sonic genre across a heterogeneous corpus of sound recordings. This preliminary study is dedicated to discussing prospective methodology. The materials under discussion are James Weldon Johnson's sermon-poems from *God's Trombones* in comparison with recordings of a "Black Diamond Express to Hell" by Rev. AW Nix.

1 Introduction

A well-known fact about TS Eliot's *The Waste Land* is that the poem was originally titled *He Do the Police in Different Voices*, before Ezra Pound struck the Dickensian allusion in favor of a more barren title. And, indeed, the most discussed aspect of *The Waste Land* is its use of heteroglossia, the murmuring and contrapuntal voices that destabilize a reader's ability to determine who is speaking any given part. Hearing Eliot perform the ragtime rhythms of "The Shakespearean Rag" and mime the working-class accents in the poem's pub scene creates an altogether different feeling than attempting to navigate the written poem; one might say that Eliot's reading even smooths the disjunction of the poem's heteroglossia. And the concept of heteroglossia, understood literally, as

Acknowledgment: I would like to thank Mark Liberman and Charles Bernstein for all their ongoing mentorship on these topics and many others. I am grateful to all of the organizers of the Plotting Poetry conference over the years I have attended, including Anne-Sophie Bories, Burkhard Meyer-Sickendiek, Petr Plecháč, and Véronique Montément, for giving me the venue to develop these ideas. I would also like to thank Erin Glass at UC San Diego, the English department at the University of Oregon, and the English department at the University of Pennsylvania for giving me a chance to refine my thinking on these topics. The Price Lab for Digital Humanities at the University of Pennsylvania provided generous support for this research.

Chris Mustazza, University of Pennsylvania, e-mail: mustazza@sas.upenn.edu

Eliot is performing the different voices in his poems, is where I would like to begin this conversation about how digital tools might provide new ways to approach poetic audiotexts and expose new dimensions of poetic form.

Modernist studies has long been fascinated by the question of the so-called “I of the poem,” the unified speaking voice that characterizes lyric poetry and which came undone during the modern period. But most discussions of this topic hold firmly to the traditional and metaphorical notion of “the voice” in poetry, understood as a singular stylistic element accessible through the written document of a poem. In the case of Eliot and countless others, sound recordings exist of poets performing the voices imbricated in their poems. In our current age of digital audio processing, we can now go beyond the idea that these voices, as performed by the poets, could provide a site for what Charles Bernstein (1998) has termed a close listening, an aural hermeneutics that can function with or against the grain of the written poem. This chapter will argue that the voices on these recordings can be studied as data, using computational means, and that these machine listenings can use large datasets to provide new readings of individual poems.

The idea of a corpus or dataset is good place to get started. We generally tend to conceive of sets of cultural objects as archives, collections sheltered for posterity in the institutional domiciles that Derrida traced to the ancient Greek archons or magistrates (1998). PennSound, for example, is the world’s largest *archive* of poet recordings (PennSound, n.d.). Now over 15 years old, it contains over 55,000 poetry recordings and 6,000 hours of audio of poets, going back to Apollinaire in 1913 and stretching all the way to contemporary recordings. In addition to those who use PennSound for close listening research or for personal edification, a set of scholars is beginning to consider this unprecedented sound collection as a dataset, pursuing questions on a scale that was heretofore impossible or approaching traditional literary historical questions via the new vector of digital sound. This is to say that these scholars work with the audio files directly, not as a way to get “closer” to a primary, written text.

Several scholars have taken up the challenge of working with literary audiotexts as digital media. Tanya Clement’s distant listening, her sonic analogue to distant reading, asks questions of sound archives at scale: do first-wave modernist poets perform differently than those in the second wave? How can we measure applause in recordings as a dimension of reader response, and how does this change at different historical moments covered by the archive (Clement, McLaughlin, 2015)? Or, a question I’ve worked on myself: can the provenance of a recording be determined by the recording noise captured on the record – is noise part of the content (Mustazza, 2015)? In distant listening, these questions are approached using machine listening techniques to parse

thousands of hours of audio in a condensed period of time. Marit MacArthur's work on "poet voice," the term she gives to the dominant reading style of the contemporary American academic poetry reading, utilizes methods from the field of phonetics to describe the sonic dynamics of poetic performance and to connect these dynamics to social and historical factors (MacArthur, 2016). Mark Liberman has developed his own visualizations of poetic prosody to study the way poets use pauses as part of their poetics (Liberman, 2016). My own work on machine-aided close listening and comparative machine-aided close listening works to render the sonic materiality of the voice legible in support of close listenings of individual poems (Mustazza, 2018). All in all, these angles of approach seek to make use of the actual voices on recordings and to offer sound-based approaches to poetic prosody.

I will begin this chapter with an observation I made while pursuing my exploration of the media history of poetic performance. While digitizing and editing early sound recordings of poetry, I was struck by the performance styles of modern poets. In contrast to the staid contemporary performances that MacArthur terms "poet voice," these performances seemed to draw on genres altogether outside of those recognized as literary recitation, including some that are and have been denigrated as sub-literary. For example, James Weldon Johnson, whom I will discuss in this essay, performed his famous sermon-poems from *God's Trombones* in the vocal cadences of African American sermons (Johnson, 2014). FT Marinetti's bombastic and bellicose performances connect him to the Mussolini speeches that would draw *Il Futurismo* into its inevitable descent into *Il Fascismo* (Marinetti, n.d.). Louise Bennett performed her poems in devastating comedic monologues that could be termed tragicomic (Bennett, 2011). And so, I wondered if machine listening could be used to identify sonic genre. This is easier said than done, obviously. To do so would mean empirically defining aspects of the genres in a manner legible to modern computing. This would include describing pitch, duration, pauses, tempo, and choosing from a range of literally thousands of measurable indices to try to describe what it is we hear when we say that something sounds like a political speech. This chapter is dedicated to beginning that process and pointing toward a larger undertaking that might use machine listening to categorize poetry by sound rather than by the written word. I will offer a case study using the sermon/sermonic poetry as an example and will focus primarily on methodology and the challenges of empirically describing the voice.

2 Sermonic voices

In 1927, James Weldon Johnson released *God's Trombones*, his collection of poems modeled on the oral tradition of African American sermons (Johnson, 1927). A work wholly modeled on speech sounds, *God's Trombones* needed to convey sound when the primary, if not only, means of distribution for poetry was writing. Following on from his background as a composer of Tin Pan Alley musicals, Johnson treated the printed page as a score, a libretto to be reconstituted – just add voice. In the preface to the collection, Johnson provides a text-to-sound legend – call it a codec for the human machine or perhaps an interface. He writes:

The tempos of the preacher I have endeavored to indicate by the line arrangement of the poems, and a certain sort of pause that is marked by a quick intaking and an audible expulsion of the breath I have indicated by dashes. There is a decided syncopation of speech – the crowding in of many syllables or the lengthening out of a few to fill one metrical foot, the sensing of which must be left to the reader's ear. (Johnson, 1927: 10–11)

In other words, Johnson used paper as a sound recording medium in the absence of the scarce resources needed to make records. But in 1934, Johnson would make sound recordings of his textual recordings that represented the speech of the preachers. He would do the clergy in different voices.

The sermon-poems of *God's Trombones* may have been inspired by the preacher AW Nix. In the preface to *God's Trombones*, Johnson notes: “I heard only a few months ago in Harlem an up-to-date version of the ‘Train Sermon.’ The preacher styled himself the ‘Son of Thunder’ – a sobriquet adopted by many old-time preachers – and phrased his subject ‘The Black Diamond Express, running between here and hell, making 13 stops and arriving in hell ahead of time’” (1927: 2). Around the time of Johnson's writing, Nix was touring the country and also releasing a series of very popular sermon records, including a performance of “Black Diamond Express to Hell.” As a way to begin approaching the question of describing the sermonic, I assembled a corpus of African American sermon recordings from this era, including several by Nix. The primary question I had for this archive was whether there is a sonic relationship between Johnson's sermonic and vocal recordings that announce themselves as sermons. If so, is it possible to define this sound prosodically through a process of hypothesis testing? And if it is, could a machine locate instances of the sermonic across a heterogenous corpus of materials, say, across the entire PennSound or an even more diverse corpus?

I would like to focus on this process of hypothesis testing as methodology. There are multiple ways to approach a machine learning project like this. One

way is by means of “unsupervised learning,” which means, in a sense, letting your machine loose on a dataset and having it locate correlations. This is problematic in a project like this for a number of reasons. For one, variables can, statistically speaking, appear to be correlated even when there is no meaningful connection between them. When Tanya Clement and I were working on an early instantiation of her project, the machine found similarities between an Allen Ginsberg recording and a snippet that sounded like a refrigerator hum. Is there a correlation? Perhaps in some machinic episteme. Perhaps less so for literary studies. For this project, I am building toward a supervised learning model that begins with an impressionistic close listening. In other words, what makes the sermons sound like sermons to me? It could be said that the human ear is the most sophisticated audio device that we have – especially if, following Marshall McLuhan, we think of all audio technology as a prosthetic extension of the human ear. So, I began with the question of what I heard when I listened to Nix, which would precede any attempts to corroborate or refute my hypothesis using digital visualizations.

I began with recordings of Nix performing “Black Diamond Express to Hell” (1995), the sermon Johnson cites as an influence. The first dynamic I noted was that Nix appeared to speak some sections in short intonational phrases that ended on an upward pitch, similar to a question, or so-called uptalk. This stands to reason, given the call-and-response structure of a sermon. The phrases’ pitches terminating in a crescendo would cue the congregation to respond within the structure. The next thing I noticed could be described as segments of high vocal effort, which I had originally and amateurishly described to a linguist collaborator as “gravelly.” Note that vocal effort is not the same thing as loudness (a psycho-acoustic condition) or intensity. It speaks to how hard the vocal cords are working to produce a tone. The high degree of vocal effort can also be explained by the social and medial conditions of the sermon. At the time Nix was giving his sermons, the era of the microphone and electrical recording was in its infancy (not becoming standard in professional recording studios until around 1925). Thus, Nix would have been conditioned to project his voice to reach the back of the church, to connect with everyone, including those farthest away. The sonic dynamics of physical space here make their way onto the record, despite the presence of the microphone in the studio. And finally, I noticed segments of the recording where a series of short, staccato intonational phrases were delivered in rapid succession with a flat pitch. These phrases sounded like they were meant to foreground a rhythmically structured prosody that backgrounded pitch in favor of audience synchronization. In other words, the rhythms of speech were meant to bring the congregation together into a common time signature of sorts.

Beginning with these hypotheses, the next step was to try to visually confirm or refute them. I will present a number of visualizations of pitch and amplitude using Praat, a linguistic measurement tool developed for phonetics research. It is worth noting here that, while I will foreground pitch and amplitude using Praat, these are only two dimensions of a vast array that could be considered. The pitch visualizations will also focus on the fundamental frequency of the voice, termed F0, and will elide the harmonic frequencies that accompany it. These harmonic frequencies contain a rich amount of information, but for a study like this, it seemed best to me to start simply, with less variables.

2.1 Call and response

The first dynamic I discussed was the open structure of the call and response in the sermon, the intonational phrases that sounded like a question to me, which is to say that they sounded like they terminated on a rising pitch. Figure 1 displays a visualization of F0 for a segment of Nix’s performance of “The Black Diamond Express to Hell.” As we can see from the pitch curves, the phrases do indeed trend upward, confirming my hypothesis. If we were to program a classifier based on this dynamic and test it on a heterogenous corpus that included some conversational speech, we might find that this density of upward-trending,

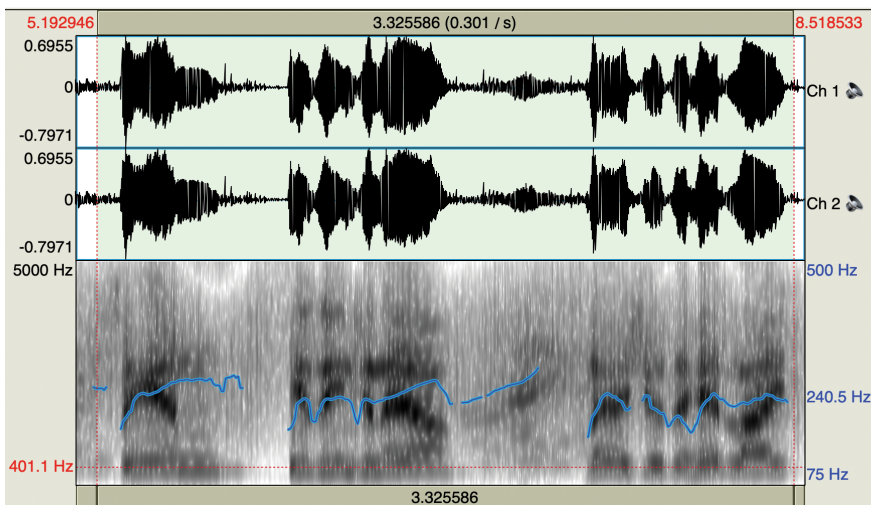


Fig. 1: Fundamental frequency (pitch curve) of a section of “Black Diamond Express to Hell” performed by AW Nix. Note the upward trend of each intonational phrase.

intonational phrases differentiates a sermon from kinds of speech that do not generally string multiple interrogatives together.

It is worth noting that these analyses cast light back on Johnson's recordings. For example, Johnson's pauses at the end of each line, which he uses to score the phrasing of the sermon-poems, mark a palpable lack: the lack of a congregation to give a response to his calls. In this way, the pauses are not the negative space to Johnson's voice but rather a forward element of poetic content – the lack of sound resounds. Nix's recordings include a theatrical response to his calls (theatrical because the recordings occur in a studio and are not live recordings), highlighting the intended structure of the sermon and differentiating it from its literary mimesis.

2.2 Vocal effort

The next dimension gives us more difficulty. I noted that Nix's recordings seemed to demonstrate sustained periods of high vocal effort or exertion of the vocal cords to produce the intensity of sound we perceive. There is no empirically decisive way to measure vocal effort (Ford Baldner, Doll, van Mersbergen, 2015), but there are many studies that attempt to do so using various proxies. Much of this research is psychoacoustic, in that it measures perceived vocal intensity. For example, one study measures how higher vocal effort might affect the perception of vowels (Liénard, Di Benedetto, 1999), which requires using multiple vocal formants to measure. In this example, I conjecture that a different proxy may be the relative curves of F0 and intensity (the latter of which can correlate with the psychoacoustic condition of loudness).

Figure 2 demonstrates a section of Nix's performance that I initially described in a conversation with Mark Liberman as "gravelly." Liberman explained to me that one explanation for such a texture could be a chaotic distribution of pitch, which comes from overexertion of the vocal cords. Another can be what is generally referred to as vocal fry, which suggests an attempt to push the pitch of one's voice lower than its range. Of course, the former explanation, the idea of a chaotic distribution of pitch, could certainly explain the sound I perceived given that Nix gets very absorbed in his performances. But periods of vocal fry are also a possibility. For example, toward the end of Fig. 2, we can see places where the pitch (the blue line) of Nix's voice drops into a plateau while the intensity of his voice (the yellow line) stays higher and pulls away from it. One possibility here is that the drop in pitch suggests a vocal fry that may differ from the vocal fry found in conversational speech or other kinds of speech due to the coexistence of a higher amplitude.

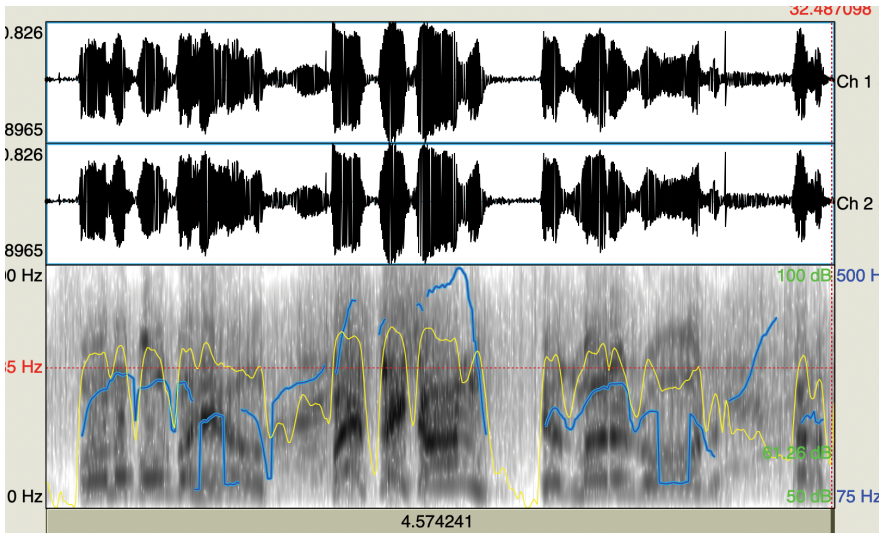


Fig. 2: Fundamental frequency and intensity of a section of “Black Diamond Express to Hell,” performed by Rev. AW Nix.

Other dimensions, such as coeval surges in pitch and amplitude, could suggest more of a chaotic distribution that would render perceptually as gravelly. In other words, both cases detailed above may hold true. So, the question (to be answered another day) will be whether it is possible, and if so how, to program a classifier that can identify this kind of vocal effort.

Again, to come back to the Johnson recordings, we do not hear this kind of vocal intensity. There could be several explanations for this. One is that, while sermons are Johnson’s model for his poems, Johnson’s readings are not sermons. Thus, his performances could be said to be more “literary,” connoting a more understated performance style that derives from reading. Another reason could be that the recordings were made in a dialect lab at Columbia University (Mustazza, 2016). The sterile conditions of a dialect lab, which was used at least in part for lingual standardization and the teaching of elocution, might have made Johnson feel uncomfortable about performing too far outside the understated speaking styles that were privileged by this space. On a related note, Johnson would have been acutely aware of the racial biases of the space. Given his fraught relationship with dialect for reasons of its reception by non-dialect speakers, Johnson may have sought to suppress the more theatrical aspects of the sermon lest these cadences be reduced to pathos or humor in their reception.

2.3 Synchronous rhythms

For all of Nix’s animated dynamics, there are sections of his performances that are subdued to the point that the contrast is noticeable. Figure 3 is an example that sounds strikingly different than any kind of conversational speech. Nix quite literally shifts into monotonous speech, suppressing the pitch dynamics of his intonational phrases. These phrases are not MacArthur’s “monotonous incantation” nor the uptalk described earlier. The intonational phrases occupy a tight range of pitch and end around the pitches where they began. At the same time, the intensity dynamics (the yellow lines) are also remarkably consistent across the phrases, creating a sonic feeling of repetition. Compare Fig. 3, for example, with Fig. 2. Between the pitch and the intensity, these phrases suggest an evenness, or a steady rhythm. This would likely be the easiest dynamic for which to write a classifier and to differentiate against conversational speech. The flattened dynamics, to my ear, seem to be a hallmark of many preachers’ styles, from Nix through Rev. Emmett Dickinson and Rev. JM Gates. My sense is that this dynamic

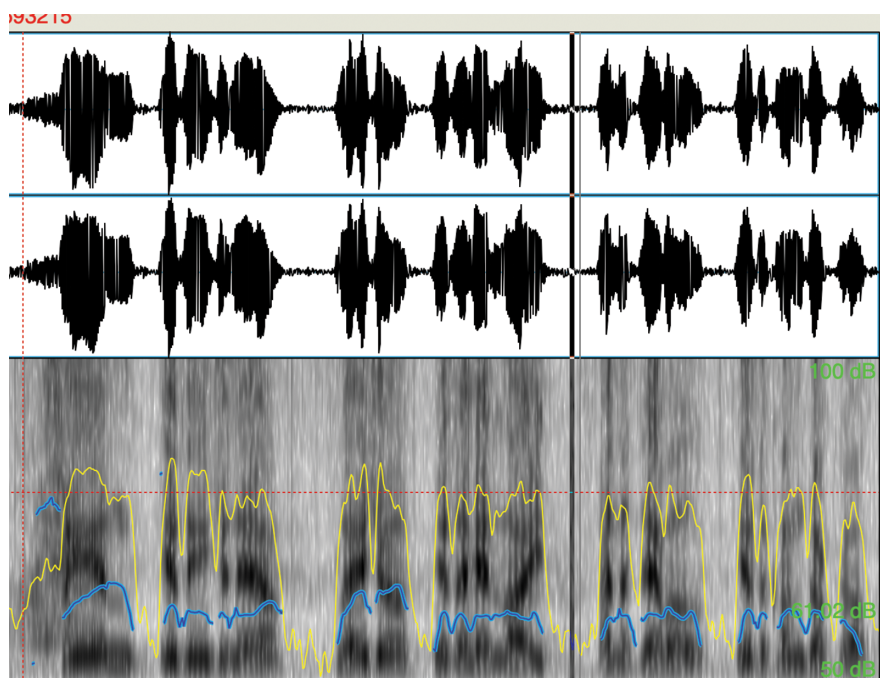


Fig. 3: Excerpt from Rev. AW Nix performing “Black Diamond Express to Hell,” with pitch (blue) and intensity (yellow) plotted.

could be the most useful when attempting to sort through a heterogeneous corpus looking for elements of the sermonic.

In addition to being potentially identificatory of the sermon genre, a notable aspect of this sonic dynamic is that it seems to foreground rhythm over pitch. By suppressing expressivity in the pitch of the intonational phrases, it could be said, the prosodic rhythms of language take center stage. Given the context, one reason to do this is to synchronize a congregation into a common rhythm, one that is organized around speech rather than music. The music that is latent in the language is expressed through intentionally plodding phrases that contrast themselves against sections such as the one displayed in Fig. 1. If there are segments of the sermon that ask for a response to an interrogatively phrased call, then these can be set against rhythmic sections meant to hold the general rhythm of the sermon, to keep the time as it were.

Returning to Johnson, we can hear similar phrases in his performances, albeit without the strong prosodic rhythms present in the sermons. In order to contrast sections where he crescendos his voice to mirror the narrative arc of the sermon, there are several understated phrases that occur within a tighter range of pitch. They do not bear the same kind of staccato feel as Nix's performance though. We could say, once again, that we are hearing the difference between a sermon-poem and a sermon, the latter of which imagines an audience present and seeks to interact with it, to move the congregation through the technology of speech rather than to enact the technology of writing.

All in all, these three examples are just proposals for where to begin. What makes sermons expressive? We will have to answer this question for ourselves before we will be able to teach a machine how to do it. But, as I have attempted to demonstrate here, we can learn a lot about the socio-historical structure of sermons and of poetry by just attempting to describe them. Description is indeed interpretation, and the domain of the empirical can be used to open up a proliferation of meanings rather than to solve a work or winnow down its possibilities. The next steps for this project will be attempting to further test these hypotheses by writing classifiers and running the code on a variety of corpora. We will learn even more from this process, especially from output that we think the machine "got wrong." In any event, paying attention to the actual sounds of poetry and speech will certainly open up new avenues for literary studies.

3 Conclusion: Collapsing the distance

Readers may notice that in each section above, where I describe the various sonic dynamics of Nix's sermon, I contrast these sounds with Johnson's performances. This may seem counterproductive, given that I am trying to explore whether studying sermons can allow us to define them as a vocal genre within poetry. My primary interest in this project is not to let an algorithm loose on a heterogeneous corpus and have it return every file/poem/performance that is itself a sermon, or even modeled on a sermon, even though such an undertaking would be valuable for a number of applications, including the ability of libraries and archives to computationally generate metadata and offer new ways to move through sound archives. Rather, I am interested in whether the sermonic or other genres can be recognized as a dimension of form within individual poems, which might contain several genres. If two or three lines of a poem were spoken as a sermon, how might we read this form as an extension or complication of those lines' content?

Here is a made-up example that returns us to Mr. Eliot. What if we were to find that elements of "The Fire Sermon" were modeled on a vaudeville monologue? Of course, we might not be shocked to hear that the "Fire Sermon" does not sound like a sermon all the way through, but we might get a new look into a dark poetics modeled on a tragicomic aesthetic. What would it mean for the disturbing story of Philomela to be rendered as an upbeat monologue or a radio play? What would it mean for Phlebas the Phoenician Sailor to take shape through a sermon genre? "He do the police in different voices," indeed. And these voices are present in the recordings, they are measurable, and they represent a dimension of form heretofore inaccessible. Perhaps by using machinic prostheses, we can learn to hear language as pure form, as its component phonemic crystals tessellated in unfurling time series. What better way to do so than to train an algorithm that is *incapable* of understanding content, that makes categorizations *only* on the basis of sound dynamics? The question of interpretation would then be left to the prosthetically extended reader, us. We can read and reread these poems together, considering new categorizations as frames that aid our hermeneutic processes rather than as Rosetta Stones to solve the poem. In other words, the algorithm and the interface are arguments. And so, the distance collapses. Scale comes to serve immediacy, the very large in support of the minute. Form becomes anything but formalist. It foregrounds the social, historical, biographical conditions of the audiotext's existence. Through it, the particular speech sounds of people, which writing can suppress or eliminate via lingual assimilation and standardization, are made audible.

References

- Bennett L. PennSound: Louise Bennett, 2011. <https://writing.upenn.edu/pennsound/x/Bennett.php> (accessed April 19, 2020).
- Bernstein C, editor. *Close Listening*. Oxford: Oxford University Press, 1998.
- Clement TE, McLaughlin S. Visualizing Applause in the PennSound Archive. Jacket2, Clipping, October 2015. <https://jacket2.org/commentary/clement-mclaughlin-pennsound-applause> (accessed August 20, 2020).
- Derrida J. *Archive Fever: A Freudian Impression*, transl. by Prenowitz E. Chicago: University of Chicago Press, 1998.
- Ford Baldner, E., Doll, E., & van Mersbergen, M. R. (2015). A Review of Measures of Vocal Effort With a Preliminary Study on the Establishment of a Vocal Effort Measure. *Journal of voice: official journal of the Voice Foundation*, 29(5), 530–541. <https://doi.org/10.1016/j.jvoice.2014.08.017>.
- Johnson JW. *God's Trombones: Seven Negro Sermons in Verse*. New York: Viking Press, 1927.
- Johnson JW. *The Speech Lab Recordings: Reading at Columbia University, December 24, 1935*, edited by Mustazza C. *Speech Lab Recordings*. Philadelphia, PA: PennSound, 2014. <https://writing.upenn.edu/pennsound/x/Johnson-JW.php>, (accessed March 19, 2020).
- Liberman M. "Poetic Sound and Silence." *Language Log* (blog), February 12, 2016. <https://languagelog.ldc.upenn.edu/nll/?p=24054> (accessed March 19, 2020).
- Liénard J-S, Di Benedetto M-G. "Effect of Vocal Effort on Spectral Properties of Vowels." *The Journal of the Acoustical Society of America* 1999; 106(1):411–22. <https://doi.org/10.1121/1.428140>.
- MacArthur M. "Monotony, the Churches of Poetry Reading, and Sound Studies." *PMLA* 2016; 131(1): 38–63.
- Marinetti FT. "PennSound: F.T. Marinetti." <http://writing.upenn.edu/pennsound/x/Marinetti.php> (accessed March 21, 2020).
- Mustazza C. "The Noise Is the Content: Toward Computationally Determining the Provenance of Poetry Recordings." Clipping, January 10, 2015. <https://jacket2.org/commentary/noise-content-toward-computationally-determining-provenance-poetry-recordings> (accessed August 20, 2020).
- Mustazza C. "James Weldon Johnson and the Speech Lab Recordings." *Oral Tradition* 2016; 30(1). <https://doi.org/10.1353/ort.2016.0001>.
- Mustazza C. "Machine-Aided Close Listening: Prosthetic Synaesthesia and the 3D Phonotext." *Digital Humanities Quarterly* 2018; 012(3). <http://www.digitalhumanities.org/dhq/vol/12/3/000397/000397.html> (accessed April 27, 2022).
- Nix AW. *Complete Recorded Works in Chronological Order*, vol. 1: 23 April 1927 to 26 October 1928. LP, vol. 1. Document Records. 1995. <https://www.discogs.com/Rev-AW-Nix-Complete-Recorded-Works-In-Chronological-Order-Vol-1-23-April-1927-To-26-October-1928/release/6220174> (accessed March 19, 2020).
- "PennSound," n.d. Accessed February 13, 2020. <https://writing.upenn.edu/pennsound/> (accessed March 19, 2020).

Éliane Delente

Can Relationships between Rhythm and Meaning in French Versified Poetry be Automated?

Abstract: Within the framework of a study on the relationships between rhythm and meaning in French versified poetry, this chapter discusses the limitations of the most common approach: aligning syntactic and metrical structures. This approach, in theoretical works as well as in automated processing, focuses on enjambments, a notion that proves unsatisfactory in several respects. I suggest analyzing the metrical expressions themselves, considering their beginning, their end, and their internal consistency while also trying to integrate the processing time of successive metrical expressions and specific reader expectations, as well as the time periods and the individual poets considered.

1 Introduction

The aim of this chapter is to examine the difficulties posed by studies on the relationship between rhythm and meaning, particularly those based on the automated processing of enjambment in versified poetry. I will first discuss the theoretical relevance of a traditional “overflow” approach, namely syntax/meter alignment, the approach most commonly used by automated enjambment detection programs. I will show how prosodic, semantic, pragmatic, and even discursive aspects of the phenomenon cannot be ignored.

In the second part, I will suggest another approach, better suited to examining the rhythmical expressions of versified discourse, and show how the aspects of the relationships between rhythm and meaning that can indeed be captured by the notion of enjambment are too few and not relevant enough.

Finally, I will try to outline the tasks we might reasonably expect an automatic program to perform.

Acknowledgment: I would like to thank Benoît de Cornulier for his accurate and meticulous reading of this chapter. I am particularly grateful for his thoughtful remarks, which have significantly improved it.

Éliane Delente, University of Caen Normandie, CRISCO – Crosslanguage Research Centre on Meaning in Context, e-mail: eliane.delente@unicaen.fr

2 Syntax/meter alignment

Most studies of enjambment, whether automated or not, focus on the syntactic boundary at the line break (see Fabb, 2015; Dell, Benini, 2021; Husein, Meyer-Sickendiek, Baumann, 2018; Ruiz et al. 2017). The units mapped are generally the sentence and its constituents on the one hand and, on the other, the line (with the notable exception of Dell and Bennini, who also consider hemistich and couplet boundaries [forthcoming]). Their alignment allows the identification of non-congruence cases between syntactic and line boundaries. Consider the following example:

- (1) Bienheureux, j’allongeai les jambes **sous la table**

Verte: je contemplai les sujets très naïfs

(Rimbaud, *Au Cabaret-vert*, *Poésies I*)

Happy, I stretched my legs out under the table,

A green one: considering the naïve prints

(Selective works in translation, A. S. Kline, www.poetryintranslation.com)

In Example 1, we generally identify an enjambment because the phrase “sous la table verte” straddles the line boundary. Such a formal characterization of line boundaries is necessary, yet its relevance with regards to syntax/meter alignment is rather limited for at least three reasons: Firstly, one wonders whether the first or second line is affected. Secondly, syntax is largely insufficient to adequately account for cases of discordance. Thirdly, by seeing these lines as two flat, complete structures, this approach fails to shed new light on rhythm as a dynamic event constituting a reader’s experience (see Auer, Couper-Kuhlen, Müller, 1999; Cornulier, 2003; Cornulier, 2009; Tsur, 2000; Paterson, 2018).

Regardless of genre, speech interpretation takes into account the flow of speech and the way speakers deal with this flow. Chafe has shown that our conscious mind can only focus on one event or state at a time, and that discourse reflects this limitation by progressively constructing intonational units (1994).

- (2) Discourse is a constant flow



- (3) Progressive construction of intonational units (.) (.) (.) (.)

Speakers and listeners focus their attention on one unit at a time according to the “one new idea per unit” constraint (Chafe, 1994: 108–119). This approach

accounts for the permanent tension between the continuity of the discursive flow, which unfolds over time, and the constant stops. William mentions “an alternation of flights and perchings” (1890, 1:243), and Cornulier “une alternance de marches et de pauses-bilans” (an alternation between walks and assessment pauses; private conversation). These speech processing conditions are true of both spoken and written language.

The same idea can be found in studies of prosody. In Simon (2011), Degand and Simon (2005), Degand, Fabricius-Hansen, and Ramm (2009), Simon and Degand (2011), and many other works as well (Auchlin, Ferrari, 1994; Golomb, 1979; Lacheret, Victorri, 2002; Martin, 2004; Rossari, 1996; Roulet, 1999; Roulet, Filliettaz, 2001; Roulet, 2002), prosodic units form the basic speech processing units. Sensible prosodic boundaries signal a space to cognitively process the preceding information and the step-by-step construction of speech coherence.

2.1 Syntax and prosody

Syntax is insufficient to account for the relationships between rhythm and meaning, whether in oral genres or in versified discourse. In a study based on oral corpora, Degand and Simon (2011) have shown how both syntax and prosody contribute to the construction of basic speech units. Some units are congruent (Example 4), while others are discordant (Example 5):

- (4) (Anne-Marie) (adore Julien)
 (Anne-Marie) (loves Julien)
- (5) (Anne-Marie adore) (Julien)
 (Anne-Marie loves) (Julien)

The choice of a grouping forms part of a discursive strategy, and we must note that cases of discordance are in no way specific to versified discourse: what is referred to as “enjambments” have a well-known prosody equivalent called “asyntagmatic units.” Such units may serve a number of purposes, of which I will mention only two here.

2.1.1 Pauses

Some pauses are expected. Others occur in unexpected places, placing a particular emphasis on what comes next, which is precisely what is called “rejet” in

versification. Still others can create a double meaning when the element added after a pause forces the reader to reinterpret the meaning of the preceding word. All these interruptions, disfluencies, and conflicts between structures deserve, I think, special attention.

2.1.2 Eurhythmic principle

In order to balance successive units, speakers tend to build up intonational units of substantially equal duration, regardless of the number of syllables.

- (6) (Marie) (voudrait du chocolat chaud),
(Marie) (would like some hot chocolate),

The imbalance in the number of syllables leads to a slower pace on “Marie” (two seconds) and a faster one on “voudrait du chocolat chaud” (eight seconds). But the eurhythmic principle (see, in particular, Martin, 2004) can also lead to different prosodic constructions and contradicting syntax:

- (7) (Marie voudrait) (du chocolat chaud)
(Marie would like) (some hot chocolate)

There are therefore prosodic and rhythmic principles at work in oral genres that interact with syntactic relationships and sometimes come into conflict with them, making it possible to construct asyntagmatic units (see Dubeda, 2004).

3 The object of study: Unit boundaries or units themselves?

A second difficulty relates to the object of study. In studies of syntax, and sometimes in studies of prosody, characterizing unit boundaries often takes precedence over analyzing the units themselves, as if boundaries were needed to demarcate and constitute otherwise undistinguishable units.

As with different oral genres, each written genre sets up its own units. With regard to poetic versified discourse, it goes without saying that:

- many of its rhythmical properties utilize the prosodic properties of the language in question;
- in the absence of acoustic clues, it is the layout in lines and line groupings that invites the reader to build units of metrical discourse;
- versified poetry adds to the rhythmic properties of language an equivalence principle, which invites the reader to perceive metrical expressions as systematically similar in some respects (metrical regularities);
- finally, the perceived equivalence of metrical expressions is an event of a temporal nature, involving the step-by-step processing of speech.

Metrical expressions are now well established, thanks in particular to Cornulier’s works. (From now on, I will refer to “metrical expressions” as “MEs.”) For example, in the quatrain below, brackets highlight different MEs, namely the hemistich, the line, the first stanza module, the second stanza module, and the stanza. For a prescribed form like the sonnet, we could also label the octave and the sestet:

1- {{Tu te plais à plonger} {au sein de ton image; }}	}	first stanza module	}	stanza
2- {{Tu l’embrasses des yeux} {et des bras, et ton cœur}}				
3- {{Se distrait quelquefois} {de sa propre rumeur}}	}	second stanza module		
4- {{Au bruit de cette plainte} {indomptable et sauvage.}}				

(Baudelaire, *L’Homme et la mer*, Les Fleurs du Mal)

- 1- You like to plunge into the bosom of your image;
- 2- You embrace it with eyes and arms, and your heart
- 3- Is distracted at times from its own clamoring
- 4- By the sound of this plaint, wild and untamable.

(Baudelaire, *Man and the sea*, *The Flowers of Evil*
(Fresno, CA: Academy Library Guild, 1954))

The reader’s mind focuses on one ME as a basic constituent, whether it is a simple verse or a complex verse (associated hemistichs). These constituents are restricted MEs, grouped into broader MEs (stanza modules, and stanzas) that express larger units of coherent information. These are units of verbal thought (Cornulier, forthcoming), units of semantic and pragmatic construction, and, for the broader levels, of constructing discursive coherence. Paterson develops the same idea: “[. . .] poets and theorists of verse should not make the mistake

of discussing poetic metre outside the context of the speech act to which it is inextricably bound” (2018: 343).

In this context, the traditional definition of enjambment used in automated processing programs, according to which a syntactic unit is divided by the line boundary, is unsatisfactory:

- (8) Now in loose Garlands thick thrown off, **the bright Pavement** that like a Sea of Jasper shone

(Milton, *Paradise Lost*)

As Cornulier puts it, since the linguistic unit is divided into two consecutive lines, it is no longer, in itself as a unit, an ME (2003). As we can see in Milton’s enjambment, “the bright pavement” is a noun phrase, not an ME, while lines 1 and 2 are the most evident MEs. This observation brings me to the third difficulty.

Although the line must be taken into account as an ME, there is no doubt whatsoever that a systematic examination of discordances would show that they differ as to their strength and distribution, depending on the level of the ME being considered. More specifically, units rather than boundaries should be favored. Therefore, the study should focus on characterizing the types of ME relative to their level in the metrical hierarchy and on the way the reader perceives them: as more or less complete – that is to say, more or less natural and familiar – or, in contrast, as more or less incomplete, suspensive, and therefore surprising, even disconcerting, etc.

4 A rhythmical approach

4.1 A work in progress

Some works defend an approach based on speech flow and its step-by-step, unit-by-unit examination, which implies a temporal construction, both in production and in reception (Mourgues, 1724; Tynianov, 1977; Golomb, 1979; Cornulier, 1977, Cornulier, 1982, Cornulier, 1995, Cornulier, 1997b; Auer, Couper-Kuhlen, Müller, 1999; Cornulier, 2000, Cornulier 2003; Tsur, 2008, Tsur, 2012; Paterson, 2018). This means that reading time is a factor inherent in the construction of MEs and the coherence of speech.

The reader’s attention cannot focus on an object as large as a poem or even a stanza but only on one ME at a time. The perception of rhythm is thus reflected in the reader’s experience by a dynamic of more or less satisfied or frustrated

expectations, which are part of the process of interpretation (see Tynianov, 1924; Golomb, 1979; Cornulier, 1977, Cornulier, 2000, Cornulier, 2009; Tsur, 2008; Paterson, 2018).

This last point is difficult to objectify since the reader's expectations, whether syntactic, prosodic, or metrical, vary depending on the extent to which the reader is acquainted with versified poetry, which implies parameters such as line length, the state of the language and the metrical system being considered, meter, the type of MEs, the type of reading (first reading or re-readings that sometimes cause rhythmical reinterpretations), as well as poetic genre. The reader's expectations in terms of rhythm/meaning relationships are less constrained in the comedy, fable, story, and funny epistle than in tragedy and lyric poetry, for example.

4.2 The reader's expectations

4.2.1 The banal "enjambment"

We have seen that a non-congruent prosodic unit like (*Marie voudrait*) (*du chocolat chaud*) is not rare in conversation. Likewise, in most poetic traditions, versified discourse produces a large number of hemistichs and lines characterized by their syntactic and semantic incompleteness. Therefore, a subject or direct object, for instance, provided its lexical and prosodic weight is sufficient, may appear outside the line of the verb that governs it. It is part of poetry readers' expectations (variable depending on the period and the poet). This phenomenon is perceived as quite ordinary by the reader, and there is no reason to suppose a stylistic effect every time an ME ends in this manner. For spectacular enjambments, Mourgues speaks of "vicious" enjambement (1724: 171), which implies that some others can be "frictionless" (Paterson, 2018: 379).

If a configuration is perceived as "frictionless," it implies the relatively natural character of whatever might be the syntactic boundary at the line break. Obviously, the syntax/meter alignment causes an over-generation of real cases of enjambment in automated processing. For example, Ruiz et al. (2017) refer to these cases as "expansions," claiming they form a particular class, but strangely enough take them as enjambments.

4.2.2 Tools for analysis

It is still rare for reader expectations to be included in the rhythmic analysis of MEs, as it is for them to be included in the analysis of other types of discourse as

- 3 Is distracted at times from its own clamoring
 4 By the sound of this plaint, wild and untamable.

(Baudelaire, Man and the sea, *The Flowers of Evil* Fresno, CA: Academy Library Guild, 1954)

The first module (lines 1–2) is the initial ME of the stanza, the second module (lines 3–4) the concluding ME of the stanza. We will see below how necessary this distinction is because MEs, depending on their function – initial or concluding – do not obey the same constraints of divergence and internal coherence.

Divergent ME at its beginning/at its end

The initial stanza module in Example 10, lines 1–2, is divergent at its end with the subject noun phrase “ton cœur” in “contre-rejet.” This is a typical indication of the loosening of metrical constraints in the mid-nineteenth century. A divergent ME at its end means that the reader expects a pause, while the ME ends with syntactic and semantic suspense. In Example 11, h1 of line 2 is divergent at its beginning, which fits the traditional notion of rejet:

- (11) Cette nymphe royale, & digne qu'on lui dresse

Des autels, tout ainsi qu'à Pallas la Déesse

(Ronsard, Églogue 1, Bergerie)

This royal nymph, & worthy of having

Altars erected to her, as well as to Pallas the Goddess

Consistent ME/inconsistent ME

A ME is said to be *consistent* if it matches a linguistic or discursive unit. In Example 10, the concluding stanza module, lines 3–4, is *consistent* because it matches a verbal phrase. In similar cases, Du Gardin speaks of a “petit sens et construction à part soy” (small meaning and construction per se; 1620: 77). An ME like “Des autels, tout ainsi” (Example 11) is considered *inconsistent* because no semantic projection of any sort can be associated with it.

Cornulier observes that an initial ME can be almost banally inconsistent (2009: 520). In addition, the lower the metrical level, the more commonplace the inconsistency. But concluding MEs tend to be consistent. This is why Ruiz et al. (2017) identify few enjambments at the end of lines 2 and 4 in quatrains. In relevant metrical terms, it means that stanza modules generally tend to correspond to large and coherent linguistic or discursive units – clauses, sentences, or periods.

Going back to Example 9, repeated in Example 12, we read:

- (12) {En beaux couplets et sur} {un rythme âpre et vainqueur}
 {In beautiful couplets and on a harsh and victorious rhythm,}

The initial ME is *divergent* at its end. The reader expects a natural speech pause, but the univocalic preposition “sur” implies a syntactic/semantic suspense. In addition, the ME does not match any linguistic unit, so it is said to be *inconsistent*.

As for the concluding ME, it is divergent neither at its beginning nor at its end, and it matches a linguistic unit, a noun phrase, which makes it a perfectly *consistent* ME. It should be kept in mind that, for any ME, including an initial ME followed by a conclusive ME, the final divergence of the initial ME does not imply an initial divergence or the inconsistency of the conclusive ME. This is demonstrated *a contrario* in Example 9, as modified in 13:

- (13) {En beaux couplets et **sur**} {un rythme âpre. Faiblesse! }
 {In handsome couplets in} {which harsh rhythm vaunted,}

(Verlaine, Prologue, Under Saturn Poems, <https://booksvooks.com/full-book/poems-under-saturn-pomes-saturniens-pomes-saturniens-pdf.html?page=11&part=4>)

h2 would not be perceived as divergent at its end, but the ME in itself, “un rythme âpre. Faiblesse!,” is likely to be inconsistent as there is no obvious way for it to match with any linguistic or discursive unit. Likewise, in the modified Example 10, repeated in Example 14:

- (14) 1- Tu te plais à plonger au sein de ton image ;
 2- Tu l’embrasses des yeux **et ton cœur tourmenté**
 3- Se distrait quelquefois de sa propre rumeur
 4- Au bruit de cette plainte indomptable et sauvage.

In terms of line boundary, we would get the same analysis: line 2, an NP subject is still separated by the line break from the VP that runs into the two following lines. For most analyses, Example 10 would include an enjambment. However, Examples 10 and 14 make a very different impression: h2 in line 2 is now perfectly consistent and could lead many readers to feel satisfied enough with it and thus to not perceive any discordance.

Therefore, an approach based on boundaries is largely insufficient because the line boundary may be perfectly marked (“end-stopped line”) but the ME may

be divergent at its beginning and/or inconsistent (Example 13). Or, the same line break can sometimes trigger a clear impression of discordance (Example 10) and sometimes be accepted as a temporary unit, as part of the natural flow of speech in a versified poem from the nineteenth century (Example 14).

Some trends can be observed. The lower the level of the MEs (hemistichs and lines), the more their consistency is morpho-syntactic. Higher-level concluding MEs (stanza modules and stanzas) are less divergent at their end and more often perceived as consistent, natural units. As larger units, their semantic and pragmatic coherence is more often discursive.

4.3 Construction of MEs

Because reading time and the reader's expectations are integral parts of meaning construction, these must be included in the analysis, as in Example 15:

- (15) 1- {Dix heures et demie,} {heure des longs services}
 2- {Divins.} Les cloches par} {milliers chantent dans l'air}
 (Verlaine, Londres, *Autres Vieux Coppées*)
- 1- Half past ten, the hour of long services
 2- Which are divine. Thousands of bells sing in the air.

If we hypothesize that these lines can be interpreted metrically as 6–6, in the course of processing $\hat{h}2$ the reader would expect a pause at the end of line 1. Despite the absence of any punctuation mark, the ME can effectively be processed as a complete and consistent unit, at least temporarily, and that is indeed the most likely initial perception. But the adjective “Divins” on the following line causes the *a posteriori* feeling that the line had been incomplete so far, inviting the reader to revise his or her previous interpretation and integrate this new information. This shows that reading time cannot be ignored since, in Example 15, the reader has no sense of discordance at the end of line 1, which is highlighted in Example 16, a modification of Example 15:

- (16) 1- {Dix heures et demie,} {heure des longs services,}
 2- {Les cloches par milliers} {chantent dans l'air divin.}
- 1- Ten thirty, hours of long services,
 2- The bells by thousands ring in the divine air.

Interpretation occurs at each metrical break, step-by-step, or during perching or pausing. And the ensuing ME processing can sometimes lead to a new interpretation. Table 1 presents very schematically these reading steps for the only concluding ME in a compound line (h2):

Tab. 1: Step-by-step reading of h2 in a compound line.

	Satisfied expectation		Frustrated expectation	
Reading h2: a certain degree of completeness is expected	h2 is perceived as temporarily complete enough and sounds natural.		The line is perceived as incomplete and sound-divergent at its end. There is the expectation of a development.	
	Confirmed	Contradicted	Confirmed	Contradicted
Reading the following line	The previous ME actually proves to be complete and natural (Example 17).	Retrospectively, the previous ME proves to be incomplete (Example 18).	The previous ME actually proves to be divergent (Example 19).	Retrospectively, the previous ME proves to be complete (Example 20).

- (17) {Mais lorsqu'on la néglige,} **{elle devient rebelle;}**
 {Et pour la rattraper} {le sens court après elle.}

But when it is neglected, it becomes rebellious;
 And to catch up with it, meaning runs after it.

- (18) {Dix heures et demie,} **{heure des longs services}**
 {Divins. Les cloches par} {milliers chantent dans l'air.}

- (19) {Murs blancs, toit rouge, c'est} **{l'Auberge fraîche au bord}**
 {Du grand chemin poudreux} {où le pied brûle et saigne,}

White walls, red roof, it is the cool Inn at the edge
 Of the great powdery path where the foot burns and bleeds,
 (L'Auberge, Verlaine)



- (20) Very rare cases. No example.

The traditional notion of enjambment proves to be of little use. It does not clearly identify which ME is being referred to, it does not clearly distinguish whether the property in question concerns the boundaries of an ME – and

which one? – or its internal consistency. It implies that any divergent ME invariably has the same effect of discordance, whether the ME is initial or concluding. Finally, it supposes that a similar feeling of cohesion (or lack thereof) between constituents can be applied from the sixteenth to the end of the nineteenth century. For these reasons, structural as well as historical, the notion of enjambment ultimately appears to be inadequate for studying the complex relationships between rhythm and meaning.

5 An example of step-by-step analysis

I have deliberately chosen this sonnet by Verlaine, which is a real metrical mess, due to the number of discordances that it displays at each metrical level that could not have occurred before the late nineteenth century and the late works of Verlaine:

- | | | |
|---|---|---------------|
| <p>1- Notre-Dame de Santa Fe de Bogota
 2- Qui vous apprêtez à faire le tour du monde,
 3- Or, mon émotion serait par trop profonde
 4- Dans le chagrin réel dont mon cœur éclata
 5- A la nouvelle de ce départ déplorable
 6- Si je n'avais l'orgueil de vous avoir, à ta-
 7- ble d'hôte, vue ainsi que tel ou tel rasta
 8- Et de vous devoir ce sonnet point admirable</p> |  | <p>octave</p> |
| <p>9- Hélas! assez, mais que voici de tout mon cœur
 10- Tel que je l'ai conçu dans un rêve vainqueur
 11- Dont, hélas! je reviens avec le bruit qui grise
 12- D'un tambourin, bruyant sans doute mais gentil
 13- D'être, grâce à votre talent de femme exquisite-
 14- Ment amusante, décoré d'un doigt subtil.</p> |  | <p>sestet</p> |

(Verlaine, À une dame qui partait pour la Colombie, *Dédicaces*)

(My grief would be too profound if I had not the pride to meet you at a table. . .)
 v8- And to owe you this sonnet not admirable
 v9- Alas! enough but here it is with all my heart

In the octave, the poet addresses his sonnet to a tambourine dancer leaving for Bogotá. In the sestet, he flatters her and treats her as a tavern aristocrat. In the octave, he evokes his supposed grief – essentially flattery – that would be too profound if “[. . .] je n’avais l’orgueil de vous avoir, à ta- / ble d’hôte, vue [. . .],” (I had not the pride to meet you at a table) “Et de vous devoir ce sonnet point admirable” (And to owe you this sonnet not admirable).

As the reader reaches the largest break between octave and sestet, he or she expects to find an end-stopped line. There is no syntactic obstacle to this expectation: the two quatrains form a coherent syntactic-semantic set. In addition, h2 in line 8, “sonnet point admirable,” seems to be consistent. Therefore, the reader is likely to retain the assessment that “the sonnet is not very good,” and nobody is expected to see it as possibly revisable since it occurs at the end of the octave.

Then, the sestet begins with “Hélas! assez, mais que voici de tout mon cœur” (Alas! enough but here it is with all my heart). Retrospectively, the reader understands his or her linguistic expectations were wrong, his or her must integrate this new information (the rejet “Hélas, assez”) and come up with a new interpretation. The consistency of h2 (line 8) now appears to be temporary. At the beginning of the sestet, what the reader thought to be a virtual noun phrase, “sonnet point admirable,” proves to be an incomplete noun phrase. They go from the assertion, “this sonnet is not admirable,” to the assertion, “this sonnet is not admirable enough,” with the inference, “This sonnet is not that bad, but nothing is beautiful enough for you.”

With a prosaic rhythm, no reader would ever have made the assertion “A sonnet is not admirable,” which is a well-known rhetoric technique: belittling oneself to better flatter the lady. In a way, the metrical release with this conspicuous rejet can mimic the laxity of customs. It can suggest the ways of “flashy foreigners” (in French *rastaquouères*): ostentatious displays of luxury and questionable practices in this disreputable world – in short: “dodgy” metrics for “dodgy” behaviors.

But in another way, it is worth noticing that speech consistency has not totally disappeared, since the reader is invited to come up with a first interpretation at the octave end, then a second, slightly different one at the beginning of the sestet. So, the traditional “turn in thought” at the boundary between octave and sestet is somehow still maintained.

These cases of reinterpretation can occur at any level and are not unusual, especially in the second half of the nineteenth century. Such an example illustrates that relationships between rhythm and meaning involve every linguistic level: phonology, morphology, syntactic, semantic, pragmatic, narrative, rhetoric, discourse, etc.

6 What can automated processing do?

1. Apply programs to homogeneous sub-corpora according to the level of MEs: hemistichs, lines, stanza modules, stanzas, composed stanzas, etc. by varying the line length, the period, and the distinction between prescribed forms and periodic forms. An interesting analysis would be to compare the internal structure of the quatrains in sequences of periodic forms and in sonnets.
2. The division between “end-stopped lines” and “run-on lines” is too sharp to be able to account for the variety of MEs from divergent to non-divergent.
3. Create homogeneous corpora and clearly distinguish MEs according to whether they are initial or concluding, so as to highlight regularities that cannot emerge without this distinction.
4. We need to bear in mind that a divergent ME at its end does not imply the next one will be divergent, or inconsistent, at its beginning.
5. Automated identification should be carried out in a step-by-step analysis: identification programs should be made to perform a two-step analysis: at the end of ME 1, then at the beginning of ME 2.
6. Finally, I suggest an interesting case study: coordination often produces rejets that are not perceived to be very divergent, as in this example, if read as 6–6:

Nos pieds glissaient d'un pur / et large mouvement
 (Verlaine, Beams, *Romances sans Paroles*,)
 (Our feet were sliding with a pure / and wide movement)

The rejet “et large” does not generate a particularly strong feeling of discordance; strangely enough, h2 “et large movement,” although slightly divergent at the beginning, appears sufficiently complete and consistent.

What are the conditions of this “sufficient consistency”? Under what conditions does this consistency cease to be perceived as sufficient? Examples of this type show that a rejet is sometimes perceived as being clearly discordant, sometimes less. Although challenging, the automated identification of divergent MEs is not impossible. In any case, this should not be conflated with the possible and varied effects of discordance they can create.

7 Conclusion

I have shown how analyzing the metrical expressions themselves is a better way to explain how versified poetry develops, where meaning is processed metrical expression by metrical expression in a dynamic of interpretative constructions and possible readjustments. The automated processing of relationships between rhythm and meaning faces some challenging obstacles, such as taking into account the temporality of processing, the reader's expectations, and the historical character of the appreciation of what a "natural expression" is. While more corpus-based research and observations need to be conducted in parallel, first applications seem possible if the constitution of homogeneous subcorpora obeys criteria such as the type and level of metrical expressions, different time periods, and individual poets.

References

- Auchlin A, Ferrari A. Structuration prosodique, syntaxe, discours: évidences et problèmes. *Cahiers de linguistique française* 1994; 15: 187–216.
- Auer P, Couper-Kuhlen E, Müller F. *Language in Time: The Rhythm and Tempo of Spoken Interaction*. Oxford: Oxford University Press, 1999.
- Chafe W. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago, London: The University of Chicago Press, 1994.
- Cornulier B. de Métrique de Mallarmé: analyse interne de l'alexandrin", dans *Analyse et validation dans l'étude des données textuelles*, éd. M. Borillo & J. Virbel, 197–222, éditions du CNRS, 1977.
- Cornulier B. de *Théorie du vers: Rimbaud, Verlaine, Mallarmé*, éditions du du Seuil, Col. *Travaux Linguistiques*, 1982.
- Cornulier B. *De Art Poétique: Notions et problèmes de métrique*. Lyon: Presses Universitaires de Lyon, 1995.
- Cornulier B. de Aspects du papillonnage métrique de La Fontaine dans *Les Amours de Psyché et Cupidon non sans un petit rappel de Maistre Clément*", dossier Agrégation, *Cahiers du Centre d'études métriques*, n°3, université de Nantes, 1997b: 73–87.
- Cornulier B. De La place de l'accent, ou l'accent à sa place: Position, longueur, concordance. In: Murat M, editor. *Le vers français: Histoire, théorie, esthétique*. Paris: Champion, 2000: 57–91.
- Cornulier B. de Problèmes d'analyse rythmique du non-métrique. *Semen: Revue de sémiolinguistique des textes et discours* 2003; 16: 107–118.
- Cornulier B. *De la métrique à l'interprétation: Essais sur Rimbaud*. Paris: Éditions Classiques Garnier, 2009.
- Degand L, Simon AC. Minimal Discourse Units: Can We Define Them, and Why Should We? SEM-05. In: Aurnague, M., Bras, M., Le Draoulec, A., & Vieu, L. (eds), *Proceedings of SEM-05*.

- Connectors, Discourse Framing and Discourse Structure: From Corpus-Based and Experimental Analyses to Discourse Theories, 2005: 65–74.
- Degand L, Fabricius-Hansen C, Ramm W. Linearization and Segmentation in Discourse: Introduction to the Special Issue “Discours [En ligne]” 2009: 4 (June 30, 2009). <https://doi.org/10.4000/discours.7292> (accessed April 20, 2022).
- Delente E. La dimension textuelle du rythme: Étude chez Verlaine. In: Monte M, Thonnerieux S, Wahl P, editors. *Stylistique & méthode: Quels paliers de pertinence textuelle?* Lyon: Presses universitaires de Lyon, 2018: 151–167.
- Dell F, Benini R. The Relationship Between Grammatical Structure and Metrical Structure in Jean Racine’s Verse: Metrics and Versification in Poetry and Song (conference paper). NordMetrik Conference, September 13–15, 2018, Stockholm, Sweden.
- Dell F, Benini R. La concordance chez Racine. *Rapports entre structure grammaticale et forme métrique dans le théâtre de Racine*, Paris, Classiques Garnier, coll. Versification, métrique et formes de la poésie, 2021.
- Dubeda T. Les unités accentuelles asyntaxmatiques en français et en tchèque. In: Bel B, Marlien I, editors. *Actes des XXVes Journées d’Etude sur la Parole. JEP 2004*. Fez, Morocco: Université de Provence, 2004. http://www.afcp-parole.org/doc/Archives_JEP/2004_XXVe_JEP_Fes/actes/jep.htm (accessed May 4, 2022).
- Du Gardin L. Les premières addresses du chemin de Parnasse, pour monstner la prosodie française par les menutez des vers français, minutees en cent reigles. Douay: B Bellere, 1620.
- Fabb N. *What is Poetry? Language and Memory in the Poems of the World*. Cambridge: Cambridge University Press, 2015.
- Garette R. *La phrase de Racine – Étude stylistique et stylométrique*. Toulouse: Presses universitaires du Mirail, 1995.
- Golomb H. *Enjambment in Poetry: Language and Verse in Interaction*. Tel Aviv: Porter Institute for Poetics and Semiotics, Tel Aviv University, 1979.
- Hussein H, Meyer-Sickendiek B, Baumann T. Automatic Detection of Enjambment in German Readout Poetry. *Proceedings of Speech Prosody 2018*: 329–333. <https://www.doi.org/10.21437/SpeechProsody.2018-67>.
- Kiparsky P, Youmans G, editors. *Phonetics and Phonology, vol. 1: Rhythm and Meter*. San Diego: Academic Press, 1989.
- Koops van ‘t Jagt R, Hoeks JCJ, Dorleijn G, Hendriks P. Look Before You Leap: How Enjambment Affects the Processing of Poetry. *Scientific Study of Literature*, 2014; 4(1): 3–24.
- Lacheret A, Victorri B. La période intonative comme unité d’analyse pour l’étude du français parlé: modélisation prosodique et enjeux linguistiques. *Verbum*, 2002; XXIV(1–2): 55–72.
- Martin P. Intonation de la phrase dans les langues romanes: l’exception du français. *Langue française* 2004; 141: 36–55.
- Martin P. Structure pramaosodique, structure de contrastes. *TRANEL: Travaux neuchâtelois de linguistique* 2007; 47: 103–116.
- Mourgues, M. *Traité de la poésie française: nouvelle édition revue, corrigée et augmentée avec plusieurs observations sur chaque espèce de poésie, 1724*; Slatkine reprints Genève, 1968.
- Paterson D. *The Poem: Lyric, Sign, Metre*. London: Faber & Faber, 2018.
- Peureux G. *La fabrique du vers*. Paris: Seuil, 2009.
- Prince A. Metrical Forms. In Kiparsky P, Youmans G, editors. *Phonetics and Phonology, vol. 1: Rhythm and Meter*. San Diego: Academic Press, 1989: 45–80.

- Rossari C. Identification d'unités discursives: les actes et les connecteurs. *Cahiers de linguistique française* 1996; 18: 157–177.
- Roulet E. La description de l'organisation du discours. Paris: Didier, 1999.
- Roulet E. Le problème de la définition des unités à la frontière entre le syntaxique et le textuel. *Verbum*, 2002; XXIV(1–2): 161–178.
- Roulet E, Fillietaz L, Grobet A. Un modèle et un instrument d'analyse de l'organisation du discours. Bern: Peter Lang, 2001.
- Roubaud J. La forme du sonnet français de Marot à Malherbe: Recherche de seconde Rhétorique. *Cahiers de poétique comparée*, 17–18–19. Paris: publications Langues'O, 1990.
- Ruiz Fabo P, Martínez Cantón C, Poibeau T, González-Blanco E. Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets. In: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Vancouver: Association for Computational Linguistics, 2017: 27–32.
- Simon AC, Degand L. L'analyse en unités discursives de base: pourquoi et comment? *Langue Française* 2011; 270(2): 45–59.
- Simon AC. L'analyse en unités discursives de base: pourquoi et comment? *Langue Française* 2011; 170 (2): 49–59.
- Tsur R. *Toward a Theory of Cognitive Poetics*, Sussex Academic Presse, 2008.
- Tsur R. The Performance of Enjambments: Perceived Effects and Experimental Manipulations. *Psychological Study of the Arts*, 2000. http://psyartjournal.com/article/show/tsur-the_performance_of_enjambments_perceived (accessed March 1, 2019).
- Tsur R. *Poetic Rhythm: Structure and Performance; An Empirical Study in Cognitive Poetics*. 2nd ed. Brighton: Sussex Academic Press, 2012.
- Tynianov I. *Le vers lui-même*, traduction coordonnée par Y. Mignot, UGE, 10/18, Paris, 1977.
- William J. *The Principles of Psychology*, 2 vol. New York: Dover Publications, 1950 [1890].

Natalie M. Houston

Rhyme Frequency in Nineteenth-Century English Poetry

Abstract: To account for the effects of rhyme, analysis of word frequencies in poetry should distinguish line-end words from words found elsewhere in the text. This research explores three methods of analyzing rhyme word frequencies that can contribute to the understanding of the conventions that shaped nineteenth-century English verse. Rhyme frequency ranking, effect size metrics, and the rhyme frequency ratio offer distinct views of a large poetry corpus and glimpses into how historical readers might have experienced rhyme's structuring force within poetic discourse.

1 Introduction

Distant reading, in Franco Moretti's oft-cited formulation, seeks to understand literature as a "collective system" rather than "a sum of individual cases" (Moretti, 2005: 4). Instead of performing close readings of a few canonical texts, Moretti argues that data-driven approaches to the study of large corpora offer new knowledge: "Distant reading: where distance, let me repeat it, is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text" (Moretti, 2000: 57). By examining small units of the literary text across large corpora, we can begin to understand the conventions and structures at work in the literary field more generally. Computational approaches to large corpora thus offer one approach to what Pierre Bourdieu identified as a challenge for literary historians: how to "reconstruct" those aspects of literature "which, because they were part of the self-evident givens of the situation, remain unremarked and are therefore unlikely to be mentioned in contemporary accounts, chronicles, or memoirs" (Bourdieu, 1993: 31). As valuable as book reviews, letters, and other documents of reception can be for understanding the significance of a particular literary text, many important aspects of the social reception of literature are unrecorded in such documents. Thus, to better understand

Acknowledgment: This study was made possible through my collaboration with the Literary Lab at Stanford University.

Natalie M. Houston, University of Massachusetts Lowell, e-mail: Natalie_Houston@uml.edu

how historical readers might have evaluated works of literature as conventional or unusual, we can apply distant reading methods to large corpora. This chapter presents three approaches to analyzing the frequency of rhyme words in a large corpus of nineteenth-century English poetry and discusses the contribution they make to understanding the conventions of rhymed verse.

1.1 Word frequency

Word frequency is one of the fundamental measures in textual analysis: it contributes to the semantic meaning of texts, to the distinguishing features of different genres of writing, and to the stylistic “fingerprint” of individual authors. In addition to the individual writer’s personal choice of words, many large-scale forces affect word frequency, such as historical changes in word usage and cultural conventions of genre and literary form (Underwood, 2016). Thus, even simple measures of lexical diversity can reveal important aspects of written texts when analyzed at a large scale.

Word frequency in natural language tends to follow certain distributional properties, conventionally described in Zipf’s law (Zipf, 1949). Zipf’s law suggests that natural language follows a power law distribution, in which a small number of words are extremely frequent in any corpus or text, and the greatest number of words occur a few times or only once. In fact, hapax legomena, words that occur only once, typically make up about half of the word types that occur in any corpus. Although these principles generally hold true at the very large scale, some analyses suggest that word frequency distributions may vary across time periods or genres (Liu et al., 2017).

Within these general linguistic parameters, the statistical analysis of word frequency used in literary texts has been shown to distinguish the works of individual authors. Stylometric analysis of the most frequent words used in particular documents has been widely used for authorship attribution (Burrows, 2002). Analysis of moderate- or low-frequency words and measures of lexical variation can be used to examine stylistic differences among particular authors (Hoover, 2008; Kim et al., 2020).

The study of literary genres and forms lies between the very general study of language usage measured by Zipf’s law and more focused studies of individual authorial style. Literary genres, modes, kinds, and forms are conventions with their own history and changing cultural value (Fowler, 1982; Guillory, 1993). Although poetry is made up of words arranged in sentences, the conventions of meter, line length, and rhyme sometimes require the alteration of natural syntax. Conventions of poetic discourse or register also constrain which words or kinds of

expression are considered acceptable in poetry at a given historical moment. Individual poets necessarily write within and against such conventions, following or modifying them to suit their expressive needs. These conventions also contribute to the “horizon of expectations” that readers bring to their evaluation of poetic texts (Jauss, 1982: 19).

Before the twentieth century, most lyric verse in English was written using line-end rhyme (McDonald, 2012). (Dramatic and narrative verse, in contrast, was frequently written in unrhymed iambic pentameter lines.) Because English is less inflected than many European languages and because of the long-standing preference for one-syllable rhymes in English, the number of possible rhyme pairs is fairly limited. In order to account for the effects of rhyme, analysis of word frequencies in poetry should include analysis of the frequency and distribution of line-end words as distinguished from words found elsewhere in the text.

1.2 Rhyme

In English poetry, rhyme is predominantly used in the final word of the poetic line. Two syllables that rhyme “have identical stressed vowels and subsequent phonemes but differ in initial consonant(s) if any are present” (Brogan, Cushman, 2012: 1184). Although the majority of rhymes commonly used in English poetry are masculine rhymes of only one syllable, feminine two-syllable rhymes may also be used occasionally. In the nineteenth century, many critics preferred exact rhymes (hat/cat) over near rhymes (hat/fact) or eye rhymes (love/prove), which are visible on the page but not heard in pronunciation.

Rhyme creates relationships of similarity and difference among the line-end words in a poem. In a poem with multiple words sharing a rhyme syllable, rhyme links not only the proximate pair of words but also a group or cluster. In a poem where lines 1, 3, and 5 rhyme, but lines 2, 4, and 6 do not, the presence of rhyme does not just connect lines 1, 3, and 5; it also highlights the difference, or creates tension, with the non-rhymed words. The non-rhyming end words in a rhymed poem are an integral element of the rhyme pattern.

Because of the general properties of words in English, even ordinary prose will eventually use words that rhyme. But for a poem to be considered an example of rhymed verse, including an occasional pair of rhyming words is not enough. Rhymed verse must contain a sufficient number of rhyming end-of-line words, typically arranged in a repetitive pattern. Human judgments about whether or not a poem is rhymed can easily be made within a 20-line window, and often the presence of rhyme may be perceived much sooner than that, as the dominant rhyme patterns in English verse include alternating rhyme (ABAB), enclosed rhyme

(ABBA), and ballad rhyme (ABCB). Even though these patterns may be combined to form stanzas of any length, the mere presence of rhyme can usually be assessed within a very few lines. For the analysis pursued here, all end words in each poem are considered in the analysis: in a ballad stanza rhymed ABCB, all four end words would be considered as part of the end word subcorpus, not just the B rhymed pair.

2 Methods

Given that word frequencies are widely recognized as useful indicators of a text's semantics and style, the present research explores three methods of analyzing rhyme word frequencies that can contribute to the understanding of the conventions that shaped rhymed English verse in the mid-nineteenth century. The first method calculates the most frequent end words in the corpus. Examining this list of frequent rhyme words reveals some of the common themes and rhetorical situations of lyric poetry.

The second method compares normalized word frequencies from two subcorpora: one consisting of all the end words and the other consisting of the remaining words in the poems. Keyness measures are frequently used in linguistic studies to discover which words are distinctive to a particular corpus. Although measures of statistical significance, such as p-values or G2 scores, are sometimes presented as measures of linguistic keyness, these scores are very sensitive to corpus size. Additionally, linguistic data do not necessarily follow normal statistical distribution. So, effect-size metrics are a more reliable measure of keyness (Gabrielatos, 2018).

The present study uses the %DIFF effect size metric (Gabrielatos, Marchi, 2011):

$$\%DIFF = \frac{(NFC1 - NFC2) * 100}{NFC2}$$

NFC1 is the normalized frequency of the word in the *target corpus*, in this case the end word subcorpus, and NFC2 is the normalized frequency of the word in the *reference corpus*, in this case the “body text” subcorpus containing all the words from other positions in the poem. In order to account for words that only appear in the target corpus (NFC1) and not in the reference corpus (NFC2), values of zero were replaced with 1e-19 (0.000000000000000001) to allow for the calculation of the effect size ratio (Gabrielatos, 2018: 237). A %DIFF value of 100 indicates that a word is twice as frequent in the target corpus as in the reference corpus; a %DIFF value of 200 would indicate that the word was three

times as frequent, and so forth. There is no upper bound to the value. Because of the substitution for zero values, extremely high %DIFF values flag words that only appear in the corpus as end words and never elsewhere in the poem.

The third approach calculates the rhyme frequency ratio, the ratio of the word's frequency as an end word to its frequency in the full text corpus (Houston, 2021). Examining rhyme frequencies as a percentage of corpus frequencies offers a statistical proxy for the likelihood a reader of the poems in the corpus would encounter a specific word as a rhyme. As noted above, in this study, all of the end words in the corpus are analyzed using the rhyme frequency ratio because each word's position at line end creates emphasis integral to the rhyme structure.

3 Materials

The present research was conducted using a corpus of rhymed English verse from the mid-nineteenth century. This corpus was derived from the Chadwyck-Healey English Poetry database, which aims to “encompass the complete published corpus by all poets listed” in the *New Cambridge Bibliography of English Literature (NCBEL)* (About, n.d.). Its 183,000 poems thus represent canonical English literary history on a large scale. In its nineteenth-century selections, the English Poetry collection goes beyond the *NCBEL* by including writers from the British Empire as well as a substantial number of American poets. Its size means not only that many of its poets are unfamiliar even to specialists in nineteenth-century literature but also that the poems included in the database range from well-known classics to texts that have been virtually ignored since the years of their first publication (Karlin, n.d.). Although there are thousands of lesser-known poets not included in the Chadwyck-Healey database, its large array of texts makes possible a distant reading of English poetry that would not be possible for conventional human reading alone.

To create the corpus used in this study, 32,233 rhymed poems were identified among the mid-nineteenth-century poems in the Chadwyck-Healey collection using a program that identifies rhymes by matching the rhyme words and syllables provided in John Walker's *A Rhyming Dictionary* (Walker, 1824; Houston, forthcoming). This historical poetics approach uses nineteenth-century rules for rhyme and pronunciation to identify rhymes in nineteenth-century texts.

For the present study, a poem over 100 lines long was defined as rhymed if over 88% of its end words were part of a rhyme pair or cluster. A poem under 100 lines was defined as rhymed if over 57% of its end words were part of a

rhyme pair or cluster. These percentages were manually confirmed as providing a threshold of reasonable certainty for rhymed verse in random samples taken from the corpus. It is possible that other rhymed mid-nineteenth-century poems exist in the larger corpus from which this study's subset was drawn, but in the absence of the human reading of every poem in the corpus, this threshold was deemed adequate for selecting the study's dataset.

This dataset contains poems by 240 poets, 196 male and 44 female. Although women poets make up 18% of the poets included in the corpus, poems by women (8,164) constitute 25% of the poems (32,233). The mean number of poems by each poet is 134, although the median is 70. There are 35 poets with more than 250 poems included in the dataset, and 67 poets with fewer than 30. The poems range from two lines to 160 lines in length, with a mean length of 30 lines. Notably, because of the popularity of the sonnet form, the corpus includes 6,551 14-line poems. In total, the corpus contains 981,712 lines of verse.

4 Results

4.1 Most frequent words

In previous work, examination of the most frequent rhyme words in a corpus of 1,244 poems suggested that rhyme word frequency follows the general patterns of natural language described in Zipf's law (Houston, 2021). In that smaller corpus, 55% of the rhyme words appeared only once. In this study, by examining word frequencies in a much larger corpus of rhymed verse, we can begin to explore how the presence of rhyme might affect overall word frequency.

The corpus used in the present study contains 31,165 word types that appear as end words. End words account for 984,336 tokens in a corpus of 6.8 million tokens (6,807,509), or about 14%. 7,117 word types, or almost 10% of the 74,404 types in the full corpus, only appear as end words.

Linguists estimate that in large corpora, 40–60% of words occur only once (Baayen, 2001). In both the full poetry corpus used in this study and the subcorpus of end words, the prevalence of hapax legomena is on the low end of that spectrum: 39% in the full corpus and 42% within the end word subcorpus. This suggests that poetic language may be more constrained in its vocabulary than ordinary speech or prose.

Another way of examining low-frequency words in the corpus is to calculate how many words only appear in one poem (42% in the full corpus and 45% in the end word subcorpus). This group of words extends beyond the hapax

legomena because end words may be repeated within a poem, particularly given the prevalence of verse forms with repeated refrains.

High-frequency rhyme words account for a significant proportion of end word tokens in the corpus: 44% (.4381) of the end word tokens in the entire corpus come from the top 250 most frequent word types, and 27% (.2663) from the top 100 most frequent word types. Because readers encounter high frequency words within individual texts, it is also useful to calculate the percentage of end words in each poem that are high frequency rhyme words. The mean percentage of end words within each poem from the top 500 most frequent rhyme words is 58% (.5818), and 44% (.4381) from the top 250 most frequent rhyme words.

Although many of the most frequent words in the end word subcorpus are also ranked as highly frequent words in the full text corpus, some of these are very common words in ordinary speech and prose as well as poetry. Examining the 30 words that appear in both the top 100 most frequent words of the end word subcorpus and the top 100 most frequent words of the full text corpus reveals pronouns like *it* and *him*, which carry little specific poetic content, but also nouns and adjectives that contribute to the themes and imagery expressed in poetry: *heart*, *love*, and *soul* (Tab. 1). Literary critics have suggested that when words appear as rhyme words, they receive particular attention from readers due to the structure of the poetic line and the sonic effect of rhyme (Brogan, Cushman, 2012).

Tab. 1: Thirty words that are ranked in the 100 most frequent words in the end word subcorpus and in the full corpus.

Rhyme word	Rhyme word rank	Full text rank	Rhyme word	Rhyme word rank	Full text rank
me	1	28	now	31	57
day	2	61	bright	33	94
thee	3	40	earth	34	87
away	4	96	by	41	24
be	5	26	life	42	71
light	6	68	soul	45	85
love	7	38	one	48	48
heart	8	43	it	56	22
night	9	90	him	63	67
there	11	52	you	68	46
more	13	62	god	78	81
see	15	83	come	80	84
eyes	17	97	us	83	73
all	23	17	will	84	53
still	26	56	o'er	97	70

Notably, the highest frequency end words in the corpus are used in almost every poem: 99% (.9882) of the 32,233 poems in the entire corpus contain at least one of the top 250 most frequent rhyme words, and 91% (.9109) contain at least one of the top 50 most frequent rhyme words (Tab. 2).

Tab. 2: Top 50 most frequent rhyme words in the corpus.

Word	Rhyme word rank	Raw frequency	Word	Rhyme word rank	Raw frequency
me	1	8038	still	26	2765
day	2	7916	alone	27	2630
thee	3	7198	land	28	2601
away	4	6005	dead	29	2594
be	5	5643	face	30	2576
light	6	5433	now	31	2552
love	7	5058	above	32	2525
heart	8	4919	bright	33	2494
night	9	4452	earth	34	2488
again	10	4290	die	35	2451
there	11	4003	high	36	2409
sea	12	3995	know	37	2398
more	13	3717	head	38	2389
sky	14	3621	pain	39	2388
see	15	3552	song	40	2367
air	16	3436	by	41	2345
eyes	17	3426	life	42	2339
way	18	3369	sun	43	2325
rest	19	3352	heaven	44	2320
eye	20	3288	soul	45	2306
free	21	3087	home	46	2227
fair	22	2957	hour	47	2220
all	23	2948	one	48	2219
breast	24	2833	go	48	2219
hand	25	2792	sight	50	2218

The 50 most frequent rhyme words in the corpus include many rhyme clusters (*me, thee, see, sea, free; light, night, bright, sight*) that, because of their frequent occurrence, helped shape the conventional sounds of English rhymed verse. Nouns predominate among the most frequent end words, including body parts, words pertaining to the natural landscape, and temporal markers. The very high frequency of *thee*, a deliberately archaic pronoun in the nineteenth century, in close proximity to *me*, the most frequent end word in the corpus, reflects not only the amatory preoccupations of much lyric verse but also the fact that poetic discourse has greater capacity for deliberate archaisms than does ordinary speech.

Thee was also often used in religious poetry to address God, following the wording of the King James Bible. Rhyme was a defining feature of nineteenth-century English poetry, and these high frequency rhyme words shaped its sounds and ideas.

4.2 Rhyme word keyness

To examine which words are distinctively used as end words, rather than appearing elsewhere in the poetic line, an effect size metric was used to measure the keyness of each word in the end word corpus. Keyness measures compare a word's frequency in two corpora in order to gauge how distinctive it is. As noted above, the %DIFF effect size metric was used to compare the end word subcorpus to the "body text" subcorpus containing all the words from other positions in the poems. A %DIFF value of 100 indicates that a word is twice as frequent in the target corpus as in the reference corpus; a %DIFF value of 200 would indicate that the word was three times as frequent, and so forth. Negative values indicate the degree to which the word lacks keyness.

In a large corpus, many words may have a high effect size but a low total frequency because they appear in a small number of poems. One optimal way to combine effect size and frequency ranking is to examine effect sizes between 900–1,500 among the top 2,500 end words. The words shown in Tab. 3 are 14 to 15 times more likely to be used as end words than elsewhere in the line but occur frequently enough in the corpus to constitute a recognizable part of nineteenth-century poetic discourse.

Tab. 3: Fifty of the 2,500 most frequent rhyme words with effect sizes between 1,300 and 1,500.

Word	Rhyme word rank	%DIFF effect size	Word	Rhyme word rank	%DIFF effect size
endure	580	1495.901	roam	374	1398.679
accord	1626	1495.484	approve	2175	1391.820
spray	464	1489.655	zight	2175	1391.820
riven	875	1480.053	supply	1857	1389.159
snare	1380	1477.557	skies	57	1388.711
clan	2228	1477.557	ray	178	1387.923
rhyme	626	1469.408	sublime	283	1384.042
restrain	2107	1468.983	fro	662	1381.404
departs	2394	1456.800	stare	1107	1378.960
tire	1367	1454.627	breast	24	1371.429

Tab. 3 (continued)

Word	Rhyme word rank	%DIFF effect size	Word	Rhyme word rank	%DIFF effect size
floor	439	1454.551	wiles	1644	1370.508
dole	2065	1452.908	mood	480	1366.851
weather	447	1451.640	delight	150	1366.309
hue	373	1449.516	guest	657	1359.876
prey	722	1445.010	nigh	233	1353.893
desire	362	1444.985	cell	638	1353.737
refrain	1571	1444.691	rill	763	1349.960
sky	14	1444.431	revealing	1267	1349.960
brink	845	1438.118	fatherland	1920	1327.961
succeed	2334	1438.118	oar	1389	1327.518
eves	2334	1438.118	bower	306	1319.801
core	1007	1427.523	holiday	1681	1319.801
play	102	1415.218	benign	1980	1315.576
mien	776	1413.209	shown	705	1312.661
bay	491	1408.174	untold	1118	1312.168

Given the predominance of iambic pentameter in English verse, it is unsurprising that most of the two-syllable words in this list are iambic and would be rhymed on the stressed second syllable: *endure*, *restrain*, *departs*, *succeed*, *delight*. Words that are especially marked by the poetic register include *eaves*, *bough*, *fro*, and *bower*. The dialect poems popular in the nineteenth century account for *zight*, a Dorset dialect spelling for *sight*. When these words appear in poems, they are 14–15 times more likely to appear as a rhyme word than to appear elsewhere in the line. Undoubtedly, most poets and readers would not have consciously recognized such patterns. But familiarity with English poetry might create an unconscious awareness of them as a structure of conventional poetic discourse.

4.3 Rhyme frequency ratio

A third approach to analyzing rhyme frequencies is the ratio of a word's rhyme frequency to its frequency in the entire corpus. This calculation produces values between 0.000058 and 1 in this corpus, representing the likelihood a reader of the poems in the corpus would encounter a specific word as a rhyme. The mean rhyme frequency ratio for the end word subcorpus is 47% (.4676).

Examining the 50 words with the greatest rhyme frequency ratios among the top 250 most frequent rhyme words in the corpus (Tab. 4) reveals some overlap with simple rhyme frequency: 21 of these words are in the top 100 most frequent

Tab. 4: Fifty of the top 250 most frequent rhyme words with the greatest rhyme frequency ratios.

Word	Rhyme word rank	Rhyme frequency ratio	Word	Rhyme word rank	Rhyme frequency ratio
shore	55	0.7901564	bed	144	0.675515
tomb	159	0.7834225	view	222	0.671829
ground	105	0.768664	away	4	0.6705
flow	117	0.7556584	way	18	0.664366
despair	195	0.7479853	die	35	0.662791
strife	107	0.7472209	glow	137	0.657335
birth	127	0.74597	air	16	0.65348
tone	210	0.7434716	hair	136	0.653176
divine	61	0.7303085	blow	171	0.649912
sky	14	0.7230431	wing	152	0.647655
play	102	0.7192029	part	98	0.646123
skies	57	0.715625	throne	170	0.634912
ray	178	0.7155172	done	100	0.624028
breast	24	0.7132427	flame	188	0.622619
delight	150	0.7125293	free	21	0.619258
nigh	233	0.7107843	fled	247	0.618689
below	72	0.6998138	wrong	227	0.618257
door	124	0.6987642	side	66	0.612974
pain	39	0.6980415	tree	64	0.607362
wall	243	0.6862745	gloom	134	0.607008
strain	217	0.6860119	again	10	0.600168
plain	190	0.683727	rest	19	0.599535
given	141	0.6826347	brow	58	0.596025
sight	50	0.6789103	tide	182	0.5902
flight	246	0.6767842	sea	12	0.589146

rhyme words, and 12 are in the top 50. But as a measure of how familiar it is to see a particular word as a rhyme word, the rhyme frequency ratio more closely approximates something about the reading experience, in which the conventions of a particular genre or form are gradually made apparent to readers through repeated exposure to its patterns. In mid- nineteenth-century poetry, *shore*, *tomb*, *ground*, *flow*, *despair*, *strife*, and *birth* appear as rhyme words 75% of the time. That means that as rhyme words they satisfy and reinforce poetic convention.

Mean rhyme frequency scores for each poem can be calculated from the rhyme frequency ratios for each rhyme word. Mean rhyme frequency ratios in the corpus range from a minimum value of .0257 to a maximum of .8372, with the dividing values for the first quartile at .3725, median at .4161, and third quartile at .4562. This wide range of values is to be expected in such a large corpus, which encompasses poems of very different styles. Calculating the

mean rhyme frequency ratio for each poem offers a method for comparing the conventionality of the rhymes used by particular poets or within other sub-groupings within the corpus.

Each of these three approaches to rhyme word frequency in a poetry corpus offers particular benefits, depending on the questions being pursued. Rhyme frequency ranking offers a simple view of common rhymes in the corpus. As indicated in Tab. 5, the %DIFF effect size and rhyme frequency ratio are highly correlated (.9518). When making comparisons between the end word subcorpus and the subcorpus of “body text,” the effect size metric is especially valuable because of the wide spread of values it produces, including negative numbers for words in the target corpus that are not distinctive when compared with the reference corpus. Because the rhyme frequency ratio compares rhyme to the full text corpus, or poetry in general, it may serve as a proxy for historical reader perceptions, and its calculation is quite simple. Researchers should consider these additional metrics as well as simple frequency measures because distinctive and high-ratio rhyme words do not necessarily fall within the corpus high-frequency ranges. No one metric is inherently superior to the others since each offers a different insight into repeated, distinctive, and familiar rhyme words.

Tab. 5: Comparison table of the three metrics: rhyme word frequency rank, %DIFF effect size, and rhyme frequency ratio.

Rhyme word rank	word	%DIFF	Rhyme frequency ratio	Rhyme word rank	word	%DIFF	Rhyme frequency ratio
1	me	128.51	0.2786	26	still	30.81	0.1811
2	day	601.04	0.5423	27	alone	589.35	0.5382
3	thee	182.04	0.3228	28	land	428.04	0.4716
4	away	1103.82	0.6705	29	dead	516.54	0.5103
5	be	38.16	0.1893	30	face	406.45	0.4612
6	light	307.26	0.4077	31	now	21.89	0.1708
7	love	61.43	0.2144	32	above	417.76	0.4667
8	heart	90.26	0.2433	33	bright	123.58	0.2743
9	night	457.40	0.4851	34	earth	98.93	0.2516
10	again	788.00	0.6002	35	die	1062.77	0.6628
11	there	82.80	0.2361	36	high	257.62	0.3768
12	sea	748.30	0.5891	37	know	191.42	0.3300
13	more	106.80	0.2590	38	head	693.09	0.5728
14	sky	1444.43	0.7230	39	pain	1267.57	0.6980
15	see	217.18	0.3490	40	song	402.79	0.4594
16	air	1015.63	0.6535	41	by	-50.09	0.0778

Tab. 5 (continued)

Rhyme word rank	word	%DIFF	Rhyme frequency ratio	Rhyme word rank	word	%DIFF	Rhyme frequency ratio
17	eyes	267.17	0.3830	42	life	32.54	0.1830
18	way	1071.00	0.6644	43	sun	231.75	0.3593
19	rest	785.66	0.5995	44	heaven	150.82	0.2977
20	eye	609.12	0.5452	45	soul	76.37	0.2297
21	free	862.18	0.6193	46	home	263.14	0.3804
22	fair	225.33	0.3548	47	hour	443.82	0.4790
23	all	-53.69	0.0726	48	one	-17.35	0.1226
24	breast	1371.43	0.7132	48	go	248.94	0.3710
25	hand	291.40	0.3982	50	sight	1150.84	0.6789

5 Discussion

Each of the approaches discussed here offers different views of rhyme word frequency in a large poetry corpus. Each could be used to locate groups of texts for further quantitative or qualitative interpretation that would combine distant analysis with close reading. Two poems are presented here as a brief example of how such metrics might open up the corpus texts for further study.

Why were you born when the snow was falling?
 You should have come to the cuckoo's calling,
 Or when grapes are green in the cluster,
 Or, at least, when lithe swallows muster
 For their far off flying
 From summer dying.

Why did you die when the lambs were cropping?
 You should have died at the apples' dropping,
 When the grasshopper comes to trouble,
 And the wheat-fields are sodden stubble,
 And all winds go sighing
 For sweet things dying.

(Rossetti, 1917: 30)

None of the rhyme words in Christina Rossetti's "A Dirge" are in the 500 most frequent end words in the corpus, and its mean rhyme frequency ratio is quite low, within the first quartile, at .2738. Feminine rhymes, consisting of an accented syllable followed by an unaccented one, are typically used in humorous

poetry or children's verse in English. Harmon even claims that "there are very few poems that employ nothing but multiple rhymes, and all of them are humorous in one way or another" (Harmon, 2012, 848). Rossetti's choice of feminine rhyme for this not at all humorous poem builds tension between the poem's form and its theme. Rossetti uses deliberately simple language to discuss the cycles of nature and the difficulty of accepting death. Rossetti also uses changes in line length and meter to emphasize the transitions in each stanza from the seasonal descriptions to the omnipresence of death. As simple as the language is, the poem is not conventional in its structure or its rhymes.

In contrast, all of the rhyme words in John William Inchbold's "Youth" are among the top 250 most frequent rhyme words, and the sonnet's mean rhyme frequency ratio is in the fourth quartile, at .4607.

In meadows bright with verdure of the Spring,
 Through which a stream pursued its lingering way,
 Changeful in hue as changed the passing day,
 A child plucked flowers and thus I heard him sing
 With voice as clear as sky-lark on the wing:
 "The jewelled year is all contained in May,
 When birds are happy and the world is gay,
 Then take whate'er the early seasons bring,
 And weave thy crown;" and as the child drew near,
 Years seemed to kiss his brow, yet left him bright, –
 And fresh flowers gathering, without a tear
 The others from his fingers fell, his sight
 Caught many more, nor was he scared with fear
 Though dark the winding river grew with night.

(Inchbold, 1876: 54)

The rhyme words alone do not make the poem, of course, nor do they determine its theme or tone. But this is a more conventional poem than Rossetti's in many respects beyond the frequency measures of its rhyme words in the larger corpus. It has the familiar rhyme pattern and meter of the sonnet form, and the conventional rhyme words used here align with the placid descriptive picture of spring. Where Rossetti's poem poses unanswerable questions about death, Inchbold's sonnet asserts the beauty of one season.

6 Conclusion

Rhyme frequency metrics might thus be used for exploratory data analysis of large poetic corpora or for large-scale studies of particular poets. Rhyme frequency ranking, effect size metrics, and the rhyme frequency ratio provide distinct views of a large poetry corpus and glimpses into how historical readers might have experienced rhyme's structuring force within poetic discourse. Readers might respond positively or negatively to conventional language and poetic style, depending on their aesthetic preferences. But their reading of poetry would inevitably have been marked by those conventions of poetic discourse. Today, access to large corpora and statistical measures, such as the ones discussed in this chapter, offer us the opportunity to investigate exactly what those conventions were: which poems were typical, and which were truly unique? Which words or phrases were unusual, and which were seen as trite because they were repeated so often? Such questions can only be examined from the perspective that distant reading provides.

References

- About English Poetry. Chadwyck-Healey Literature Collections, n.d. http://collections.chadwyck.co.uk/marketing/products/about_iloc.jsp?collection=e_poetry (accessed August 27, 2020).
- Baayen RH. *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers, 2001.
- Bourdieu P. *The Field of Cultural Production, or: The Economic World Reversed*. In: Johnson R, editor. *The Field of Cultural Production: Essays on Art and Literature*. New York: Columbia University Press, 1993: 29–73.
- Brogan TVF, Cushman S. Rhyme. In: Greene R, Cushman S, Cavanagh C, editors. *Princeton Encyclopedia of Poetry and Poetics*. Princeton: Princeton University Press, 2012: 1182–1192.
- Burrows JF. “Delta”: a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing* 2002; 17: 267–287.
- Fowler A. *Kinds of Literature: An Introduction to the Theory of Genres and Modes*. Cambridge, MA: Harvard University Press, 1982.
- Gabrielatos C. *Keyness Analysis: Nature, Metrics and Techniques*. In: Taylor C, Marchi A, editors. *Corpus Approaches to Discourse: A Critical Review*. New York: Routledge, 2018: 225–258.
- Gabrielatos C, Marchi A. *Keyness: Matching Metrics to Definitions* (conference paper). *Corpus Linguistics in the South* 1, University of Portsmouth, 5 November 2011. <http://eprints.lancs.ac.uk/51449> (accessed September 27, 2020).
- Guillory J. *Cultural Capital: The Problem of Literary Canon Formation*. Chicago: University of Chicago Press, 1993.
- Harmon W. *Masculine and Feminine*. In: Greene R, Cushman S, Cavanagh C, editors. *Princeton Encyclopedia of Poetry and Poetics*. Princeton: Princeton University Press, 2012: 848–849.

- Hoover D. Quantitative Analysis and Literary Studies. In: Schreibman S, Siemens R, editors. *A Companion to Digital Literary Studies*. Oxford: Blackwell, 2008: 517–533.
- Houston NM. Exploring the Idiom of Victorian Rhyme Through Applied Historical Poetics. In: Bories A, Purnelle G, Marchal H, editors. *Plotting Poetry: On Mechanically-Enhanced Reading*. Liège: Presses Universitaires de Liège, 2021: 41–55.
- Houston NM. Thinking with/in Forms. In: O’Sullivan J, editor. *Text Analytics for Literature: Tools & Methods from the Digital Humanities* (forthcoming).
- Inchbold JW. *Annus Amoris*. London: Henry S. King & Co., 1876.
- Jauss HR. *Towards an Aesthetic of Reception*. Minneapolis: University of Minnesota Press, 1982.
- Karlin D. Victorian Poetry and the English Poetry Full-Text Database: A Case Study, n.d. <http://collections.chadwyck.co.uk/marketing/products/karlin.jsp> (accessed August 27, 2020).
- Kim S, Tak J, Kwak EJ, Lim TY, Lee SH. Implications of Vocabulary Density for Poetry: Reading T.S. Eliot’s Poetry Through Computational Methods. *Digital Scholarship in the Humanities* (2020): <https://doi.org/10.1093/llc/fqaa009>.
- Liu C, Zhang S, Geng Y, Lai H, Wang H. Character Distributions of Classical Chinese Literary Texts: Zipf’s Law, Genres, and Epochs. *Proceedings of the 2017 International Conference on Digital Humanities 2017*. <https://dh2017.adho.org/program/abstracts/> (accessed March 2, 2020).
- McDonald P. *Sound Intentions: The Workings of Rhyme in Nineteenth-Century Poetry*. Oxford: Oxford University Press, 2012.
- Moretti F. Conjectures on World Literature. *New Left Review* 2000; 1: 54–68.
- Moretti F. *Graphs, Maps, Trees: Abstract Models for Literary History*. London, New York: Verso, 2005.
- Rossetti C. *The Poetical Works of Christina G. Rossetti*. Volume 1. Boston: Little, Brown & Co., 1917.
- Underwood T. The Longue Durée of Literary Prestige. *MLQ: Modern Language Quarterly* 2016; 77(3): 321–344.
- Walker J. *A Rhyming Dictionary; Answering, at the Same Time, the Purposes of Spelling and Pronouncing the English Language, on a Plan not Hitherto Attempted*. New Edition. London: William Baynes and Son; Edinburgh: HS Baynes and Co., 1824.
- Zipf GK. *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley Press, 1949.

Jonathan Armoza

Hayford's Duplicates: Cobbling a Model of Melville's *Moby-Dick*

Abstract: In “Unnecessary Duplicates,” renowned Melville editor Harrison Hayford discusses duplicate and vestigial settings, events, and characters in *Moby-Dick*, showing how they serve both wider aesthetic and structural effect in the novel. Their seemingly “unnecessary” nature prompts Hayford to offer several compositional hypotheses to explain their origins. With no extant manuscripts, Hayford relies on stylistic evidence, inductive reasoning, and Melville’s personal letters to posit three primary draft stages of the novel. As with Hayford’s duplicates, evidence of phenomena that produces data can be scant or absent. Multiple data sources may be patched together to paint a fuller picture of that phenomena. Model-based collaborative filtering methods can be used to characterize sparse data. Their outputs are recognizable in the recommendations of commerce and streaming websites. One such method, nonnegative matrix factorization (NMF), has proven successful in these tasks. Factorizing a table of numeric records approximates two factor matrices that potentially produced the table, thus estimating latent features not explicitly present in the data. NMF’s constraint of producing matrices with nonnegative values better isolates those features and produces more human-interpretable models. This chapter demonstrates how a probabilistic variant of NMF can be used to characterize the parts of speech in the sentences of *Moby-Dick*. Hayford’s ideas on Melville’s draft stages are explored and tested via the NMF model.

Acknowledgments: This writing and all of its figures, data, and code are available for viewing and download in the “Hayford’s Duplicates” GitHub project at: https://github.com/jarmoza/hayfords_duplicates.

I would like to thank Dr. David M. Blei (Columbia University, Statistics and Computer Science) for his consultation on nonnegative matrix factorization and the vector profiling method I developed to parse and analyze NMF models. I would also like to thank Drs. Jennifer J. Baker (New York University, English) and Thomas Augst (New York University, English) for their consultation on *Moby-Dick* and Harrison Hayford’s writings. Special thanks go to Matthew Honnibal, Ines Montani, and the developers of *spaCy*, as well as Marinka Zitnik and Blaž Zupan, the developers of *Nimfa*.

Jonathan Armoza, New York University, e-mail: Jia237@nyu.edu

1 Melville; or, The Grumbling Carpenter

I don't like this cobbling sort of business – I don't like it at all; it's undignified; it's not my place [. . .]. I like to take in hand none but clean, virgin, fair-and-square mathematical jobs, something that regularly begins at the beginning, and is at the middle when midway, and comes to an end at the conclusion; not a cobbler's job, that's at an end in the middle, and at the beginning at the end.

– The Pequod's carpenter, Chapter 126 “The Life Buoy” in *Moby-Dick; or, The Whale*

In a lecture-turned-essay entitled “Unnecessary Duplicates,” long-time Herman Melville editor and critic Harrison Hayford takes the grumbling artisan above as the author's stand-in aboard the Pequod in *Moby-Dick*. Hayford writes on the many strange instances of duplicate and sometimes vestigial settings, events, and characters in the book – this, one of the “greatest” of American novels.¹ Though these duplicates are sometimes lost to even experienced *Moby-Dick* readers in light of the novel's enormity and reputation, Hayford explains how these duplicates serve as both some of *Moby-Dick*'s most glaring imperfections and wonders. The seemingly “unnecessary” nature of many of these duplicates informs Hayford's own set of hypotheses as to how they might have come about. Three primary stages of drafting are central to Hayford's idea of the novel's composition. And as no extant manuscript can verify claims about Melville's drafts, Hayford relies on incidental evidence of language use and inference as well as the occasional comment Melville provided in personal letters discussing the novel's progress.

Authorship attribution via the computational study of writing style often proceeds with evidence gathered from similarly oblique angles of observation. This metaphor of cobbling vs. carpentry is an interesting one with which to experiment in that light. Can the computational analysis of Melville's writing in the novel be used to detect differences across whole drafts or to identify specific prose deployed for different effect and purpose? Below I attempt one exploration of this general question using Hayford's ideas on the drafting of *Moby-Dick* and a probabilistic variant of a modeling method called nonnegative matrix factorization (NMF) to help determine the most likely, consistent applications of Melville's phrasing. First, I present Hayford's hypotheses on draft stages, covering some of the broader strokes of those hypotheses and then fill in their details where appropriate.² Afterward, I show how the likelihood of part-of-speech patterning

¹ *Moby-Dick* is seen as partially responsible for the idea of the “Great American Novel,” something that bears mention while modeling some of its noted imperfections (see Buell, 2016).

² I list these as chronological to the plot as permitted. Hayford's explanation of the potential drafts and his evidence in the essay is at times (perhaps beneficially) circuitous.

in Melville's writing revealed via NMF can be used as a means to test Hayford's inferences about the drafting and roughly finalized structure of the novel.³

The stage 1 draft presents a plain narrative in two sections, one in which Ishmael potentially (1) sets out from New York to one whaling port, New Bedford, then to another port, Nantucket, and then (2) leaves for sea on a whaling ship. Absent from this narrative are the three central characters Queequeg, Ahab, and Bulkington, and in their place are Peleg⁴ and Bildad, with Peleg a newer captain of the ship (not yet necessarily dubbed the "Pequod") and Bildad the owner/former captain/pilot. Characters like Queequeg, Tashtego, and other "savages" may have been aboard, but they were likely neither central characters nor harpooners. We will return to why this is likely in a moment, but what is important to note about this first stage is that Hayford suggests that much of the sea narrative (Chapters 22–Epilogue) was written before some of the passages in the shore narrative (Chapters 1–21) that are now celebrated. With so little definitive evidence as to the shape of this stage of the draft, Hayford produces an outline using subtractive logic that cuts away from stages 2 and 3 of the drafting process. It should be noted that Hayford suggests that there may have been other stages and other smaller substages for which he does not account in this writing. He lists several questions he feels that the plot holes and duplicates of *Moby-Dick* elicit (a few of which follow below), including some that remain beyond the ambit of his essay.

Stage 2 introduces what Hayford calls "roles" for characters in the story involving the protagonist and the voyage (Hayford, 2003: 52). He characterizes the introduction and prompt exit of Bulkington in the third chapter at the Spouter Inn in New Bedford as an obvious graft onto the overall plot. Melville (or Ishmael) issues a mere four paragraphs, providing a tongue-in-cheek motivation for Bulkington's introduction: "the sea-gods had ordained that he should soon become my shipmate" (Melville, 2016: "Chapter 3. The Spouter-Inn"). From Bulkington's description, we get a sense of the potential roles that Melville had in mind for the character:

His face was deeply brown and burnt, making his white teeth dazzling by the contrast; while in the deep shadows of his eyes floated some reminiscences that did not seem to give him much joy. ("Chapter 3. The Spouter-Inn")

³ Like many works, a finalized edition of *Moby-Dick* is still up for scholarly discussion given its different published editions. See *The Herman Melville Electronic Library* for more on various editions (<https://melville.electroniclibrary.org> [accessed April 15, 2017]).

⁴ Hayford guesses that Melville originally dubbed him the clichéd "Pegleg" in an early, rough draft (2003: 54).

Hayford proposes that Bulkington was to take on two different roles in the novel, one of a seasoned whaleman who is teacher and “comrade” to Ishmael in the ensuing ocean voyage, and one of a “truth-seeker” harboring a deep resentment toward nature (i.e., the whale) (2003: 47).⁵ However, as readers know, Bulkington does not take on either of these roles in the final draft of *Moby-Dick*. The now famously vestigial Bulkington is given a short passage of literary grandiosity – the entirety of Chapter 23, “The Lee Shore” – and then summarily dismissed from the rest of the novel without explanation. Hayford ventures that it may have been that Melville could not bear to part with this good bit of writing, going back to the aforementioned introduction in Chapter 3 and later adding to Ishmael’s claim about Bulkington being his future shipmate the parenthetical “though but a sleeping-partner one, so far as this narrative is concerned” (Melville, 2016: “Chapter 3. The Spouter-Inn.”) to suture shut the cut.

Stage 3 splits the roles of Bulkington between Queequeg – a new or possibly re-used member of the whaling ship’s crew – as a seasoned mentor and comrade to Ishmael, and Ahab as a brand-new captain who takes on this role of troubled “truth-seeker.” Peleg and Bildad both become vestigial, duplicate owners/mates/pilots. Hayford delivers persuasive evidence for their duplicate status given their actions and dialogue at the signing of Ishmael onto the Pequod’s crew, as well as their presence during the piloting of the ship out of the harbor, and their calling of the crew aft. For instance, the latter act is entirely duplicated by Ahab later in Chapter 36, “The Quarter-Deck.” He, like several other characters throughout the novel (e.g., Fedallah and his men; Queequeg), seems to have been subsequently written in by Melville and has been “hiding out” in earlier chapters with only allusive, Bulkington-like grafts to indicate his presence aboard the ship (Hayford, 2003: 51). The naming of the ship as “The Pequod” after a defeated and scattered Algonquin tribe, the centralizing of Pacific Islander Queequeg, and the humanization and promotion of non-white characters like Tashtego and Daggoo to the role of harpooners are all deemed to be additions made during this draft. And while the chapters featuring Queequeg and his befriending of Ishmael in the shore narrative are rich, Queequeg’s comradeship *and* centrality almost entirely drop away once the ship is underway in Chapter 22. Why is this? Looking at a letter in which Melville alludes to the novel’s progress, Hayford deduces that it is because the rough sea narrative of stage 1 was completed long before stage 3, at which point Melville went back and inserted Queequeg’s fleshed out character and, subsequently, a spate of seemingly unnecessary duplicates. There are two

⁵ Hayford admits his own premise conflates the possible roles of Peleg and Bulkington, and thus also the possibility of other unmentioned drafting stages.

ports, two inns, two landlords, two pre-voyage sleeps (on the bench at the Spouter Inn and then in the bed with Queequeg), two to three captains (Peleg, Bildad, and Ahab), and two signing scenes (one for Ishmael while Queequeg is “hiding out” on his 24-hour “Ramadan” and one for Queequeg himself).⁶

In a painstaking examination of the prose and dialogue surrounding Queequeg and Ishmael in later chapters, Hayford also shows that there is no longer any strong comradery between the two – neither in scenes that evoke codependence, like when they work together weaving a sword-mat, nor in more fatalistic ones, like when a distressed Queequeg instructs the carpenter to build him his own coffin. In the former, Queequeg and Ishmael do not even regard each other, and in the latter Queequeg does not call upon Ishmael’s help while ceremoniously preparing himself for death at sea, but rather turns to nameless crew members to assist him. Ishmael is a mere onlooker. In the case of the famed monkey-rope scene where Ishmael and Queequeg are physically tied together, there is a similar lack of comradeship. Ishmael merely refers to Queequeg as a “savage” (Hayford, 2003: 57). Hayford notes that a few prefixed additions of words in these scenes, like “my” or “poor” before “Queequeg” are the only local evidence that brings us a sense that Ishmael regards Queequeg as anything other than an unusual crewman, similar to the language used to describe Tashtego (2003: 57). The virtual absence of this comradery also suggests a cutout of the now-invisible stage 2 drafting, where it is likely that Bulkington’s dialogue as harpooneer and “truth-seeker” was redistributed, the latter being reassigned to Ahab. The assignment of the title of harpooneer to Queequeg, Tashtego, and Daggoo also becomes a possibly haphazard late-stage addition. Hayford points out that they are mentioned as resting in the forecabin with the other men, away from their specified quarters in Chapter 33 “The Specksynder” – where the harpooneers of the Pequod are said to “take their meals in the captain’s cabin, and sleep in a place indirectly communicating with it” (Melville, 2016). This provides a bit of narratorial cover, the harpooneers being potentially sequestered away from the rest of the crew for chunks of the voyage. Nonetheless, the humorous, touching, and novel-lengthening prose of Queequeg’s addition to the shore narrative vs. his smaller role in the larger sea narrative warrants some attention. After noting all of these various instances of possible duplication and authorial cobbling, one is left wondering if there is a way to test Hayford’s hypotheses at a larger scale. Is there a thread that ties some, if not all, of these incidences together? For instance, is there a way to measure the difference between Queequeg as comrade vs. Queequeg as harpooneer?

⁶ Hayford notes more duplicates, but these are enough to mention here.

2 Probable cobbling

Seat thyself sultanically among the moons of Saturn, and take high abstracted man alone; and he seems a wonder, a grandeur, and a woe. But from the same point, take mankind in mass, and for the most part, they seem a mob of unnecessary duplicates, both contemporary and hereditary.

– Ishmael (?), Ch. 107, “The Carpenter” in *Moby-Dick; or, The Whale*

There is certainly more than one way of testing Hayford’s hypotheses, but what follows in the subsequent pages is an experiment as to how one might do so. One of the challenges of working with linguistic information is the sheer scale of possible patterning – semantics, phonetics, and syntax to list a few of the undergirding structures. *Moby-Dick* is over 210,000 words long with 17,000 to 20,000 unique words (depending on what you count as “unique”), all of which have been arranged in repetitive, systematic fashion to tell the tale of Ishmael, Queequeg, Ahab, the white whale – and much more. Melville is explicit in his reference to hermeneutics, devoting the whole of Chapter 99 “The Doubloon” to the subject. And similar to that coin-cum-symbol, the search for “whatever significance might lurk” in the book’s “strange figures and inscriptions” is unending (Melville, 2016: Chapter 99 “The Doubloon”). It is a complex book. To look at the kinds of Melvillian drafting tactics Hayford has proposed, we need a modeling method capable of teasing that complexity apart. But where to even start our investigation? Hayford’s premises are admittedly incomplete, though they do offer rough boundaries and points of contact on which to base a quantifiable study. He also notes which drafting scenarios are more likely than others. Hayford has supplied a bevy of evidential concepts, one of which is the functional roles that Melville might have had in mind for the characters of his novel. Concomitant with that concept is the notion that the novel can be roughly divided spatially in a way that reflects the chronologies of its development. Since there is no remaining manuscript with which to contrast the finalized version of the novel,⁷ I consider the split between a shore narrative and a sea narrative. In this arrangement, the latter drafting stage is represented more by the shore narrative and the earlier stage by the sea narrative. The versions of Queequeg that remain are also split across those two parts. This division can be complex given the at times uncertain identity of the novel’s narrator or, in other cases, the narration being very clearly a retrospective act (see, e.g., Melville, “Chapter 54. The Town-Ho’s Story”), but

⁷ As noted, *Moby-Dick* was originally published in two different versions, one British and one American. Subsequently, Hayford and his student Herschel Parker have contributed significantly to editing scholarly editions of the book. This experiment utilizes the digital edition available on Project Gutenberg. See the bibliography for more information on these editions.

Hayford's premise and evidence suggest that the shore vs. sea dichotomy suffices as the most identifiable representation of the novel's drafts. With this measuring concept in mind, the subsequent questions are: How do we operationalize it? What do we measure and how? One means would be to count words, maybe using something a bit more complex to account for document length and word frequency, like Burrow's Delta. But since Melville's own style is unlikely to change from draft to draft, some other aspect of language must be considered. From Hayford's perspective, it is clearly the type of words and how they are employed that allow for his reading of Queequeg's changing role. One possibility is to look at a more functional aspect of linguistic information, parts of speech (POS), to see what kinds of words are being employed to that effect. Hayford makes the case that distinct sequences of such particles reflect a mix of Melville's conscious and subconscious language patterning. As mentioned earlier, Hayford claims Ishmael and Queequeg's close friendship is oversold because of the combined effect of the later written/edited shore chapters and small insertions of affectionate phrases in the sea chapters. For instance, here is Hayford disputing the general claim that Chapter 72 "The Monkey-Rope" presents evidence of the "bosom friendship" between Ishmael and Queequeg:

[Queequeg's] special relationship to Ishmael is specified by epithets at two points: the first reference is "*my particular friend Queequeg, whose duty it was, as harpooneer . . .*"; the second is "*my dear comrade and twin-brother, thought I.*" [. . .]. In the light of what follows, I argue that Melville later inserted "*my particular friend*" and "*my dear comrade*" [. . .]. Nothing else in that scene of the chapter is written in a way that presumes or requires the pair to be comrades already [. . .]. [I]t suggests the likelihood that it was his writing of this scene and this metaphor that opened to Melville the possibility of making the pair bosom friends in the shore sequence when he removed Bulkington from the comrade role. (Hayford, 2003: 57–58, my emphasis)

What is notable about this particular, hypothetical edit is that it visits upon several central aspects of Hayford's ideas on the drafting. The first is that Melville went back during a later stage and inserted phrases that would seem to superficially bolster the relationship between the two characters. The second is that this physical tie-turned-metaphor – the "monkey-rope" – is also a possible point of origin (Hayford uses the term "likelihood") for Melville's rewritten shore narrative, where Queequeg becomes a richer character and takes on the larger role of Ishmael's comrade. Of note is that this hypothesizing all turns on a set of phrases. The two I have highlighted in the above excerpt have a similar construction:

< *possessive pronoun* > < *adjective(s)* > < *noun(s) or proper noun* >

One measurement option could be to look for such sequences of words with those parts of speech in the sentences of *Moby-Dick*. While it is true that their sequence informs their function, another way to think of this is merely to account for the proportional presence of each particle in a sentence. Though Hayford has presented this particular case, it is not certain that such a restrictive filter would not miss other possible categorically or functionally similar combinations of POS. At this low level of the text, the question following that measurement becomes, “Can the presence of particular amounts of POS in sentences reflect style?” The trouble with style detection that relies only on the measurement of counting is that, at the analytical stage, once those counts are made, the method for analysis relies on visual inspections and/or formulae with unidimensional results (e.g., the Burrows’ Delta that detects style distinctions via proportions of word frequency), often obscuring small-scale evidence based on that larger patterning across large swaths of text. In other words, once text-comprehensive frequency tallies and their respective distance metrics are derived, there is no easy way of going back to the relationships between words in the individual sentences where those words were deployed. This is why computational text analysis has turned to ever-more complex modeling techniques that can attempt to find the latent distributions of things like “topics” by producing multi-dimensional factors of difference. Probabilistic modeling’s suitability for this problem becomes apparent given the possibility that, as Hayford’s hypothesizing suggests, Melville’s pattern formation reflects a multi-stage process and in its final form appears as inconsistently distributed or unstable.

The first measurement for probabilistically modeling *Moby-Dick* in terms of Hayford’s ideas will be to count the POS in each sentence of the novel. I use the POS tagger of a language modeling program called *spaCy*.⁸ It should be noted that the act of POS tagging itself is always a probabilistic venture. Even manual tagging is prone to error or different interpretations. In the case of modern POS tagging programs like *spaCy*, a model of language use is created and “trained” via neural networks (e.g., machine learning). To a certain degree, the more data that flows into the networks the greater the possibility it has to learn which POS is most likely the correct one to assign words given their in-sentence context. The version of *spaCy* used below is trained on a corpus of over one gigabyte of digital

⁸ This use of *spaCy* accounts for proper nouns, punctuation, determinants, adjectives, nouns, adverbs, spaces, conjugations, verbs, participles, adpositions, numbers, pronouns, interjections, and symbolic characters. All remaining words unable to be tagged by *spaCy* are accounted for in a category, “X.”

English language texts.⁹ Once the tagging is done for the text, I represent each sentence by the counts of the number of instances of POS categories in them. The produced observations are thus a series of numbers stored in order of those categories.

3 Underlying factors

In linear algebra, a sequence of related numbers (often regarded as existing within a spatial coordinate system) is referred to as a vector. Each sentence can thus be considered a POS vector and can be said to exist within a space of the novel's sentences. The next task is to use some method of establishing meaningful relationships between those vectors. The suggestion itself is not so bold in a literary studies context. It is also inherently decompositional. Sentences are components of systems that are paragraphs that are chapters that are novels. What I would like to understand is if sentences and their POS serve particular roles in those systems. One patchwork of methods that has been used successfully to identify clusters of related vector data is known as collaborative filtering. These methods have been used in several well-known, contemporary contexts such as recommendation systems. Netflix, for instance, sponsored a collaborative filtering contest in 2006 that challenged researchers to exceed the company's algorithms' accuracy for predicting user ratings of movies. The general premise of the problem area is determining what attributes between Netflix's users and its movies draw one to the other and elicit a response. Some basic questions that they were interested in answering via models of their data included: (1) What about a user makes them watch a movie and/or give it a particular rating? and (2) What about a movie makes users watch and/or give those ratings? It is proposed that the answers to those questions may be characterized or approximated by recovering latent factors that helped produce movie ratings. Note that the factors are not the answers, but they are suggestive of them. This is the power and weakness of presenting a system in this way. The model encodes assumptions about a problem set and proceeds to approximate answers to questions one might have, given what is observable.

⁹ I want to take care to describe the instrument with which I am making measurements of the language of *Moby-Dick*. It is not a small point to be overlooked. For each measurement, there is a probability of error and one of success. At the time of my use, the creators of *spaCy* claimed it had a 93% tagging accuracy (Honnibal, 2016a).

One set of collaborative filtering methods helpful in locating distinct features within a data set is, in the very same sense, decompositional: matrix factorization. Here, it will be enough to briefly describe the mathematics of this process for its application to *Moby-Dick*. Each POS vector represents a linear equation (i.e., the equation of a line in n-dimensional space) with a set of known coefficients – the POS counts by category. A vector with three coefficients could be written as: $y = 2a + 6b + 3c$.¹⁰ You can get a sense of the space inhabited by all of the vectors you have on hand by stacking a set of vectors together into a table. This establishes a relationship between them. In our case, we have the sentences of *Moby-Dick* and the space they inhabit. In linear algebra, this stack of vectors is referred to as a matrix.¹¹ Going back to the Netflix example provides an initially more sensible context to discuss the factorization of a matrix. Here, the POS counts are replaced by ratings, each vector representing a Netflix user. In reality, there may be large holes (zeroes) in this matrix due to missing values. You can easily imagine that most people have not watched every movie on the streaming service. One task could be to infer what those missing ratings might be, based on the ratings of other users who have watched movies similar to those watched by a user of interest. Another task could be simply to group users to recommend a movie that has characteristics similar to the types of movies they have watched and highly rated. The latter is what matrix factorization is more typically used for.¹² In either case, the idea is to estimate how much latent factors contributed to the values of this matrix, and to do so the matrix is “factorized,” attributing some portion of each of its values to latent user factors and the remaining portion to latent movie factors. The process of factorization produces the two most likely smaller factor matrices; in this case, one for users and the other for movies. The coefficients of these factor matrices, when multiplied together, *approximately* produce the known data: the stack of vectors containing each user’s movie ratings. There will almost always be some amount of error or discrepancy between the factor matrices’ product and the original matrix. The process by which those factor matrices are determined is iterative, and the idea is to be able to reduce that amount of error as much as possible. Nonnegativity is also an important characteristic of and limit placed upon the entries in these factor matrices for nonnegative matrix factorization. Eliminating the possibility of negative values helps to isolate distinct topography among the numeric

10 Multiplying coefficients 2, 6, and 3 by any values for a, b, and c produces the location (or vector) in four-dimensional space: (a, b, c, y).

11 In formal notation, the variable names are erased for clarity.

12 Inferring missing values has been found to produce the unfortunate result of making the computation of these problems increasingly difficult.

features in the data and also removes the illogical possibility of negative factor values – and subsequently, negative feature values.

For the purposes of this experiment, I will use a variant of NMF called “probabilistic nonnegative matrix factorization” (PNMF)¹³ that treats the entries of my matrix as samples from a multinomial statistical distribution of data – that is to say, it assumes that each sentence can be associated with some categorical, relatively unique sentence type. Those sentence types are the latent factors in this model. PNMf also uses the Euclidean distance between the matrix produced by multiplying the two potential factor matrices and the matrix containing the original vectors (e.g., the POS counts in *Moby-Dick*) to help minimize that amount of error. Matrix factorization, in general, solves the problem of high-dimensional data. One implicit but important property of this decomposition is that, in order for the factor matrices to be multiplied, they must share a common dimension. And that shared dimension is something that can be parameterized by the person/algorithm doing the factorizing. For instance, if one matrix has a width of three entries, the other has to have a height of three vectors, and so forth. This mathematical requirement has the effect of allowing us to take a matrix of, say, 50 million users and 50,000 movies, and greatly reduce one dimension in each of the resultant factor matrices. The suggestion of this lower dimension value (it could be 1,000 or 2 or any integer greater than 0) is determined through iterative experimentation. While this reduces the computational complexity of determining the potential factor matrices, the lower the suggested shared dimension, the more detail is lost. The hope is that some details are not helpful or are even distracting to the pattern detection problem one is interested in solving.

The Netflix example helps us to understand this method, but users and items are only one metaphor for this kind of model. Another intuitive metaphor more appropriate to our POS-tagged data set is one proposed by biologists: patients and genes. In this modeling metaphor, one tries to identify latent factors that are “latent patients” and “latent genes.”¹⁴ The task is then to identify clusters of patients with similar genetic traits or, the inverse, genetic traits with similar patients. The different approach simply depends on what each of your vectors represent. Are they a patient with entries for each gene? Are they genes with entries that tally patients’ medical data? Overlaying this metaphor on the text to POS problem

13 See “Probabilistic Nonnegative Matrix Factorization” at <https://nimfa.biolaab.si/nimfa.methods.factorization.pmf.html> (accessed November 15, 2016; Zitnik and Zupan) for more detailed information on this particular variant method.

14 This is the metaphor employed by a set of researchers using the matrix factorization code library, *Nimfa*, used for the PNMf results below. See bibliography and <http://nimfa.biolaab.si/> (accessed November 15, 2016).

yields “latent sentences” or “latent POS.” In this case, the task is to identify latent sentences around which PNMf can group actual sentences from the novel as being somehow linked by their similar POS usage.¹⁵ Running the factorization algorithm several times to produce a consensus of plausible coefficients of those matrices helps to determine a hierarchy or tree of those consensus results – with the sentences at the leaf level. Dividing that tree’s branches up produces a cluster assignment for each sentence. These results give us much more than a mere cluster assignment ID though. In fact, they associate each sentence with one of those latent sentences. Each latent sentence is proportionally representative of the features of each of the sentences assigned to it. In this case, the features are POS counts. The latent sentences are thus also far more descriptive of the collection of objects as a whole than, say, a unidimensional distance between sentences or groups of sentences. I refer to these representations as PNMf-produced POS profiles. PNMf-POS profile vectors (e.g., the “latent sentences”) are averages of the POS vectors of those sentences grouped together in the PNMf model. Profile vectors enable understandings of a collection of objects that are simultaneously more comprehensive and individually representative.

4 Queequeg’s carpentered and cobbled roles

To turn back to the question of whether there is a way to determine the stylistic difference between Queequeg as comrade and Queequeg as harpooneer, this broader understanding of a phenomenon is exactly the hope. As discussed, Hayford’s understanding of the draft stages of *Moby-Dick* are both locally and globally informed. The years he spent editing, writing about, and teaching Melville and *Moby-Dick* have allowed him this more comprehensive perspective. But not everyone is Harrison Hayford. Here, the model is allowed to tune itself instead through suggested parameters (i.e., shared matrix dimension, statistical prior distribution) and through adjustment via error correction across multiple runs. For this experiment, the PNMf model and such accompanying suggestive and corrective modeling functions are supplied by a matrix factorization program designed

¹⁵ This process always begins with some prior assumption about the data. In this case, the prior statistical distribution given to seed PNMf is nothing but standard bell-curves (a.k.a. normal or Gaussian distributions). This prior information given to PNMf tells it roughly how it should expect possible factor matrix vectors to be statistically distributed. That statistical expectation is quickly overtaken as more and more plausible factor matrices and their products’ distance from the original data are computed.

by computational biologists called *Nimfa* – the authors' acronym for NMF. *Nimfa* identifies 393 semi-distinct POS sentence patterns in a model of the sentences of *Moby-Dick*. Just looking at explicit mentions of Queequeg in these sentences¹⁶ and framing them by Hayford's major divisions of drafts between shore and sea narrative, a plot of the instances of sentences identified by the POS sentence groupings produces the graph in Fig. 1.

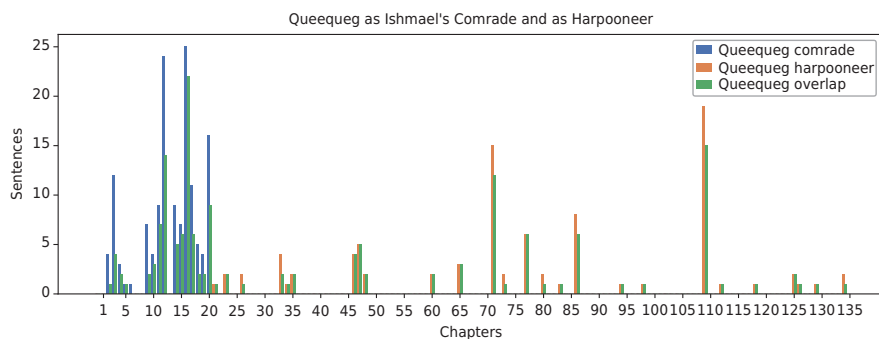


Fig. 1: Sentences mentioning Queequeg split across Hayford's division of *Moby-Dick's* shore (Chapters 1–21) and sea (Chapters 22–134) narratives. Here, blue represents Queequeg on shore as Ishmael's "comrade," and orange represents Queequeg at sea as the Pequod's "harpooneer." The green shows overlap in the POS pattern groups as identified by a PNMF model of the novel's sentences.

From the outset, this distant view in Fig. 1 makes clear the sheer sparsity and comparative dearth of Queequeg's presence in the novel after the Pequod leaves Nantucket. All sentences on the shore are marked in blue and those on the sea in orange. The green indicates when those sentences share a sentence group ID in the PNMF model. At times, that overlap is almost complete, but in other notable chapters there are significant amounts of uniquely grouped sentences. They include both much-critiqued and some lesser-critiqued Queequeg chapters including "Chapter 3. The Spouter-Inn," "Chapter 12. Biographical," "Chapter 16. The Ship," "Chapter 20. All Astir," "Chapter 72. The Monkey-Rope," and "Chapter 110. Queequeg in His Coffin." The question is how to meaningfully access the model's underlying data and when to take the step or carry out the operation that will provide a comparison with the inductive proposals of Hayford's hypotheses. At

¹⁶ The programmatic identification of characters is a notoriously challenging problem, and one with which even human readers can, at times, produce inconsistent results. See, for instance, Vala et al., 2015.

first glance, taking a look at the unique sentences across that shore-sea division seems an obvious step. If Queequeg is indeed written differently across that division, albeit with emendations once the bosom friendship of the shore narrative had been conceived in the stage 3 draft, then there should be some sort of stylistic or linguistic difference that conveys a richer characterization. The statistical leap here, though, is to question why one should move away from considering the mode or most common sentence types. Just like with the count data, it would seem that consistent styling would be more worthy of consideration than inconsistent styling. Moreover, humanities-style hypothesizing frequently likes to consider the functional/aesthetic value of outliers – something that contradicts the statistical notion employed here: central tendency. Looking at sentences in the most highly used POS patterns (i.e., the most highly identified PNMF groups) through a superlative lens is not very helpful in this problem context. Disregarding sentences with shared PNMF groups, however, is precisely the outcome of the measurement we would like here and can help determine what possibly makes these writings different. This is akin to throwing out that central bell of a bell curve.

However, one does not have to entirely disregard the most common POS patterns. They can also be looked at in a sort of statistical relief – what their profundity among all of *Moby-Dick's* sentences denotes. It turns out that comparing the sentences of Queequeg on shore with those of Queequeg on the sea reveals Melville writing in a more diverse manner in the former. On shore, writing mentioning Queequeg comprises 142 sentences that are assigned to 73 PNMF groups. While on the sea, writing mentioning Queequeg consists of just 91 sentences across 47 PNMF groups. When we subtract the overlapping PNMF groups, shore-Queequeg has 39 unique groups across 56 sentences and sea-Queequeg has only 13 unique groups across 15 sentences. Recall that the shore narrative includes 21 chapters and that the rest of the novel has another 114 plus an Epilogue. Where the plain sentence to chapter ratio denotes a more obvious Queequeg density issue, the PNMF groupings tell us of another. Combining those two measures of density – length and POS usage – we see that not only is Queequeg comparatively absent in mention, the way he is being written about once on the sea is also about 17% less diverse. So how can we talk about this diversity in a more meaningful way?

When looking at the uniquely grouped sentences of both sections, we find that they are pretty uniformly distributed. Each sentence was different enough in its POS usage to be assigned to a different group in the PNMF model.¹⁷ While

¹⁷ This is somewhat unsurprising given the statistical, tail-like nature of these outlier sentences, and while it could be suggested that their unique labelling is a product of overfitting, the iteration on parameterization and error reduction performed with *Nimfa* suggests otherwise.

each of those sentences' POS counts could be considered, the analytical aim here is to balance the operationalization of a particular part of Hayford's hypotheses with his original and broader theoretical conjecture regarding the novel's drafting. To stick with the PNMf model – the reasoning behind why sentences were identified as unique – I instead consider the PNMf-POS profile surrogates and attempt to determine the combined POS tendency of those “latent” sentences. Figures 2 and 3 below provide a look at the average of the POS profiles of those sentences identified as having POS patterns unique to the shore (Queequeg as “comrade”) and sea (Queequeg [mostly] as “harpooneer”) narratives.

Figures 2 and 3 give a sense of the dynamics of what, on average, distinguishes these sentences that were identified as unique from the other shared POS

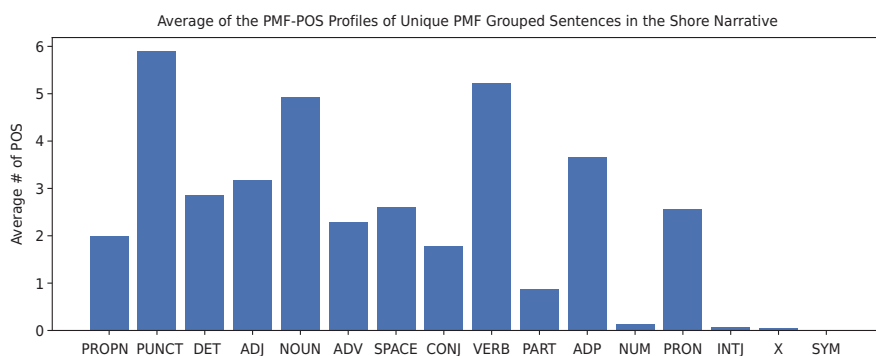


Fig. 2: PNMf-POS profiles of sentences with POS patterns unique to the shore narrative are averaged together to produce this POS sentence count dynamic.

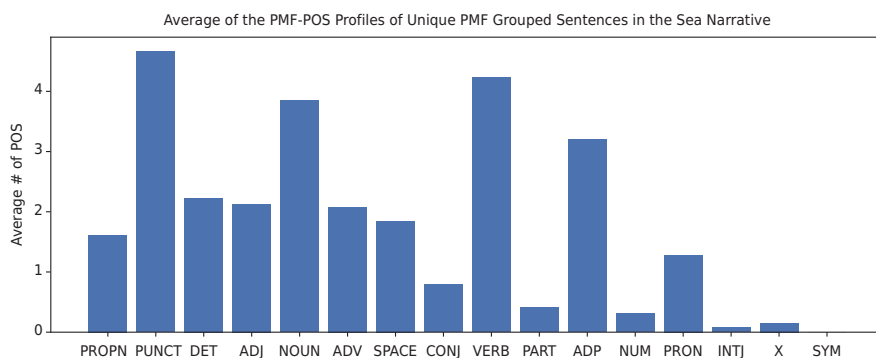


Fig. 3: PNMf-POS profiles of sentences with POS patterns unique to the sea narrative are averaged together to produce this POS sentence count dynamic.

patterns of the Queequeg sentences. The most immediate observation is a disparity in sentence length. In the shore narrative, these Queequeg sentences have a full nine to ten more words more than their sea narrative counterparts (38.136 average words to 28.872 average words). Once that difference is accounted for by normalizing the vectors,¹⁸ the percentage differences between the PNMF-POS profiles of these Queequeg sentences become perceptible. The hope is that the operation of taking their average does not obliterate too much of the outlying POS features that make these sentences unique. Instead of concentrating on the exact values, the idea at this juncture is to use the changes in POS across the separate parts of the novel to guide an investigation of the sentences.

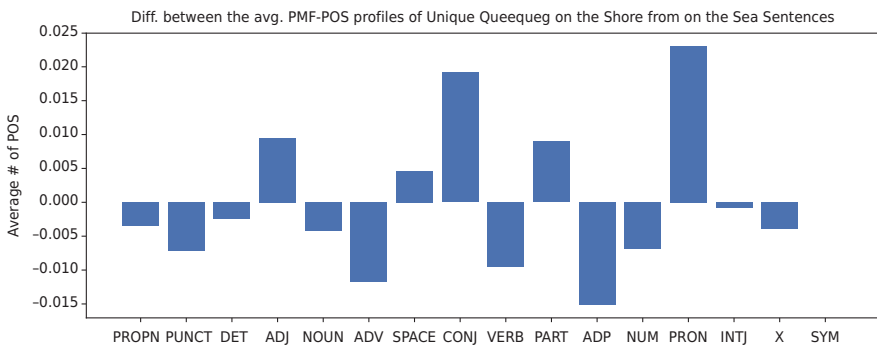


Fig. 4: The percentage difference in POS counts between Figs. 2 and 3. This displays the difference in Queequeg sentences across the stage 1 to stage 3 drafts. (When the order of subtraction is flipped, the percentage dynamics are reversed.)

In Fig. 4, the biggest swings are an increase in conjunctions, adpositions, and pronouns moving from the proposed earlier stage 1 sea narrative and the later stage 3 shore narrative. The larger sentence size of the shore sentences ostensibly explains the increased conjunctions and maybe even the adpositions. The largest shift in POS is in the increased number of pronouns. An optimistic and educated guess here is that the increased number of pronouns reference Queequeg, possibly “he” or “his,” and lead to a personalizing effect rather than making references to him using more objectified language as Hayford notes (i.e., “the savage”).

Which sentences from either narrative are most similar in POS proportion, not just certain POS category count differences? As if to answer for Hayford’s

¹⁸ The proportional normalization of vectors is accomplished by summing the POS category counts of each profile separately and then dividing each separate count by that sum.

claims about the kinds of characterization that Melville deployed across draft stages, the model replies. The most “comrade”- or shore-stage-like Queequeg is found in the fitting “Chapter 12. Biographical”: “In vain the captain threatened to throw him overboard; suspended a cutlass over his naked wrists; Queequeg was the son of a King, and Queequeg budged not” (Melville, 2016).

The story in this passage is retrospective, but it is a scene that depicts a Queequeg with some of the most narrative agency he has in the entire novel. Here, a young Queequeg flees his homeland for adventure, climbing aboard an American ship and defying the captain’s order to disembark. The increased pronoun usage appears as guessed alongside the sense of agency imbued in the character. One sentence later, the captain relents, observing, “*his* desperate dauntlessness, and *his* wild desire to visit Christendom” (Melville, 2016; my emphases). If one believes Hayford’s premises, it is not too difficult to imagine the entire chapter among several adjacent ones entirely focused on Queequeg – “Chapter 10. A Bosom Friend,” “Chapter 11. Nightgown,” and “Chapter 13. Wheelbarrow” – being constructed out of whole cloth in a late draft for similar ends. Conversely, when looking at what sentence in the sea narrative closely fits the difference in POS proportion, the model corroborates Hayford’s premise of Queequeg as secondary, patronized, as a less-than-equal crew member. Stubb speaks sarcastically, “[F]irewood? – lucifer matches? – tinder? – gunpowder? – what the devil is ginger, I say, that you offer this cup to our poor Queequeg here” (Melville, 2016).

Here, some of the limits of a POS tagger are exposed given the unclear punctuation and capitalization. *spaCy* is unsure of the bounds of the sentence. But as if to also temper skepticism about counting categories like POS – which somewhat erode the rich qualities of the words beneath them – the PNMf model shows that it at least correlates with Hayford’s suspicion about the imbalance in the qualities of Melville’s prose. In the last phrase, the possessive pronoun *our*, matched with *I* and *you*, sets an *us* off against Queequeg as other. Here, again, is that familiar sequence of *possessive pronoun (our)*, *adjective (poor)*, and *proper noun (Queequeg)*. It seems as if the model has connected enough dots to point to yet another example of Melville’s stage 3 drafts. Is this another insertion, a case of Melville’s “cobbling” as Hayford proposed? Or is this mere coincidence? The probabilistic weighting of NMF’s iterative, pattern-matching operations provides more *confidence* that it is not coincidental. And that degree of belief is itself the threshold between Hayford’s concept and our complex measurement.

5 Conclusion

I would take caution here, however, lest these results appear too conclusive. The model and the framing of the questions asked of it harbor the initial presumptions of Hayford's hypotheses about *Moby-Dick*: that duplicates exist because of a multi-stage drafting process and that Melville's phrasing can provide some evidence of those stages. This kind of warning is not exactly new when it comes to the precarity of literary evidence, but the introduction of a probabilistic model calls into question our authority as readers. These perhaps unnecessary duplicates that the PNMf model has located can still only be identified via an initial concept and set of operations meant to put that concept into practice. In order to fully test Hayford's premises, a number of plausible suggestions counter to Hayford's suggested division of three drafts across the shore and sea narratives of *Moby-Dick* would need to be made and then tested as to their comparative likelihood. Nevertheless, it seems that, in this experiment, the PNMf model of POS has uncovered some probable and promising evidence that Hayford was on to something with regard to the inequitably styled characterization of Queequeg in *Moby-Dick*.

References

- Bryant J, Kelley W, Ohge C. The Melville Electronic Library. 2008–2017; <https://melville.electroniclibrary.org/> (accessed April 15, 2017).
- Buell L. *The Dream of the Great American Novel*. Cambridge, MA: Harvard University Press, 2016.
- Burrows J. Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing* 2002; 17(3): 267–287.
- Halkidi M, Batistakis Y, Vazirgiannis M. On Clustering Validation Techniques. *Journal of Intelligent Information Systems* 2001; 17(2/3): 107–145.
- Hayford H. Unnecessary Duplicates. In: Parker H, editor. *Melville's Prisoners*. Evanston, IL: Northwestern University Press, 2003: 39–68.
- Honnibal M. Citation Information #272, February 22, 2016a. <https://github.com/explosion/spaCy/issues/272> (accessed November 15, 2016).
- Honnibal M, Montani I. spaCy, 2016b; <https://spacy.io/> (accessed November 15, 2016).
- Koren Y, Bell R, Volinsky C. Matrix Factorization Techniques for Recommender Systems. *Computer* 2009; 42(8): 30–37.
- Lee D, Seung HS. Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature* 1999; 401(6755): 788–791.
- Melville H. *Moby-Dick; or The Whale: The Project Gutenberg Ebook Edition*, 2016, edited by Lazarus D, Jonesey, Widger D, editors. <https://www.gutenberg.org/files/2701/2701-0.txt> (accessed April 15, 2017).

- Moretti F. Literature, Measured. Pamphlets of the Stanford Literary Lab 2016; 12: <https://litlab.stanford.edu/LiteraryLabPamphlet12.pdf> (accessed April 15, 2017).
- Moretti F. "Operationalizing": Or, the Function of Measurement in Modern Literary Theory. Pamphlets of the Stanford Literary Lab 2013; 6: <https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf> (accessed April 15, 2017).
- Vala H, Jurgens D, Piper A, Ruths D. Mr. Bennet, His Coachman, and the Archbishop Walk into a Bar but Only One of Them Gets Recognized: On the Difficulty of Detecting Characters in Literary Texts. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 769–774. <https://www.aclweb.org/anthology/D15-1088> (accessed April 15, 2017).
- Zitnik M, Zupan B. Nimfa: A Python Library for Nonnegative Matrix Factorization. Journal of Machine Learning Research 2012; 13: 849–853.
- Zitnik M, Zupan B. Nimfa. 2012; <http://nimfa.biolab.si/> (accessed November 15, 2016).

Georgy Vekshin, Egor Maximov, and Marina Lemesheva

Poeticisms and Common Poetic Discourse in the Digital *Russian Live Stylistic Dictionary*

Abstract: The use of a word in a specific sociocultural environment makes it a marker of that context and of the corresponding typical speech role. Is it possible to create an automatic detector of the poet's role in a text? The Russian poeticisms discussed in this chapter constitute a layer of vocabulary and phraseology that is optional for poetry but indispensable for authors who position themselves as poets and try to make their texts sound as poetry-like as possible. In Russian culture, this stratum is mainly used in common poetic discourse, the popular tradition of naive versification. The technology for poeticism detection implemented in the *Russian Live Stylistic Dictionary* and described in this chapter opens up possibilities for the essential stylistic differentiation of poems and the preliminary assessment of their aesthetic quality.

1 Introduction

The rapid advances made by corpus linguistics in recent years have allowed us to set ourselves the task of creating electronic dictionaries that automatically

Acknowledgments: We are grateful to all of those with whom we have had the pleasure to work on this project: Polina Morozova, one of the project's initiators, linguists Valentina Ledeyova and Elena Kukushkina, web designers Mikhail Gertzev and Nikolai Tzapkin, and others.

The work has been carried out within the framework of the St. Petersburg University research project, "Study of Vladimir V. Nabokov's Literary Heritage in an Interdisciplinary Perspective Using Information Technology Methods" (ID 72828386).

This project was supported by RFBR, grant no. 17-04-00421: Linguistic Development and Creation of the Electronic *Russian Live Stylistic Dictionary*.

Georgy Vekshin, Department of Russian Linguistics and Literary History, Faculty of Philology St. Petersburg University Saint Petersburg, Russian Federation/Moscow Polytechnic University, Moscow, Russian Federation, e-mail: philologos@yandex.ru

Egor Maximov, Department of Aerophysics and Space Research, Moscow Institute of Physics and Technology, Moscow, Russian Federation, e-mail: egor.maksimov@phystech.edu

Marina Lemesheva, Department of Russian Linguistics and Literary History, Moscow Polytechnic University, Moscow, Russian Federation, e-mail: lemesheva.m@list.ru

create a multidimensional stylistic portrait of a language unit, taking into account all the features of its sociocultural use. This solves the problem faced by the compilers of traditional dictionaries when it comes to describing the stylistic potential of words. Existing printed dictionaries now provide clues (stylistic marks), which, firstly, do not cover all types of stylistic coloring of the word; secondly, such clues are not based on an objective picture of the communicative practice of society but on the individual vision of the idea held by the compilers; finally, printed dictionaries do not have time to follow the real changes in the sociocultural and affective meaning of words, and are thus unable to quickly represent the social life of a word in its dynamics. This situation could be changed by a modern digital dictionary, representing stylistic variations of a word on the basis of its fixed applications in characteristic contexts, texts, and collocations.

The means to solving this problem from different angles stems from the tradition of the sociolinguistics of genres (M. Bakhtin, A. Wierzbicka); the study of register variation (M. Halliday, D. Biber); the tradition of semantic speech analysis within the framework of Prague functionalism and the Russian “theory of styles,” with its emphasis on the sociolinguistics of institutional spheres (K. Hausenblas, M. Kozhina); and the French tradition of stylistic semantics (C. Bally, P. Guiraud); as well as the corpus study of sociolects and ethnolects, and work on corpus research into tonality and topic modeling. At the same time, there is an extremely wide range of methods on offer to describe language sociolinguistically and semantically. We still have no universally accepted criteria for describing and defining the socially and affectively determined semantics of a word or other language units.

This chapter proposes taking an approach to the description of semantic structure and to the automatic identification of lexical units determined by one of the spheres of sociocultural interaction universal to European culture – the field of verbal art, the specificity of which is most clearly presented in the field of poetry, which in the mind of a naive speaker is equated with verse-composing practices. In accordance with the method described below, it is poetic works with their most obvious features of versification (accentual-syllabic meter, rhyme) that will be included in the corpus of the dictionary we propose.

The techniques developed for the automatic identification of poetically determined semantics and pragmatics of a word in the *Russian Live Stylistic Dictionary* (<http://livedict.syllabica.com>; hereinafter referred to as the “*Live Dictionary*”) project may prove useful for the prospect of using corpus methods to identify words and other linguistic units as deictic pointers to typical sociocultural contexts and as markers of communicative image, social status, and the cultural “self” of the author.

2 Theoretical background

2.1 Denotative core and stylistic periphery of meaning

It is well known that meaning as a linguistic phenomenon is the result of the use of a sign in speech contexts: “the meaning of a word is its use in the language” (Wittgenstein, 2009: 25^e). The speaker is guided by the memory of the sign, extracting the word from its repository as already marked by its typical use and then using it in a real, unique situation. The description of a word’s or idiom’s semantic features accepted in this work takes into account the fact that the meaning of a word is formed by the restrictions and preferences for its use within corresponding utterances in typical situations, including not only nearby pragmatic contexts but also typical contexts of institutional action and personal emotive condition. This description is based on a three-level model of the semantic structure of linguistic units, which distinguishes between layers 1) denotative-significative semantics (objective logical core), 2) stylistic coloring (semantic periphery of level 1 – socio-cultural and affective contextual-role semantics), and 3) the connotative semantic periphery of level 2, which following Apresyan (1995) is understood as associative semantics formed by nationally specific contexts.

The area of semantics that the *Live Dictionary* corpus and the dictionary itself is designed to reveal is an area of stylistic sociocultural and emotive coloring. The stylistic coloring of linguistic units (cf. Bally, 1921; Leech, 1974; Dolinin, 1987; Vekshin, 2017) is formed on the basis of the indexical ability of the linguistic sign under the influence of typical situations and roles of two kinds: typical socio-cultural frames and roles (both those universal to culture and more specific), and typical affective states and the corresponding emotive roles (“I am a scientist”; “I am a professional”; “I am a woman”; “I appreciate,” etc.).

Using templates for pragmatic word description (Wierzbicka, 1996; Goddard, 2019), information expressed by stylistic coloring can be described as follows using the example of poetry:

1. I know that the same thing can be said in different ways depending on the tasks of the speaker and the conditions of communication.
2. I say this as poets and people who write poetry usually say it.
3. I want you to believe that it is a poet speaking and that we are in a situation of poetic creativity.

For Russian socioculture, the following universal typical contexts (those that inevitably determine the life and behavior of any bearer of a given national culture) are considered the most influential: 1) the context of family relations (intimate communication and cognition) in contrast to distant, societally institutionalized

communication; 2) legal and official relations associated with the state (the exercise of state power is an objective pole of the social space); and 3) political and ideological relations (maintaining and redistributing power – the subjective pole of the space). These are accompanied by three contexts that provide cognitive activity and are formed by it: 4) science (the rational-logical mastery of nature), 5) religion, and 6) art (the objective and subjective “poles” of irrationally exploring the physical and metaphysical world) (cf. Shapir, 1990). The area of semantics that the *Live Dictionary* corpus and the dictionary itself is designed to reveal is its stylistic sociocultural and emotive coloring.

These spheres of the sociocultural space not only relate to the life of every bearer of modern, primarily European culture but together simultaneously form the communicative competence of any person. A person may give preference to some of these areas or specialize in some of them, but they cannot completely avoid activities in at least one of them. A person may not be a scientist, but they cannot but possess basic scientific concepts, for example, they cannot not understand what “temperature above zero” means; they may not be a believer, but they cannot completely isolate their mind from the category of “God.” This allows us to speak about the universal nature of these basic contexts. A person may not speak the language of a certain profession, not know the territorial dialect; they cannot be an aristocrat and a peasant, an adult and a child at the same time. However, in order to be a fully-fledged bearer of culture, any bearer of it must, to a greater or lesser extent, live everyday family life and obey the laws of the state, etc. In total there are six such universal spheres. They are grouped in a certain way and make up a system (see Fig. 1). While state and political activity constitute the objective and subjective poles within the single social space, in the same way, religious and aesthetic activity form two the poles of the mythological, extralogical knowledge of absolute and metaphysical reality. This is not the place to talk about the peculiarities of interaction or the axiological properties of these spheres and semiotic systems in different cultures, where their properties differ. It is enough for us to point out that six main institutional contexts determine the universal segmentation of the communicative space of European cultures, which is relatively independent of their more special communicative spheres and is superimposed on them.

In addition to these six basic contexts, the *Live Dictionary* is designed to identify typical, non-universal contexts (those defining social life and behavior but not necessarily encompassing the lives of each member). These are 1) social-estate contexts (bourgeois, peasant, aristocratic, lumpenproletarian, intellectual, etc.), 2) professional (programmers, school-teachers, etc.), 3) geographical (Russian South, St. Petersburg, etc.), 4) gender (contexts of male and female communication),

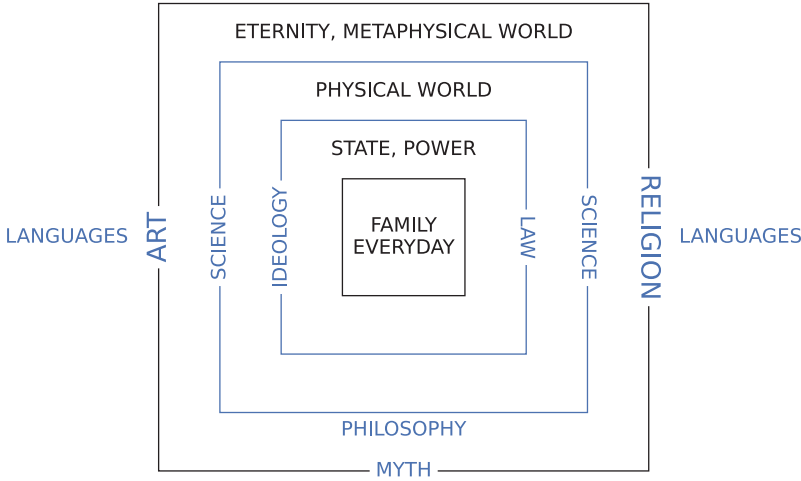


Fig. 1: The basic sociocultural frames and languages of culture.

5) age (contexts of the elderly, children's communication, etc.), 6) chronological (archaic, of the modern era, the latest contexts), 7) xenological (contexts putting in focus the opposition between cultural nativity and alterity – what is one's own, national, and what is alien, foreign), and 8) genre (typical situations in relation to typical goals and typical means and rituals of communication). For modern, at least Russian culture, these eight types of typical non-universal contexts can be considered the most significant, although this number is not finally determined. All of them are revealed as a relatively closed (chronology, gender) or open (profession, genre etc.) series of features.

Moreover, there are four main modal-evaluative contexts that form the affective component of stylistic coloring, which never denotes emotion but does express an emotional assessment as the speaker's point of view, distant from the essence of the subject, the conceptual content of the sign. We conceive of the emotion of pleasure/displeasure as the most universal, linguistically reflected affect forming the core meaning of desirability/undesirability (approval/disapproval) (1) within stylistic coloring (cf. Osgood, 1990; Russell, 1991, 2003; Wierzbicka, 1999). In the usual interaction with this opposition but also independently of it, three more types of affective meaning dimension in Russian are characteristic of stylistic coloring and are realized in the corpus tagging: (2) evaluation (importance/unimportance), (3) distance (intimacy, proximity/detachment), and (4) accommodation (friendliness/aggressiveness).

Thus, the layout of the *Live Dictionary's* corpora is shaped by 18 basic contextual role parameters – stylistic primitives of Russian speech.

2.2 *To be and to appear in culture and communication: The basic approach to the formation of a stylistically relevant corpus*

Humans are characterized by their desire not only to be but also to appear. These two basic sets of communicative behavior are not independent of each other, but they are polar in their extremes. Society is structured in such a way that appearances and even the imaginary, all kinds of relatively or purely formal indicators of the speaker's role, status, and function (not always real or sincerely fulfilled by them), are an integral attribute of communicative behavior in general. When "the actor identifies with the socially objectivated typifications of conduct in actu, but re-establishes distance from them as he reflects about his conduct afterwards," "the roles, objectified linguistically, are an essential ingredient of the objectively available world of any society" (Berger, Luckmann, 1966: 91). Born as socially objectified scenarios of behavior in typified situations of communication, they are further performed as tools of social self-presentation, relatively free from the pragmatics of the activities and contexts that gave rise to them. This property of the role of behavior is emphasized in the definition of the Jungian concept of the persona as "a complicated system of relations between individual consciousness and society, fittingly enough a kind of mask, designed on the one hand to make a definite impression upon others, and, on the other, to conceal the true nature of the individual" (Jung, 1966: 264).

The peripheral social meaning in the semantic system of a language like Russian is primarily formed by roles as images, more than by roles as functions and identities. Fundamentally, a text that upholds the conventions of academic writing is not the same thing as a scientific one, just as a text that rhymes and is saturated by poeticisms does not yet provide an aesthetic effect:

If the specific properties of scientific speech were entirely derived from scientific needs proper, then, obviously, the articles and books of the most outstanding scientists would be the most typical examples of the academic writing style. Meanwhile, there is more likely an inverse relationship: real scientists are often inclined to violate the unwritten norms of the academic writing style, while works that are weak in terms of content are most often written quite academically. (Dolinin, 1987: 75)

The success of the act of scientific communication is achieved by means of the language of science – as an operational semiotic system of devices, tactics, and strategies – the rigor of logical constructions, and the explanatory power of conclusions. The pragmatics of science may demand a concentration of terminology in the text if it is needed to build a conceptual and categorical system of knowledge. However, young scholars often do not notice that they are only

inventing a complex term to emphasize their innovativeness, or that they are complicating their syntax too much to simply manifest the depth and complexity of their thought. This paradox can be very clearly observed in Russian scholarly practice. (The authors of this chapter, writing in English, are not always able to successfully fight the stylistic inertia of Russian scientific communication – for example, the excessive use of passive constructions to “objectify” knowledge, sadly noticing that they speak more complexly than pragmatics and simple common sense require.) Of course, this does not negate the possibility or even the appropriateness of using stylistically colored units to achieve the aims of scientific knowledge, but stylistically marked academic elements as such (as well as the exclusion of neutral expressions) are required especially where the writer diligently signals scientific discourse or even feigns it. Thus, there is a tendency toward asymmetry between the scientific quality of thought and the academicity of writing. A similar trend can be observed in literary discourse.

So, the stylistic meaning of linguistic units cannot be deduced directly from the essential pragmatics of institutional communicative acts. Moreover, it is impossible to build and tag a stylistically adequate corpus by relying on the formal classification of texts alone according to their qualification by the author or the bibliographer (this is how most corpora with genre-stylistic tagging are arranged). The basis of the stylistically relevant corpus should be formed by the texts and text fragments that most consistently and clearly manifest a typical social role and actively signal the relevance of verbal action in institutional contexts. To identify the sociocultural significance of linguistic units, not all texts that are nominally related, for example, to science or poetry, must be included in the corpora.

The fact that the sociocultural pragmatics of the text and the nature of the stylistic coloring of its elements do not depend directly on each other can be observed, for example, in advertising texts. Thus, the task of promoting cosmetics, which has nothing to do with the pragmatics of scientific knowledge, is often performed using academic phraseology and syntax (Diez-Arroyo, 2013). Such a text works as a persuasive advertisement because the image of the speaker is built as the image of a scientist, a professional. Instead of expressions such as “for fine and supple skin,” terms and verbal nouns will appear in a text: “Sharp temperature changes, environmental pollution and stress make our skin lose its optimum level of moisturization” (Glacier Essence, Sensilis, leaflet; Diez-Arroyo, 2013: 202); “[t]his eye serum contains marine kelp that’s meant to lift skin while retinol stimulates collagen to plump the area” (Murad advertisement). A person buying this product does not need to know what retinol or collagen is, but they should have a feeling that it was a professional who advised them to use it. Such advertising

tactics may even be explicit: “Discover a dermatologist’s way to reveal fresh, new, healthy skin” (L’Oreal advertisement).

If the text is saturated with units of a scientific coloring, sustained in a single academic writing style, and corresponding to an overall image of the author as a scholar, even though it might never be formally attributed to science, it can enter the scientific corpus of the Stylistic Dictionary. Conversely, popular scientific texts, which are as accessible as possible, explaining the nature of things in a trusting, friendly tone, will not be included in the scientific corpus of the *Live Dictionary*, since the role-playing, image side of these texts correlates with the image of a close friend, not a scientist. If any text or fragment of text, regardless of its institutional pragmatics, is kept in a single informal register and embodies the image of a loved one with the help of colloquial markers, it will be included in the conversational corpus. Such a dictionary corpus, for example, in relation to its other corpuses, will automatically detect colloquial markers that tend to be used in everyday contexts.

2.3 Poetry, poeticity, and poetic corpus: The semantic structure of Russian poeticism

The language of literary art as a semiotic operational system, as a technique of “estranging” the cognition of verbal and extra-linguistic reality in its metaphysical perspective, with its unique techniques and tactics (Shklovsky, 1990; Hansen-Löve, 1978, etc.), is not accessible to every native speaker. However, everyone has some idea of what poetry is, of how it differs from other types of speech. The national literary language as a public domain includes elements that native speakers, regardless of their ability to understand art, associate with artistry, poetry as a cultural institution, and with the way they think a typical poet should speak. For example, a composer of amateur congratulatory poems will be guided by this norm, including accentual syllabic meter, rhyme – albeit flawed – and a certain kind of vocabulary and phraseology; these are popular markers of poetic diction. We label these markers poeticisms. Their presence in the text does not necessarily mean that we are dealing with verbal art, although it does not automatically mean that we are dealing with a sample of amateur writing. Modern Russian poetry may use poeticisms as well as other linguistic means within the frame of its artistic tactics, which may include the task of portraying the typical speech role of the author as a poet (along with any other possible roles). However, poetry and the poetic on the one hand and poeticity and the poetical as signals of the poet’s speech role on the other are radically different concepts.

Poeticisms, being part of mass cultural consciousness and containing poetical coloring as a component of their stylistic semantics, are partially taken into account and described as such by traditional dictionaries. Meanwhile, the accuracy and adequacy of their presentation in conventional dictionaries entirely depend on the mindset of their compilers, which is not only subjective but also quickly becomes obsolete. Words marked “poetic” usually include those implying the meaning of “high,” “solemn.” However, genuine Russian poeticisms as exponents of a poet’s speech role go far beyond these stylistic classes (Vekshin, Lemesheva, 2019).

The *Live Stylistic Dictionary*, aimed in particular at the objective automatic recognition of poeticisms, uses a poetic subcorpus that, owing to the style set technique (see 3.2), primarily includes texts that manifest the speech role “I am a poet, a composer of verse” and that are recognized as poetry due to their poeticality. In this regard, the poetic corpus of the *Live Dictionary* covers the widest range of poetic texts – from high poetry to graphomania – but, first and foremost, those that the majority of Russian speakers will qualify as typical verses and lyrics expressed by way of a poet’s typical speech, for poetry outside the verse and lyric genre is practically inexistent for naive native speakers.

The poetic subcorpus of the *Live Dictionary* is part of the corpus of fiction, characterized by a combination of tags such as fiction, verse, and lyrics. One necessary and sufficient sign of a poeticism is its poetic social coloring. Such, in particular, are the lexical and phraseological markers of poetry: *бесбрежный* (shoreless), *безмолвие* (quietude), *брожу* (I roam), *былое* (yore), *взор* (gaze), *вериги* (chains), etc. The main semantic component of poeticism is actualized, for example, in the following constructions: *У нее не взгляд, а взор; Мы говорили не о прошлом, а о былом; День был не удивительный, а дивный*. Something similar to these expressions can be represented in English sentences: *It was not a holiday, but a feast. He was not crying, but weeping*.

3 Related work

3.1 General approaches

In works on automatic genre identification (cf. in particular Stamatos et al., 2000; Santini, 2007; Sharov, 2018, etc.), a fully automated approach based on the n-grams method has been proposed, which was designed to capture nuances of style, including lexical variation. However, grammatical and formal indicators (verb, substantivity, share of functional words, average word length, sentences,

etc.) are considered the main ones, while lexical indicators (high-frequency words) are treated as non-universal and are used only as an addition to the main set of parameters since they are considered subject-dependent (Ljashevskaja, Sharov, 2009). Contextual role-based sociocultural parameters of speech are obviously in some coordination with the topic, but they act as an independent and powerful factor in style formation. When combined with stylometric data, they can be very useful in the attribution of texts and the determination of authorship and individual style.

In contrast to the approaches adopted in automatic genre identification, the *Live Dictionary* fundamentally distinguishes between speech genres as culturally patterned and rigidly pragmatically determined; textual practices (complexes of typical textual means in typical situations to achieve typical goals) (cf. Bakhtin, 2011; Wierzbicka, 1985; Günthner, Knoblauch, 1995; Vekshin, 2017; and others) on the one hand and the complexes of universal sociocultural role markers (in Russian and Czech traditions often called “functional styles”) on the other. Thus, identifying the text as belonging to the genre of the church sermon (an important reference point here is the formal name of the genre of the text and its typical pragmatics), which makes it possible to assign the genre tag “sermon,” does not interfere with the text being simultaneously assigned to the religious corpus, if the speech image of the preacher is primarily constructed as “I am a believer,” or to the spoken corpus, if the dominant speech role in the text is “I am a person close to you.” Genres and style are phenomena of a different order, which is why the genre and style markup of texts for the *Live Dictionary* corpus are carried out independently of one another.

The sociocultural style, with its exceptional contextual role determinant, and the speech genre are phenomena not only of a different hierarchical order but also of a different nature. This is not usually taken into account in works on register analysis (Halliday, Hasan, 1985; Martin, 1993; Biber, 1993) and is also reflected in corpora classification and tagging systems. In the *Russian National Corpus* (RNC; <http://www.ruscorpora.ru>), prose is included in the main body, and poetry is presented as a separate one, along with dialect (the subcorpus of territorial varieties of the language) and newspapers (the collection of texts of any genre limited to a specific print source). The RNC poetic subcorpus is made up exclusively of high-quality, professional poetry, striving to overcome the canon, often intentionally creating stylistic contrasts, combining elements of different sociocultural styles to implement artistic tasks. This corpus may be a source of data on the frequency of words used in Russian poetry at different times, on the keywords of certain authors, but we can only partially judge the stylistic semantics of a word to the extent that these texts embody a poet’s typical, stable speech role (despite the fact that professional poetry normally does not use such make-believe tactics).

We hope that, to understand the sociocultural use of the word as a whole, the *Live Dictionary* corpus, compiled based on the role context factor and using the styleset method, which will be described below, is much more indicative.

The list of the 50 most significant Russian poeticisms obtained on the basis of the *Live Dictionary* is a series in which we do not find a single random element and which includes, in addition to frequency, poetic concepts and formal operators, pure carriers of poetic sociocultural coloring: *ты* (you), *словно* (as if, like [poet.]), *мне* (me), *сердце* (heart), *над* (over, above), *солнце* (the sun), *небо* (sky, heaven), *снег* (snow), *всё* (all, everything), *свет* (light, world), *как* (as, like), *лишь* (only [poet.]), *осень* (autumn), *где* (where), *ночь* (night), *любви* (love [dat., gen., abl. loc.]), *жизнь* (life), *иль* (or [poet.]), *чтоб* (so, for [poet.]), *душа* (soul), *вдруг* (suddenly), *вновь* (again [poet.]), *ветер* (wind), *сквозь* (through), *тобой* (you [abl. instr.]), *будто* (as if, like [poet.]), *дождь* (rain), *ни* (nor), *боль* (pain), *любовь* (love), *душе* (soul [dat., abl. loc.]), *глаза* (eyes, eye [gen.]), *снова* (again), *миг* (moment, blink, about time), *тебя* (you [gen.]), *как будто* (as though), *твой* (your), *не* (not), *птицы* (birds), *счастье* (happiness), *мной* (me [abl.]), *моей* (my [abl., fem.]), *ночи* (night [gen., abl. loc.]), *души* (soul [gen.]), *он* (he), *дом* (home), *ль* (if, whether [poet.]), *мой* (my [masc.]), *моя* (my [fem.]), *лес* (forest) (see comparative data on frequent lexemes in Russian naive poetry and lexemes dominant in the RNC poetic subcorpus in Bonch-Osmolovskaya, Orekhov, 2013).

3.2 Method of corpora formation

The theoretical apparatus described here is the basis of the methodology for the formation and labeling of the *Live Dictionary* corpus. Eighteen types of Russian elementary stylistic meanings, which reflect the corresponding types of sociocultural contexts and emotional states, require the building of 18 dictionary corpora. For texts reflecting non-universal contexts, subcorpora (for example, professional or genre) are collected. Each of them should include at least 1,000 texts. Crucial for fulfilling the main tasks of the *Live Dictionary* are six universal, basic sociocultural contexts (conversational, administrative, ideological, academic, literary/poetic, and religious). We have assembled these cases most deliberately. Since the markup of any text includes 18 tag types, other corpora will also be formed in the process of compiling the six main corpora; however, the deliberate choice of texts for a particular corpus remains most effective since the principle of maximum stylistic uniformity of the text is being observed here.

To build the corpora, experts are using the “styleset” principle (Avamilova, Vekshin et al., 2019). The main feature of this method is that it excludes certain selections of texts according to their formal classification and explicit attribution

and requires only those texts that most typically exhibit the typical speech role of the speaker. That is why, for example, not all articles published in scientific journals can be selected for the scientific corpus. Only those papers and their fragments that actively use the style of a word to create the typical speech role of a scientist will get into the corpus. And in this case, articles by novice scientists who are very concerned about their speech role and seek to demonstrate their scholarship will make their way into the *Live Dictionary* over texts by major researchers. Thus, the compilers of the *Live Dictionary* are guided in principle by texts where the author seeks “to appear” much more than “to be.” These texts turn out to be the most saturated with stylistically specific vocabulary and phraseology. And the frequent appearance of any word in such contexts will ultimately be a guide for the stylistic identification of a unit as a result of machine learning.

The second feature of the styleset method is the expert’s work algorithm, which involves the initial formation of search queries consisting of five to seven words or expressions exclusively specific to this context and speech role. The expert’s next main partner is then the web search engine, returning texts from which the expert selects those that are stylistically most homogeneous, with the most clearly expressed desire on the part of the author to play a corresponding speech role. The expert, firstly, selects these texts for the corpus (sometimes not the whole texts, because we require the most stylistically typical fragments). Secondly, in these texts, he or she looks for the most striking markers of contexts and roles, then uses them for new queries.

To give an idea of this process, we will try to use the English poetic styleset we have chosen intuitively: *misty purple wane glory light restless*. The algorithm of action will be as follows: after sending the request, poetic texts are returned. These are, in particular:

- The Complete Poems of Emily Brontë (https://en.wikisource.org/wiki/The_Complete_Poems_of_Emily_Brontë)
- Songs of the Sea Children / Bliss Carman [electronic text] (<https://quod.lib.umich.edu/a/amverse/BAC8020.0001.001?view=toc>)
- Forest Buds: From the Woods of Maine, Elizabeth Akers Allen (<https://quod.lib.umich.edu/m/moa/ABK0842.0001.001?rgn=main;view=fulltext>);
- Victorian Women Writers Project: The Dream, and Other Poems, Caroline Sheridan Norton, 1808–1877. (<http://purl.dlib.indiana.edu/iudl/vwwp/VAB7052>);
- Songs – Song – Wedgeblade.net (collection of lyrics); and others.

Further, in Bliss Carman’s cycle “Songs of the Sea Children,” chosen because of its general stylistic poeticity (regardless of its pragmatics – ironic or serious), we find the most poetical words and phrases: *joyous soul, golden April, fare-*

thee-well, twilight on hills, without thee, rose of dawn, hollow jar, and others. They will fall into the styleset base for the formation of new stylesets and will also be used to expand the further search for texts. Moreover, the most stylistically specific poetic texts will be selected for the poetic corpus. Please note that, as stated above, Russian poeticisms are undoubtedly more active in modern speech as indicators of the role of the poet than in English, and the status of a poeticism in modern English speech differs greatly from its status in Russian – it is more of an exotic element than a fact of modern literary language and mass versification practices. Therefore, in the case of a similar search on the Russian internet, we will receive a large number of today's amateur poems in which the author seeks to sincerely implement his speech role as a poet.

Stylesets include predominantly poetically colored units as well as thematic conceptual words, and, finally, words and phrases that are simply frequent in poetry. To make the styleset base more complete and objective, an expert could resort to the data of lexicography, which widely uses the mark “poetic,” as well as “high” (Kourova, 2016). However, it should be noted that the stylistic marks of dictionaries often suffer from inaccuracy and much more subjectivity than the intuition of a modern native speaker, and are also archaic and usually do not take into account many new trends in the use of words (Vekshin, Shilikhina, 2017). Therefore, we draw from these sources with great care.

In addition, to add typical poetic concepts and characters to the database, dictionaries of poetic language may serve as a support (*Dictionary of the Language of Russian Poetry*, 2001–; Ivanova, 2004; Pavlovich, 2007) as well as the most significant linguistic studies of poetry and authors' individual style.

The most stylistically homogeneous texts or text fragments selected from the search results are further subjected to double processing. Firstly, units are extracted from them to form new stylesets and further replenish the corpus. In the immediate context (within the limits of one poem), for example, other poetical words and phrases are supposed to appear, obeying the rule of stylistic attraction. An expert can verify the correctness of the “linguistic flair” by using an additional web search, which allows us to understand whether a given word or combination of words is mainly unique to poetic texts or is also regularly found in utterances of other pragmatics, texts of non-poetic genres (for example, religious). When solving this problem, the poetic subcorpus of the *National Corps of the Russian Language* is also of great help.

Secondly, the selected texts are tagged by experts in accordance with the established parameters, which are divided into two blocks: 1) factual information about the text and 2) its stylistic features. Factual information includes attribution: name, source, author (name, gender), and date. These data can serve, in particular, as guidelines for automatically reconstructing the picture of the dynamics of

the use of a language unit. In addition to the main sociocultural features, the stylistic tagging of texts requires us to determine their narrower social and genre specificity: gender, age, profession, estate, areal, xenological, chronological (in relation not to the actual historical period but to the one recreated in the text), and, finally, genre proper. Xenological stylistic coloring is a specific semantic parameter of the Russian language unit, which is used as a deictic indication of its belonging to a foreign cultural environment (first of all, European, due to which the word also forms a modality of importance), when its foreign cultural origin is tangible to native speakers. These include, for example, Gallicisms and the latest Anglicisms in Russian, and Church Slavisms in archaic vocabulary. Semantics of the latter type, combined with the coloring of poeticism, strengthens it in this status and generates the meaning of “high.”

Thus, a combined stylistic portrait of a linguistic unit, which can be compiled with the help of a dictionary, will reveal not only poeticisms in general but also those that are characteristic, for example, of female poetry or a folkish poetic style.

3.3 Style identification

The word classification problem can be considered a word representation learning procedure. The majority of modern word representation algorithms are based on neural networks (Joulin et al., 2016; Mikolov et al., 2013; Devlin et al., 2019; Peters et al., 2018). These algorithms are able to learn word representation using the context. However, the aforementioned methods are unsupervised. That means we would not be able to obtain the necessary style features of the word from its representation. Due to this disadvantage, we cannot use these methods in our dictionary. At the same time, every single word may be considered as a text that contains only one word. This allows us to use a text classification model to predict word labeling.

The majority of modern approaches to the text classification problem are based on recurrent or convolutional neural networks (Zhang et al., 2015; Liu et al., 2016; Conneau et al., 2016; Howard, Ruder, 2018; Lai et al., 2015). This means that the model takes into account not only a single word but also the word order. For this reason, the model cannot be used in our case.

Another group of algorithms uses topic modeling as a preprocessing step for text classification (Neogi et al., 2020; Li et al., 2018; Pavlinek, Podgorelec, 2017). In these approaches, a text is represented as a vector in a low-dimensional feature space. Some topic modeling algorithms, like LDA (Blei et al., 2003) or PLSA (Hofmann, 2013) are based on the co-occurrence of words in texts. These approaches

are implemented in different libraries for NLP (Vorontsov et al., 2015; Egorov et al., 2019; Loper, Bird, 2002; Rehůřek, Sojka et al., 2011). However, topic models are also unsupervised, so we cannot control the result of the representation.

To solve our problem, we require supervised text classification approaches that take into account the word presence but not the word order. The classical approaches based on the bag-of-words model (Zhang et al., 2010; Harris, 1954) have these properties. The more advanced approach uses TF-IDF (Spärck, Jones, 1972).

4 Proposed corpus

The first important task in creating such a complicated computer system as the *Live Stylistic Dictionary* is to collect data on which the machine learning model can later be trained. To solve this problem a new corpus needs to be created. This corpus must contain texts of different styles and genres, written by different authors in different periods. One can download an up-to-date version of the *Live Dictionary* corpus from our official website (<https://livedict.syllabica.com>).

4.1 Overview

Each text should be properly labeled according to the following features: title; source; date of writing; type of source (e.g., internet, newspaper, etc.); gender of the author; typical social and pragmatic context affiliation (style); social stratum; age; occupational, regional, gender, xenological, and chronological specificity; speech forms (dialogue/monologue; verse/prose; phrase/text); genre characteristics.

The corpus was created and labeled by our group of linguists – experts in stylistics. It currently contains more than 8,000 texts. A labeling process is carried out on a website. An expert pastes in the text and then fills out a simple form where the correct category for each feature must be selected. For some features, such as occupational specificity, a single text may belong to multiple categories. Such texts, for example, may be written by representatives of several professions. The “style” feature is considered essential to the *Live Dictionary*. This feature has been labeled more accurately by the experts, and all the algorithms will be tested on it first. According to this feature, the text may be classified by six categories: colloquial, business, ideological, scientific, religious, fiction (Fig. 1). The latter includes a poetic subcorpus, which is formed as a combination of texts with the following features: fiction + verse + lyrics. Other features also contain different subcategories, but their labeling is a work in

progress. The number of texts in the corpus so far has been modest. However, the corpus is constantly growing, and more ideological and fictional texts, as well as texts with different genre tagging, will soon be added. Therefore, all the advantages will be shown below on the example of text “style” feature.

4.2 Structure

All the data is stored in the SQLite database file. The data is stored in the “question” table. Each text is described with 23 features from “field0” to “field22.” Each categorical feature has its own description in the appropriate table.

To speed up computation, each text is stored in its own separate file. The filename is stored in the “question” table in the “field6” column. In the event

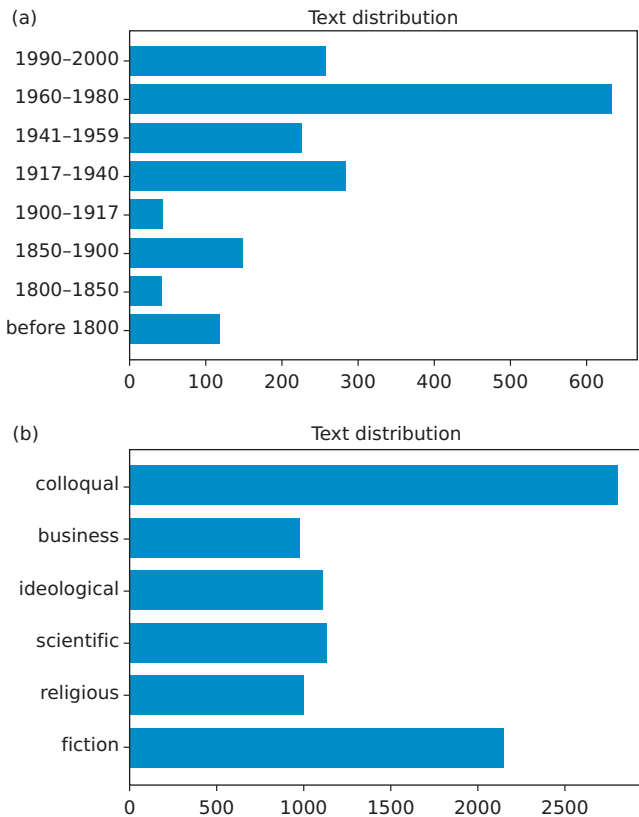


Fig. 2: Distribution of texts in the corpus by a) date and b) style.

that a single text has multiple labels for some features, these labels are separated by a comma.

5 Method

The *Live Stylistic Dictionary* service is based on two different assignments: text style identification (a multitask text classification problem) and single-word style identification. While we do have a labeled corpus to solve the text classification problem, we do not have any labeling for single words, so we have to take a kind of semi-supervised approach to word classification.

There is a huge variety of approaches to perform text classification (Zhang et al., 2015; Kowsari et al., 2019; McCallum et al., 1998; Ikonomakis, Kotsiantis, Tampakas, 2005). These methods perform rather well, and we will not discuss them any further. The main goal for us was to build a word classification pipeline.

5.1 Basic approach

We chose a classic approach based on the bag-of-words model and TF-IDF. TF-IDF is used for feature selection. We trained a logistic regression algorithm to predict the text category. As the algorithm is trained on a word presence vector, the weights of the model indicate the importance of a single word for obtaining a classification result. The probability that the text belongs to a certain category may be described using the following formula:

$$P(T) = \sigma \left(\sum_{i=1}^N \omega_i I(\omega_i \in T) + b \right)$$

In this formula, T denotes the text, ω_i is the weight of the words ω_i , $I(\cdot)$ is the indicator function for a word ω_i to appear in the text T , $\sigma(x) = \frac{1}{1+e^{-x}}$.

If ω_i is positive, the i -th word is more likely to appear in the text. We use the *Live Dictionary* to store the information about each term. This ID determines where the term's weights are held among all model weights (Fig. 2). For each term there is a weight corresponding to the specific category of a certain feature.

We also use the 2-3-gram model (Broder et al., 1997). This may help to improve the classification quality and allow us to classify different word combinations such as *в шаговой доступности* (a stone's throw), *превзойти ожидания* (to exceed expectations), etc.

This approach poses some challenges. The first problem is that, if we face an unknown word, we are not able to say anything. The second problem is the multiple forms of single words. For this approach, no stemming or lemmatization is used, because each form of the word may contain extra information that could help to classify the text. Nevertheless, it is also a problem, for the *Live Dictionary* becomes extremely large. A dataset of 4,000 texts contains more than 8 million unique terms. Moreover, in that case, we cannot say anything at all about some rare forms of a common word.

5.2 Morpheme-based approach

To overcome the difficulties discussed above, we use another approach to text preprocessing inspired by a number of authors (Joulin et al., 2016; Schütze et al., 1993; Sennrich, Haddow, Birch, 2016).

Every single word consists of letters and combinations of letters (character n-grams). These character n-grams form larger segments of the word that are called morphemes. There are several kinds of morphemes in the Russian language (prefix, root, suffix, postfix, and flection), which are located differently within the word. Moreover, a word might not contain any morphemes except the root or may have more than one morpheme of the same type (excluding flection and postfix).

Different morphemes can carry some stylistic information. Let us look at different forms of a single word. The word *кот* (cat) has many different derivatives such as *котёнок* (kitten), *котик* (pretty cat), *котейка* (nice cat, mostly used on the internet and in feminine discourse), *котэ* (cat, used on the internet by young people), *котяра* (something akin to a large, old cat, mostly used in masculine discourse), *котенька* (lovely little cat, mostly used in feminine discourse or in folklore), *котофей* (cat, in folklore), *котище* (large cat, used in common speech), etc. Some of the morphemes used in these words are absent in all morpheme dictionaries as they have emerged on the internet, where the language used is quite different to ordinary language.

Using morpheme features in classification may give us more accurate classification results. To find all the possible morphemes, we count all character n-grams of length three to six presented in the word. We also include prefixes and suffixes of lengths of up to four in the model as we have assumed that prefixes and suffixes contain important information. The morpheme length parameters are chosen on cross-validation by grid search.

This approach allows us to reduce the dimensionality of a feature space from more than 8 million terms to about 1 million terms. This makes the learning

process much quicker and also improves the quality of the text classification algorithm to 87% accuracy on the style feature.

6 Evaluation and discussion

6.1 Word style identification

There are two different variants of a single-word classification task for a model trained on morphemes. The first approach is to make a prediction for a text of a single word and then subtract the result from the classification result of a text with no words. Here, a classification result for an empty text represents the prior distribution of classes in the corpus learned by the model, and the difference represents the influence of a text on a classification result. This approach also helps us to classify word combinations.

The second approach is to take the weighted sum of the model's weights as it is done in the bag-of-words approach. Here, another hyperparameter appears – the weights of the character n-grams. If the weights are equal, we take an average model. It is rather simple, but it considers all morphemes and parts of words to be equally valuable for classification, which can hardly provide good results. Therefore, another rule may be used: the longer the character n-gram is the more valuable it is. That is better because having some information about the whole word is far more important than having some information about the suffix. On the other hand, if the word is unknown to the system, and we do not know anything about its major part, it may be classified according to its morphemes.

A comparison of these approaches to word classification is presented in Fig. 3. Here, we show the comparison of three different approaches to word classification by columns: character n-gram (Char n-gram), TF-IDF vectorizer without stemming (TF-IDF), and TF-IDF vectorizer with stemming (TF-IDF+stem). The first word was classified quite correctly using all three approaches. The only model to classify the second word correctly was the TF-IDF vectorizer as this model had already seen the word in this form. The third word was classified correctly by the first two approaches, and the third gave us a rather uncertain result. In the fourth word, there was an error, and neither the second nor the third model was able to overcome it. However, the first classifier made the correct decision. The first classifier interpreted the fifth word incorrectly, but the second one was not able to give any result. This is because the required form of the word was not presented in the dataset. Therefore, stemming generally gives us the worst result among all the classifiers. We can only apply it to classify the words occurring in the different forms,

but it may lead to errors. The second approach performs rather well, but only with words it has encountered before. The first approach gives us some misinterpretations, but it can work with words it has never encountered before.

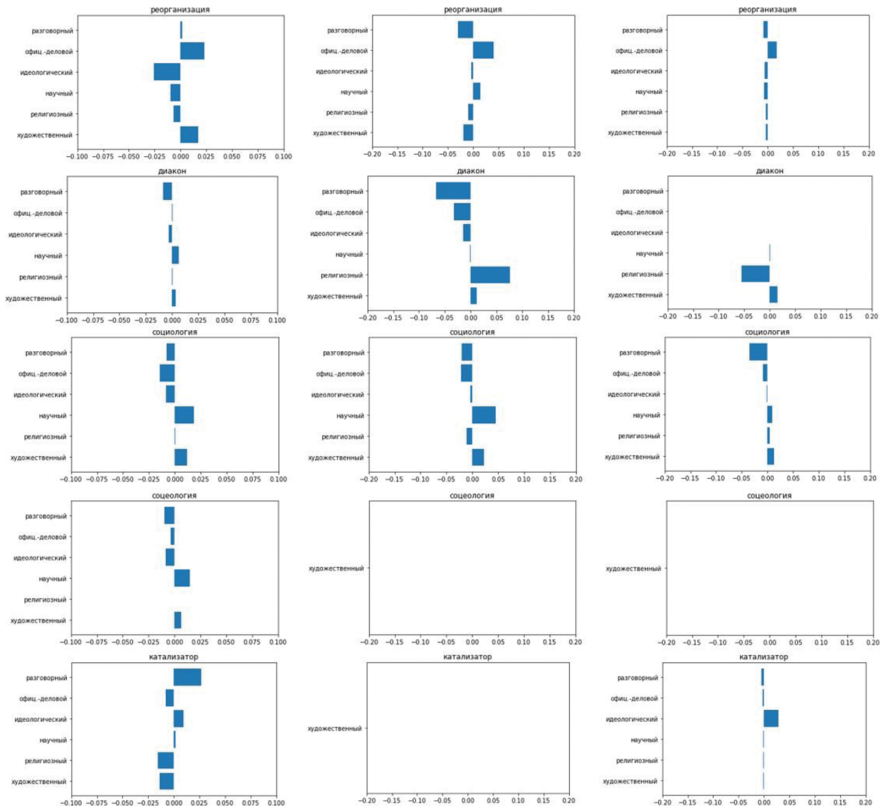


Fig. 3: Predictions of different models by columns: 1. Char n-gram; 2. TF-IDF; 3. TF-IDF+stem
Glosses: реорганизация (reorganization), диакон (deacon), социология (sociology), катализатор (catalyst).

6.2 Text classification

As previously mentioned, linear models trained on a bag-of-words model or character n-grams can be applied in text classification. Models trained on character n-grams show higher accuracy on cross-validation as they count not only the cases of word presence but also the forms of words. Tab. 1 shows a comparison of the accuracy of these methods. However, these methods cannot provide

Tab. 1: Comparison of different word style identification models in terms of text classification.

Model	Accuracy score on 5-fold cross-validation
Count vectorizer 1-gram	0.8662
Count vectorizer 1-2-gram	0.8574
Count vectorizer 1-3-gram	0.8514
TF-IDF vectorizer 1-gram	0.8033
TF-IDF vectorizer 1-2-gram	0.7635
TF-IDF vectorizer 1-3-gram	0.7396
TF-IDF vectorizer 1-gram+stemming	0.8423
TF-IDF vectorizer 1-2-gram+stemming	0.7920
TF-IDF vectorizer 1-3-gram+stemming	0.7767
Character 3-7-gram+1-4 prefix+1-3 suffix	0.8751

superior quality as they are based exclusively on word presence but do not take into account the sequence of words.

7 Conclusion

In the *Russian Live Stylistic Dictionary*, we offer a number of approaches that can significantly improve the identification of words and phrases as holders of social stylistic meaning, particularly poeticisms – words and expressions with a poetic social coloring, typical of the poetry subcorpus in the *Live Dictionary*'s fiction corpus. From a linguistic perspective, the basis for effectively recognizing Russian poeticisms is the criterion of the unity of the contextual role of the texts included in the poetic corpus. The styleset method helps to extract such texts from online resources and could be the basis for the future automatic detection and parsing of stylistically homogeneous texts and for replenishing corpora. It thus becomes possible to monitor changes in the use of the word and to trace the dynamics of its stylistic meaning, which has not been possible for traditional dictionaries. At the same time, the *Live Dictionary* aims to define the stylistic dominant of a text (the Style Prompter option – <https://livedict.syllabica.com/text>). A high concentration of poeticisms allows us to speculate that a text is of low artistic value. Combining the features of such texts could present

a universal, very stable repertoire of people who exhibit their sociocultural status and roles, such as poets and scientists. It is thus becoming possible to make a composite sketch of a typical poet, a portrait of the author of mass poetry or naive literature. On this basis, we can carry out primary diagnostics of a poem's artistic merit: compliance with the “zero idiosyncrasy” norm (a lack of personality in style) can point out the mediocrity of the poem, and deviations from it can suggest originality and even a text's uniqueness.

References

- Apresyan JD. Konnotatsii kak chast pragmatiki slova [Connotations as Part of Word Pragmatics]. In: Apresyan JD. *Izbrannyye trudy* [Selected Writings], vol. 2. Moscow: Jazyki russkoj kultury, 1995: 156–177.
- Avamilova EA, Vekshin GV, Kretov A, Maksimov ES. Algoritmy sostavleniya korpusov “Zhivogo stilisticheskogo slovarja russkogo jazyka” i ego programmaja razrabotka [Algorithms for Building the Corpora of the Russian Live Stylistic Dictionary and its Software Development]. In: *Kniga v sovremennom mire: mesto v kilturnoj paradigme obschestva v uslovijakh tzifrovoj revoliucii*. Voronezh: VGU, 2019: 4–19.
- Bakhtin M. The Problem of Speech Genres. *Literary Criticism* 2011; 4(15): 114–136.
- Bally C. *Traité de stylistique française*, vol. 1. Heidelberg: C. Winter, 1921.
- Berdyaev NA. *Sudba Rossii* [The Fate of Russia]. Moscow: Filosofskoye obschestvo SSSR, 1990 [1918]. https://imwerden.de/pdf/berdyaev_sudba_rossii_1918_1990__ocr.pdf (accessed April 1, 2022).
- Berger PL, Luckmann T. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Garden City, NY: Doubleday, 1966.
- Biber D. Representativeness in Corpus Design. *Literary and Linguistic Computing* 1993; 8(4): 243–257.
- Biber D, Conrad S. *Register, Genre, and Style*. Cambridge: Cambridge University Press, 2009.
- Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 2003; 3: 993–1022.
- Bonch-Osmolovskaya A, Orekhov B. Nekotoryye primeneniya korpusnykh metodov r naivnoy poezii [Some Applications of Corpus Methods to Naive Poetry]. 2013. http://www.ruthe.nia.ru/leibov_50/article_b-osm_orexov.html (accessed April 1, 2022).
- Broder AZ, Glassman SC, Manasse MS, Zweig G. Syntactic Clustering of the Web. *Computer Networks and ISDN Systems* 1997; 29(8–13): 1157–1166.
- Conneau A, Schwenk H, Barrault L, Lecun Y. Very Deep Convolutional Networks for Text Classification. Preprint, submitted in 2016. <https://arxiv.org/abs/1606.01781> (accessed April 1, 2022).
- Chloupek J, Nekvapil J, editors. *Studies in Functional Stylistics*, vol. 36: Linguistic and Literary Studies in Eastern Europe. Amsterdam: John Benjamins Publishing, 1993.
- Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

- Technologies, vol. 1. Stroudsburg, PA: Association for Computational Linguistics, 2019: 4171–4186.
- Diez-Arroyo M. Scientific Language in Skin-care Advertising: Persuading Through Opacity. *Revista Espanola de Linguistica Aplicada* 2013; 26: 197–214.
- Dolinin KA. *Stilytika frantzuzskogo yazyka [The Stylistics of the French Language]*. Moscow: Prosveschenije, 1987.
- Egorov E, Nikitin F, Alekseev V, Goncharov A, Vorontsov K. Topic Modelling for Extracting Behavioral Patterns from Transactions Data. In: International Conference on Artificial Intelligence: Applications and Innovations (IC-AIAI) Artificial Intelligence: Applications and Innovations (IC-AIAI) 2019: 44–49. <http://www.machinelearning.ru/wiki/images/6/69/Egorov19behavioral.pdf> (accessed April 1, 2022).
- Fedotov GP. *Stikhi dukhovnye [Spiritual Poems]*. Moscow: Progress, Gnozis Publ., 1991.
- Galitsky B, Ilvovsky D, Kuznetsov SO. Style and Genre Classification by Means of Deep Textual Parsing. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2016: 171–181. <https://www.dialog-21.ru/media/3390/galitskyba.pdf> (accessed April 1, 2022).
- Goddard C, Taboada M, Trnavac R. The Semantics of Evaluational Adjectives: Perspectives from Natural Semantic Metalanguage and Appraisal. In: *Functions of Language* 2019; 26(3): 308–342.
- Grice P. Logic and Conversation. In: Cole P, Morgan J, editors. *Syntax and Semantics 3: Speech Acts*. New York: Academic Press, 1975: 41–58.
- Grigor'ev VP, Shestakova LL. *Slovar' yazyka russkoy poezii XX veka [Dictionary of the Language of Russian Poetry of the Twentieth Century]*. Moscow: Znack, 2001.
- Guiraud P. *Essais de stylistique*. Paris: Éditions Klincksieck, 1969.
- Günthner S, Knoblauch H. Culturally Patterned Speaking Practices – The Analysis of Communicative Genres. *Pragmatics* 1995; 5: 1–32.
- Halliday MAK, Hasan R. *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective*. Geelong: Deakin University Press, 1985.
- Hansen-Löve AA. *Der Russische Formalismus: Methodologische Rekonstruktion seiner Entwicklung aus dem Prinzip der Verfremdung*. Vienna: Austrian Academy of Sciences Press, 1978.
- Harris ZS. Distributional Structure. *Word* 1954; 10(2–3): 146–162.
- Hausenblas K. On the Characterization and Classification of Discourses. In: *Travaux Linguistiques de Prague* 1966; 1: 67–83.
- Hofmann T. Probabilistic Latent Semantic Analysis. Preprint, submitted in 2013. <https://arxiv.org/abs/1301.6705> (accessed April 1, 2022).
- Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification. Preprint, submitted in 2018. <https://arxiv.org/abs/1801.06146> (accessed April 1, 2022).
- Ikonomakis M, Kotsiantis S, Tampakas V. Text Classification Using Machine Learning Techniques. *WSEAS Transactions on Computers* 2005; 4(8): 966–974.
- Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of Tricks for Efficient Text Classification. Preprint, submitted in 2016. <https://arxiv.org/abs/1607.01759> (accessed April 1, 2022).
- Jung CG. *Collected Works, vol. 7: Two Essays on Analytical Psychology*, transl. by Hull RFC. Princeton: Princeton University Press, 1966.
- Kourova OI. *Slovar' tradicionno-pojeticheskoy leksiki i frazeologii pushkinskoj epohi [The Dictionary of Traditional-Poetic Words and Phrases in the Age of Pushkin]*. Shadrinsk: Shadrinskij gos. ped. in-t, 2001.

- Kourova OI. Tradicionno-poeticheskaja leksika i frazeologija kak termin stilistiki [Traditional-Poetic Vocabulary and Phraseology as Stylistic Concept]. *Vestnik Cheljabinskogo gosudarstvennogo pedagogicheskogo universiteta* 2016; 8: 167–170.
- Kowsari K, Meimandi KJ, Heidarysafa M, Mendu S, Barnes L, Brown D. Text Classification Algorithms: A Survey. *Information* 2019; 10(4): 150. <https://www.doi.org/10.3390/info10040150>.
- Lai S, Xu L, Liu K, Jun Zhao J. Recurrent Convolutional Neural Networks for Text Classification. In: Twenty-ninth AAAI Conference on Artificial Intelligence, 2015. <https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/download/9745/9552> (accessed April 1, 2022).
- Leech G, Garside R, Bryant M. CLAWS4: The Tagging of the British National Corpus. In: COLING, vol. 1: The 15th International Conference on Computational Linguistics. 1994. <https://www.doi.org/10.3115/991886.991996>.
- Leech G. *Semantics*. Suffolk: Richard Clay, 1974.
- Li X, Li C, Chi J, Ouyang J, Li C. Dataless Text Classification: A Topic Modeling Approach with Document Manifold. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management 2018: 973–982. <https://www.doi.org/10.1145/3269206.3271671>.
- Liu P, Xipeng Qiu X, Huang X. Recurrent Neural Network for Text Classification with Multi-task Learning. Preprint, submitted in 2016. <https://arxiv.org/abs/1605.05101> (accessed April 1, 2022).
- Ljashevskaja ON, Sharov SA. Chastotnyj slovar' sovremennogo russkogo jazyka (na materialah Nacional'nogo korpusa russkogo jazyka) [Frequency Dictionary of Modern Russian (based on the Russian National Corpus)]. Moscow: Azbukovnik, 2009.
- Loper E, Bird S. Nltk: The Natural Language Toolkit. Preprint, submitted in 2002. <https://arxiv.org/abs/cs/0205028> (accessed April 1, 2022).
- Martin JR. A Contextual Theory of Language. In: Cope B, Kalantzis M, editors. *The Powers of Literacy: A Genre Approach to Teaching Writing*. London: Falmer Press, 1993: 116–136.
- McCallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification. In: AAAI-98 Workshop on Learning for Text Categorization, vol. 752. Citeseer, 1998: 41–48.
- Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and Their Compositionality. In: *Advances in Neural Information Processing Systems* 2013: 3111–3119.
- Neogi PPG, Das AK, Goswami S, Mustafi J. Topic Modeling for Text Classification. In: Mandal JK, Bhattacharya D, editors. *Emerging Technology in Modelling and Graphics*. Singapore: Springer, 2020: 395–407.
- Osgood CE, Tzeng O. *Language, Meaning, and Culture: The Selected Papers of CE Osgood*. New York: Praeger, 1990.
- Pavlinek M, Podgorelec V. Text Classification Method Based on Self-training and *l*-Topic Models. *Expert Systems with Applications* 2017; 80: 83–93.
- Pavlovich NT. Slovar' poeticheskikh obrazov: Na materiale russkoj hudozhestvennoj literatury XVIII–XX vv. T. 1–2 [Dictionary of Poetic Images: Based on Russian Fiction, vol. 1–2]. 2nd ed. Moscow: Editorial URSS, 2007.
- Peters ME, Neumann M, Lyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep Contextualized Word Representations, 2018. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1. New Orleans, LA: Association for Computational Linguistics, 2018: 2227–2237.

- Rehůřek R, Sojka P. Gensim Statistical Semantics in Python. In: EuroScipy 2011, Paris, Aug. 25–28. 8. 2011. <https://www.fi.muni.cz/usr/sojka/posters/rehurek-sojka-scipy2011.pdf> (accessed April 1, 2022).
- Russel JA. Culture and the Categorization of Emotion. *Psychological Bulletin* 1991; 110: 26–450.
- Russel JA. Core Affect and the Psychological Construction of Emotion. *Psychological Review* 2003; 110(1): 145–172.
- Santini M. Characterizing Genres of Web Pages: Genre Hybridism and Individualization. In: 40th Hawaii International Conference on Systems Science (HICSS-40 2007). Abstracts Proceedings, 3–6 January 2007, Waikoloa, HI: IEEE, 2007: 71. <https://www.doi.org/10.1109/HICSS.2007.124>.
- Schütze H. Word Space. In: *Advances in Neural Information Processing Systems*. 1993: 895–902. <https://proceedings.neurips.cc/paper/1992/file/d86ea612dec96096c5e0fcc8dd42ab6d-Paper.pdf> (accessed April 1, 2022).
- Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1: Long Papers. Berlin: Association for Computational Linguistics, 2016: 1715–1725. <https://www.doi.org/10.48550/arXiv.1508.07909>.
- Shapir MI. Yazyk byta / yazyki dukhovnoy kultury [The Language of Everyday Life / Languages of the Spiritual Culture]. *Russian Linguistics* 1990; 14(2): 129–146.
- Sharoff S. Russian Frequency Lists. June 2008. <http://corpus.leeds.ac.uk/serge/frqulist/> (accessed April 1, 2022).
- Sharov SA. Using Machine Translation for Automatic Genre Classification in Arabic. In: *Computational Linguistics and Intellectual Technologies 2018 Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”* Moscow, May 30–June 2, 2018: 153–163. https://www.dialog-21.ru/media/4292/bulyginmv_sharoffsa.pdf (accessed April 1, 2022).
- Shklovsky V. Art as Device. In: V. Shklovsky. *Theory of Prose*. Elmwood Park: Dalkey Archive Press, 1990 [1919]: 1–14.
- Spärck Jones K. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation* 1972; 28(1): 11–21.
- Stamatos E, Fakotakis N, Kokkinakis G. Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics* 2000; 26(4): 471–495.
- Vekshin GV. *Osnovy stilisticheskoi semantiki [Basics of Stylistic Semantics]*. Moscow: MUT Publ., 2017.
- Vekshin GV, Shilihina KM. Ob istochnikah slovnika “Zhivogo stilisticheskogo slovarja russkogo jazyka” [About the Sources of Glossary of the Russian Live Stylistic Dictionary]. In: *Vestnik VGU (Lingvistika i mezhkul'turnaja kommunikacija)*. Voronezh: VGU, 2017; 3: 16–20. <http://www.vestnik.vsu.ru/pdf/lingvo/2017/03/2017-03-02.pdf> (accessed April 1, 2022).
- Vekshin GV, Lemesheva MM. Poet kak rechevaya rol: k semantike I pragmatike russkogo poetizma [Poet as a Role: On the Semantics and Pragmatics of Russian Poeticism]. In: *Vestnik RUDN. Ser.: Teoriya yazyka. Semiotyka*. Semantika 2019; 10(4): 1067–1087. <https://www.doi.org/10.22363/2313-2299-2019-10-4-1067-1087>.
- Vorontsov K, Frei O, Apishev M, Romov P, Dudarenko M. Bigartm: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. In: *International Conference on Analysis of Images, Social Networks and Texts*. Berlin: Springer, 2015: 370–381.

- Wierzbicka A. A Semantic Metalanguage for a Crosscultural Comparison of Speech Acts and Speech Genres. *Language in Society* 1985; 14(4): 491–514.
- Wierzbicka A. *Emotions Across Languages and Cultures*. New York: Cambridge University Press, 1999.
- Wierzbicka A. *Semantics: Primes and Universals*. New York: Oxford University Press, 1996.
- Wittgenstein L. *Philosophische Untersuchungen = Philosophical Investigations*, transl. by Anscombe GEM, Hacker PMS, Schulte J, rev. 4th ed. Oxford: Wiley–Blackwell, 2009.
- Zhang X, Zhao J, LeCun Y. Character-level Convolutional Networks for Text Classification. In: *Advances in neural information processing systems* 2015. Cambridge: MIT Press, 2015: 649–657. <https://www.doi.org/10.48550/arXiv.1509.01626>.
- Zhang Y, Jin R, Zhou Z-H. Understanding Bag-of-Words Model: A Statistical Framework. *International Journal of Machine Learning and Cybernetics* 2010 (1): 43–52.
- Zhivoj stilisticheskij slovar' russkogo jazyka [The Russian Live Stylistic Dictionary]. Vekshin GV, Gertsev MN, Maksimov ES. 2020. <http://livedict.syllabica.com> (accessed April 1, 2022).

Benjamin Krautter, Janis Pagel, Nils Reiter, and Marcus Willand

Properties of Dramatic Characters: Automatically Detecting Gender, Age, and Social Status

Abstract: This chapter investigates the relationship between interpretative literary character types, such as the schemer, and descriptive character properties, such as gender, age, and social status. This relationship is crucial to studying dramatic characters quantitatively across a corpus of plays, as both properties and types can be used to guide a rational comparative or diachronic analysis. To this end, we first discuss the principles of character types in drama history and theater practice, connect them to gender, age, and social status, and finally discuss their possible operationalization using machine learning.

1 Introduction

Introduced to the world of theater by his friend Christlob Mylius and above all by fellow writer Christian Felix Weiße, the young student Gotthold Ephraim Lessing developed a pronounced interest in the practical side of theater, which would accompany him in the following decades. Lessing attended rehearsals, was in close contact with Friederike Caroline Neuber's troupe, and even translated plays for her stagings (cf. Barner, 1980: 108–109). Later, when he was employed by the Hamburg National Theater as a critic, Lessing discussed and reflected on current theater practice in a periodical titled *Hamburgische Dramaturgie* (1767–1769), which was initially published on a weekly basis and eventually – due to financial problems and a lack of interest from the public – as a collected volume (cf. Barner et al., 1998: 185; Harris, 1992: 229). In a 1776 letter to his brother Karl, Lessing, now associated with the Mannheim-based national theater, specifically asks for his brother's help searching for actors. Addressing his brother, he mentions the

Acknowledgment: The work in the project QuaDramA is being generously funded by the Volkswagen Foundation.

Benjamin Krautter, University of Cologne, e-mail: benjamin.krautter@uni-koeln.de

Janis Pagel, University of Cologne, e-mail: janis.pagel@uni-koeln.de

Nils Reiter, University of Cologne, e-mail: nils.reiter@uni-koeln.de

Marcus Willand, Heidelberg University

roles of a father, a mother, a male lover, a female lover, a servant, and a girl that he is looking to cast. He even adds that above-average acting skills would be sufficient (cf. Lessing, 1988: 842). These, among other examples,¹ underline Lessing's understanding of theater practice in the second half of the eighteenth century. As a playwright, he is said to have written plays for the theater he knew and roles for the actors he knew (cf. Rilla, 1977: 270–271). Accordingly, Edward P. Harris states that Lessing's knowledge of an ensemble's instrumentation and its specific skills was a key component of his dramaturgy (cf. 1992: 232).

We view Lessing as an exemplary author whose creative period coincided with the advent of the *Rollenfach* – the actor's role type – which was established as an important tool in German theater practice in the eighteenth century (cf. Maurer-Schmooch, 1982: 157; Detken, Schonlau, 2014: 9). In his still influential but rather broad definition, Bernhard Diebold explains that a role type consists of a totality of roles that are similar (1978: 9–10).² This similarity can be literary, in that it is based on the relation of its underlying character types created by the poet. Diebold mentions heroes, fathers, and schemers. But it also refers to the means of artistic performance, i.e., the individuality of the actor's acting style (cf. Doerry, 1926: 2, 5, 9). In many cases, similar literary types – for example different servants – coincide with the actor's role type (cf. Jannidis, 2007: 712). On the one hand, this demonstrates that a role type seems to be comparable to the concept of character types (cf. Maurer-Schmooch, 1982: 159; Kretz, 2012: 106–108) or even stock characters.³ On the other hand, however, the system is both quite flexible and historically variable (cf. Mehlin, 1969: 351–356; Detken, Schonlau, 2014: 19–20). It includes, for instance, a category that is specifically tailored to – so to speak – dynamic, round, complex, or individual characters, i.e., characters that are specified by their individual traits or their potential to develop new attributes (cf. Fishelov, 1990: 422–426). Thus, the “tension between the individuality of a character and the fact that this very individual is an ‘intersection’ of abstract typical traits” (Fishelov, 1990: 422) seems to apply to the role type as well. The specific role types, therefore, should not be viewed as clearly defined or as having fixed limits (Diebold, 1978: 70).

Lessing followed the conventions of the role type system to guarantee the performability of his plays. Harris argues that this is one of the keys to his

1 Lessing, e.g., reflects upon the revisions needed to transform a staged play into a written drama (cf. Lessing, 1988: 352–353).

2 The German wording of the definition reads: “Aus einer Gesamtheit – von in gewisser Beziehung – ähnlichen Rollen besteht das Rollenfach” (Diebold 1978: 9).

3 The *Oxford Dictionary of Literary Terms* defines stock characters as “stereotyped character[s] easily recognized by readers or audience from recurrent appearances” (Baldick, 2015: 342).

success at the time. For his plays *Emilia Galotti* (1772) and *Minna von Barnhelm, oder das Soldatenglück* (1767), he composed typical German characters within an ensemble's usual spectrum of roles. Although he respected the boundaries of the system, they did not restrict his plays' artistic refinement or its effect on spectators (cf. Harris, 1992: 232, 234–235).

The role type's origins lie in the *commedia dell'arte*'s improvisation troupes, which were adapted by the German *Wanderbühne* (traveling theaters) in the course of the seventeenth and early eighteenth centuries (cf. Hinck, 1965: 74–80; cf. Winter, 2014: 33–38). As the theaters' repertoires changed from improvisation pieces to French repertoires and later to regular German plays, the role type evolved as well, and a rather solid system developed. As soon as a theater group settled into a city, it had to expand its repertoire many times over to maintain visitor interest – even in small cities, a repertoire had to include at least 40 different plays (cf. Harris, 1992: 222–225). As a consequence, the system had to obey certain conventions regarding the different types of roles an actor could play. However, there were still variations depending on the theater group's economic basis and repertoire: while smaller theater troupes had to get by with only four actresses and four actors playing both tragic and comedic roles (Harris, 1992: 227), ensembles would have ideally consisted of 16 different actresses and actors (or even more) and their individual role types (cf. Diebold, 1978: 57–61; Barner et al., 1998: 82).

Table 1 illustrates the Mannheim theater's ensemble from 1778 until 1780 under the direction of Wolfgang von Dalberg (cf. Barner et al., 1998: 82; Harris, 1992: 231). The table combines the ensemble's actors with a prototypical list of role types that Johann Christian Brandes submitted when applying for the role of director of the National Theater in Mannheim in 1779 (cf. Diebold, 1978: 58–59; Maurer-Schmooch, 1982: 163). It also includes one of Lessing's plays, *Minna von Barnhelm, oder das Soldatenglück*, with Major Tellheim – the first male lover – and Minna – the first female lover – as the main characters. Lessing's play is considered a prime example of a common role type cast (cf. Schonlau, 2010: 83).

Understanding how influential the role type was for the character configuration of plays and theater practice in the eighteenth- and early nineteenth-century German theater (cf. Detken, Schonlau, 2014: 7) poses the question of whether the connection between a literary character type and the actor's role type that Diebold (1978: 10) mentions could be used for analytical purposes. To this end, we must consider two underlying issues: firstly, is it possible to find literary character types that correspond to an actor's role type and to establish consistent groups of dramatic characters that match those types? A character

Tab. 1: Ensemble of the Mannheim national theater (1778–1780).

Brandes's role types ⁴	Actress/actor	Lessing's <i>Minna von Barnhelm</i>
Tender father	Meyer	–
Comical old man	Beil	Landlord
Reasoner	Herter	Bruchsal
First lover (m)	Böck	Tellheim
Second lover (m)	Zuccarini, Beck	Paul Werner
Savant (<i>Gelehrter, Experte</i>)	Beck, Beil	Riccut
First servant	Backhaus	–
Second servant	Beil	–
Character role (schemer, etc.)	Böck, Brandes, Iffland	Just
Tender mother	–	–
Comical mother	Syler, Pöschel	Grieving lady
First lover (w)	Wallenstein, Pöschel	Minna
Second lover (w)	Seyler, Brands	–
Third lover (w) and naive roles	Toscani	–
First and second soubrette	Toscani, Kummerfeld, Pöschel	Franziska

type would then be defined – rather loosely⁵ – as a group of similar characters sharing certain properties and characteristics that make them delimitable (cf. Eder, Jannidis, Schneider, 2010: 38–42).⁶ Secondly, can we then distinguish, for instance, between typical fathers and typical mothers, or between typical lovers and typical schemers? And is it possible to do so automatically, based on quantitative analysis?⁷

4 Our translation (cf. Harris, 1992: 231).

5 The extent to which the common distinction between personification, character type, and individual character could be incorporated into this approach would need further discussion.

6 As an example, Eder, Jannidis, and Schneider describe the Prince in Lessing's *Emilia Galotti* both as “representative of the bad ruler, for he is distracted from his duties of good rule by his private feelings” and as “representative of a ruling class who uses their privileges for the satisfaction of their personal wishes and desires” (2010: 42).

7 We have pursued this from a similar perspective in Krautter et al., 2020.

2 Second-hand criticism

To perform quantitative analyses of dramatic characters, we have revived Franco Moretti's original idea of distant reading, which he first introduced in his essay *Conjectures on World Literature* (2000). Moretti imagines distant reading as second-hand criticism, as a "patchwork of other people's research, *without a single direct textual reading*" (Moretti, 2000a: 57). With his ambitious focus on world literature, he rather polemically proposes that we skip reading primary literature in favor of secondary literature and the expertise of research networks. He does so in an attempt to find a suitable method for analyzing literary history without depending "on its canonical fraction, which is not even one per cent of published literature" (Moretti, 2000a: 55, cf. Moretti, 2000b: 226).

Moretti's idea, though methodologically under-specified, seems to be useful for our approach toward analyzing literary characters. After all, the automatic identification of different character types requires both a theoretical framework for a sound typology of character types and, subsequently, annotation data that labels the literary characters. This is the groundwork required for further analyses. Considering the different role types, one could, for instance, distinguish between fathers, lovers, servants, schemers, and so on. To identify these highly specific types in dramatic texts, however, a more generalizable typology is needed as a structuring framework. For this, the traits and attributes that distinguish one group from another group of characters must be singled out. In his "ground-breaking" (Jannidis, 2014: 35) *Theory and History of Folklore* (1928), Vladimir Propp proposed a classification based on "seven areas of action" (Jannidis, 2014: 35; cf. Propp, 1968: 79–83). The actions of characters are, however, only one possibility to distinguish between different groups of characters. Other qualities that are potentially distinctive include genre, gender, nationality, social status, and age (cf. Detken, Schonlau, 2014: 11; cf. Schonlau, 2010: 82).

During our manual research process, each individual character eventually had to be assigned to a certain type of character or to certain property classes. The manually annotated data then served as input for the task of automatic classification. In order to obtain this annotation data, we followed Moretti's proposal: we carried out second-hand criticism. To find the unique properties and traits of the dramatic characters, we used attributions made by other researchers and the playwrights themselves. We searched through the plays' *dramatis personae*, different literary histories, encyclopedias, and handbooks and singled out the aforementioned character traits and attributes to label the characters with different properties. Those properties serve as a foundation for the allocation and identification of character types that are associated with a role type.

The corpus of plays we used for this chapter to automatically predict character properties consists of 41 plays written between 1730 and 1850.⁸ It includes a total of 247 annotated characters. On average, we annotated 6.0 characters per play. So far, we have focused heavily on a relatively short time span of about 120 years. There are two underlying reasons for this: on the one hand, the production of German plays was overshadowed by the success of Realist prose genres from around 1850 until 1880 (cf. Begemann, 2007: 10). On the other hand, the role type and its influence on theater practice changed in the course of the nineteenth century: new types of roles emerged, their boundaries blurred even more, and the size of ensembles changed – as repertoires grew, more actresses and actors had to be employed (cf. Doerry, 1926: 33–38).

3 Annotation and operationalization

In the following, we will give a short example of our annotation process and the categories we used to annotate the dramatic characters. For this purpose, we will focus on Friedrich Schiller's *Maria Stuart*, which premiered in 1800. As indicated earlier, the annotation process itself was twofold. Firstly, we singled out the specific traits of each character in question. With this knowledge in hand, we then matched the character with the best-fitting character types.

The character we are focusing on first is Queen Elisabeth, one of the two main characters in Schiller's *Maria Stuart*. Elisabeth is the antagonist to the eponymous character Maria Stuart. Table 2 shows Elisabeth's characteristics according to six different sources and the *dramatis personae*. Although some sources are more comprehensive than others, they are rather consistent: all together, Elisabeth seems to be a power-hungry queen and Maria's jealous rival.

Operationalization describes the process of making theoretical concepts measurable. The annotation and classification of literary characters is a major operationalization task. However, literary concepts such as different character classifications tend to be complex and are in most cases not intended to be measured (cf. Moretti, 2013: 1, 9). As indicated above, different typologies have been proposed to classify characters and their specific attributes. In 1927, EM Forster suggested distinguishing between round and flat characters: "Flat characters were called 'humorous' in the seventeenth century, and are sometimes called types, and sometimes caricatures. [. . .] In their purest form, they are constructed round a single idea or quality: when there is more than one factor

⁸ For this, we make use of the *German Drama Corpus* (Fischer et al., 2019).

Tab. 2: Qualities and traits of Elisabeth in Schiller's *Maria Stuart* according to seven sources.⁹

Source	Qualities and traits
<i>Dramatis personae</i> (Schiller, 1948: 2)	Queen of England
<i>Geschichte des Dramas</i> (Fischer-Lichte, 1999: 366–381)	Queen, competitor, woman, protestant, virgin, ruler, regent, renunciation of instincts, ethics of performance, exemplary, envious, power-hungry, addicted to ruling
<i>Kindlers Literaturlexikon</i> (Sautermeister, 2009: 520–521)	Queen, arrogant, rival
<i>Geschichte des deutschen Dramas</i> (Mann, 1969: 283–288)	Not beautiful, queen, jealous, tormentor
<i>Geschichte der deutschen Literatur</i> (Kurz, 1861: 436–437)	Woman, queen, character based on hypocrisy and pretense, vanity
<i>Die deutsche Nationalliteratur</i> (Hillebrand, 1875: 442–444)	Queen, enemy, Protestant, drawn in spiteful light, reduced to the lowest level of common passion, shows no royal dignity
<i>Schiller-Handbuch</i> (Vonhoff, 2011: 153–168)	Queen, miserable, criminal, driven by power, addiction and will to power

in them, we get the beginning of the curve towards the round” (Forster, 1927: 103–104). This dichotomy was perceived as “highly reductive, obliterating the degrees and nuances found in actual works of narrative fiction” (Rimmon-Kenan, 2002: 40–41). In light of this critique, more sophisticated taxonomies have been developed. Baruch Hochman, e.g., lists eight dichotomous categories, such as “Wholeness” and “Fragmentariness,” “Dynamism” and “Staticism” or “Stylization” and “Naturalism” (Hochman, 1985: 89; cf. Fishelov, 1990: 424) to distinguish literary characters. Drama analysis generally uses three concepts to categorize characters, ranging from one-dimensional to multi-dimensional characters: personifications, character types, and individual characters (cf. Pfister, 2001: 243–245)

After investigating the annotated data, which is summarized in Tab. 3, we established a basic operationalization consisting of three different properties that serve as a starting point for the identification of character classes (cf. Bullard, Ovesdotter Alm, 2014: 11–12): gender, age, and social status. According to Anke Detken and Anja Schonlau, these properties are important indicators that can be

⁹ Our translation.

Tab. 3: Distribution of annotated character types.

Character Type	Frequency
aristocrat	41
servant	30
daughter	24
married person	21
schemer	20
beloved	18
citizen	18
father	18
lover	18
military	18
son	15
unmarried person	15
official	14
confidant	12
monarch	12
low social status	10
opponent	10
scholar	10
mother	9
cleric	8
brother	7
widow	7
artist	5
judiciary	5
sister	5
old person	4
tyrant	4

Tab. 3 (continued)

Character Type	Frequency
loving person	3
republican	3
strategist	3
tender father	3
messenger	2

used to differentiate between various role types (cf. 2014: 11).¹⁰ In the following step, we were able to combine them with further context, for instance, genre information and the distinction between major and minor characters.¹¹ In total we established 18 different classes to which a dramatic character can be assigned (cf. Tab. 4). These classes, however, are not to be equated with character types. Rather, they form the basis for locating the more specific character types in one of the classes. Consequently, we opted for an approximate operationalization using properties that are related to the concept of character types but not congruent with them (cf. Reiter, Willand, 2018: 54, 69).

The difficulty with a typology such as this is maintaining a good balance between complexity and feasibility. A binary distinction between male and female characters alone would not be complex enough with respect to literary characters, nor would groups of clearly distinguishable character types emerge. It could only serve as a starting point. Distinguishing a virtuous father from a tender one, a noble, grumbling, or benevolent father (cf. Doerry, 1926: 11), however, is too finely granular for a computer-based approach that builds on training data. And even for literary scholars, the subtle differences can be hard to grasp.

While mapping each character to a character class, the premise was to find the best-fitting matches considering both the transformations that have taken place in drama history, e.g., the establishment of the bourgeois tragedy, and the single fictional reality of the play. In the case of Elisabeth, we annotated her as a middle-aged female character of the upper class.

¹⁰ Bullard and Ovesdotter Alm use these criteria for a binary classification of dramatic characters into male and female, young and old, as well as upper-middle class and lower class (2014: 11–16).

¹¹ Our preliminary work includes the automatic identification of main characters (cf. Krautter et al., 2018: 1–56).

As a good balance between complexity and feasibility is hard to achieve, annotation is an ongoing, iterative process, where adjustments to the categories and the typology are still being carried out (cf. Pagel et al., 2018: 31–36).

4 Experiments

In the following section, we explain how we used the annotations described above for a set of experiments. Each of the three experiments aimed to classify characters according to the different classes of one of the three properties gender, age, and social status. We represented each character as a set of features that have been shown to successfully help to classify other kinds of character properties, such as being a protagonist (cf. Krautter et al., 2018). Using machine learning, we then classified each character according to one of the property classes (binary for gender and tertiary for age and social status) using their set of features. The results indicate how well a feature is suited to classifying a certain character type. A detailed analysis using feature importance in the form of a Shapley analysis gives insights into single characters and how their features led to a certain outcome in the machine learning model.

4.1 Experimental setup

In total, the data comprises 41 plays with 247 annotated characters. The distribution of classes per property is displayed in Tab. 4. There are some class imbalances as, e.g., male characters make up two-thirds of the gender class.

Tab. 4: Overview of character classifications.

Gender	Social status	Age	Frequency
male	high	young	5
male	high	middle-aged	33
male	high	old	15
male	middle	young	24
male	middle	middle-aged	39
male	middle	old	25

Tab. 4 (continued)

Gender	Social status	Age	Frequency
male	low	young	9
male	low	middle-aged	11
male	low	old	6
female	high	young	7
female	high	middle-aged	16
female	high	old	3
female	middle	young	23
female	middle	middle-aged	8
female	middle	old	7
female	low	young	9
female	low	middle-aged	3
female	low	old	4

For our machine learning algorithm, we opted for a support vector machine (SVM) (Steinwart, Christmann, 2008). To counter the class imbalances, we applied SMOTE sampling (Chawla et al., 2002). SMOTE generates artificial data points by using the closest neighbors in the feature space of the existing data points. This increases the size of the underrepresented class(es). For training and testing, we used ten-fold cross validation and iterated the process ten times to ensure that the distribution of training and test set split would have no impact on the results. We evaluated each model according to precision, recall, and its F1-score. Additionally, we made use of a “Shapley analysis” (Shapley, 1953), where the distribution of the features’ importance is plotted for a single character. This allowed us to investigate the impact each single feature value had on predicting a certain class. A negative coefficient ϕ indicates that the feature value was detrimental for assigning the class; a positive ϕ indicates that a feature value was beneficial for assigning the class. Greater values of ϕ indicate a greater influence of the feature value in making a decision in favor of a class.

We represent each character as a set of features, which we outlined in Tab. 5. The *tokens* feature displays the number of tokens that are uttered by a character. The feature is normalized by the total number of tokens in the play. The normalization ensures comparability between different plays. *Utterances* represent the total number of utterances a character expresses. An utterance is defined as the

Tab. 5: Feature set overview.

Feature set	Feature name	Possible values	Short description
	tokens	0–1	Normalized number of tokens a character utters
Utterance	utterances	Natural Numbers	Number of utterances of a character
	utteranceLengthMean	Positive Rational Numbers	Mean length of utterances of a character in tokens
	TTR	Positive Rational Numbers	Type-token ratio
Centrality	degree	0–1	Degree centrality
	wdegree	Natural Numbers	Weighted degree centrality
	between	0–1	Betweenness centrality
	close	0–1	Closeness centrality
	eigen	0–1	Eigenvector centrality
Presence	actives	0–1	Active presence
	passives	0–1	Passive presence
Topics	T1–T5	0–1	Topic model
WF	Liebe, Familie, Religion, Ratio, Krieg	0–1	Word fields: love, family, religion, reason, war
	nfig	Natural Numbers	Number of characters in play
	lastAct	Boolean	Whether a character appears in the final act

text a character utters until the speaker switches. *UtteranceLengthMean* is the mean length of a character’s utterances. We also computed the type-token ratio (TTR) per character. In order to capture the role a character assumes in the configuration of a play, we computed several centrality measures based on the plays’ co-presence graphs. A co-presence graph comprises a set of nodes and edges. Each node represents a character in the specific play. Two nodes are connected via an edge if the two characters co-occurred in a scene (cf. Trilcke, 2013: 224–225). The number of joint occurrences determines the weight of the respective edge. It is now possible to compute centrality measures of the graph. Centrality measures comprise a set of measures to determine how central a node is for a

graph. In our case they determine the centrality of a character in the play's configuration.¹² We also measured the characters' presence in the play, both active and passive. Active presence describes the number of scenes in which a character is present, i.e., uttering at least one token in a scene. Passive presence is defined as the number of scenes in which a character is not present but is mentioned by another character by name. Both presence values are normalized by the total number of scenes. In order to get an idea of the kinds of topics a character talks about, we applied two strategies. Firstly, we trained a topic model on the 41 plays, excluding stop words, which resulted in five topics. For each character, the probability of them talking about a particular topic is calculated. Secondly, we manually created lists, consisting of about 70–110 words each, for five word fields: *Liebe* (love), *Familie* (family), *Religion* (religion), *Krieg* (war), and *Ratio* (reason), which we assume were common in plays between 1770 and 1830 (cf. Willand, Reiter, 2017: 84–85). We then counted how often the characters utter a word from a particular word field. Lastly, we included the feature *nfig* to capture how many characters are present in a play, and *lastAct*, a Boolean feature that expresses whether a character is present in the final act of a play or not.

To investigate the effect of the features on classification, we grouped similar features into feature sets.

4.2 Experiment 1: Gender classification

In our first experiment, we investigated how well the features predict gender. When classifying each character using an SVM and the described feature sets, we obtained the scores shown in Fig. 1 for predicting male and female characters.

Some observations can be made: using all features together yields by far the best results. The features thus seem to add different information, which is required in order to classify gender. Next to all features, using the centrality-based and word-field-based features led to the second-best results. Consequently, male and female characters appear to differ in the way they talk semantically as well as in the way they interact with other characters. To explore the differences between male and female characters with respect to the various features, we can look at the distribution of features per class, which is displayed in Fig. 2.

This graph shows the number of characters in a certain class that exhibit a specific feature value. For instance, many female and male characters have a

¹² The centrality measures used are described in detail in Krautter et al. (2018).

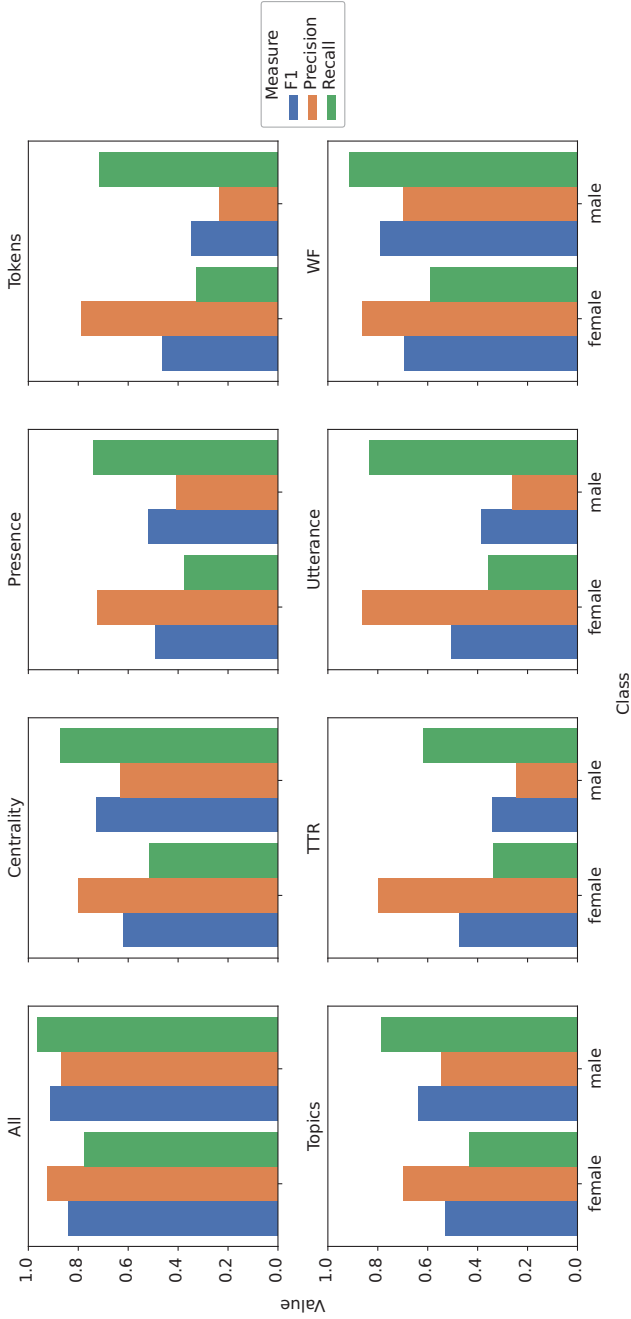


Fig. 1: Classification performance for female and male characters.

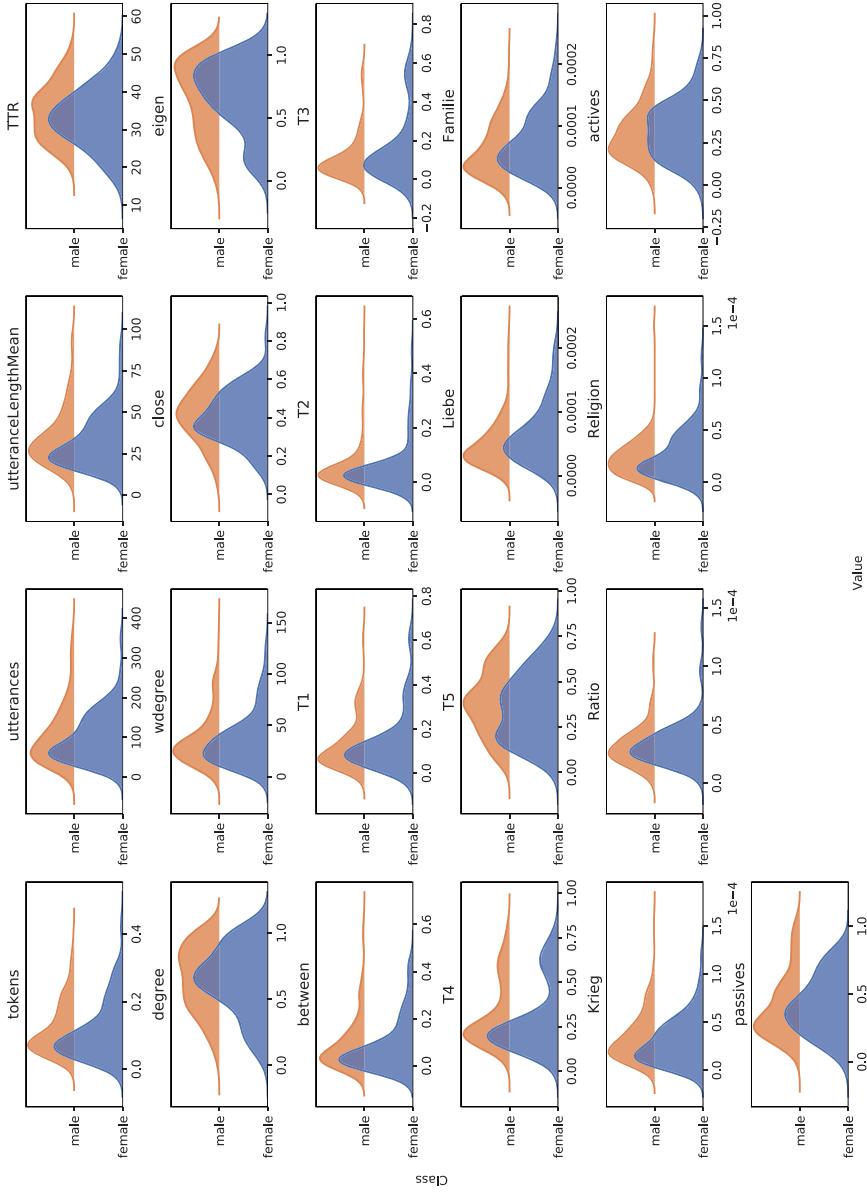


Fig. 2: Feature value distribution.

token value close to 0.1, while some female characters have a higher token value than any of the male characters. This shows that, for the majority of characters, the number of spoken tokens is not a differentiating factor between male and female characters, but for the very characters that speak a larger portion of the tokens, it is. A distribution such as this allows us to examine which features are most distinctive within the classes and which features might be the most telling for the SVM during classification. The distribution of centrality features displays some clear differences, showing that a few male characters reach higher values than female characters. We can also identify some substantial differences in the distributions of presence features, but, apparently, these differences alone did not help the SVM to distinguish between the classes very well (otherwise, performance scores for the centrality feature set would be higher; cf. Fig. 1).

A more direct way of investigating the importance of features for classification is what is referred to as feature or variable importance. Feature importance measures how much the performance of the model drops when a feature is left out, thus giving insight into the importance of the feature for the classification. We trained a new model on the complete data set using a random forest classifier (Ho, 1995), which offers a direct way of obtaining feature importance. Fig. 3 shows the feature importance values in a bar plot.

Apparently, the word field and topic-model-based features, as well as the type-token-ratio, are most important for the classifier's decision. Accordingly, the gender classes seem to differ mostly in the way female and male characters speak. Centrality-based features, on the other hand, did not have the biggest impact on the model that uses all features, although they performed quite well on their own. This illustrates that, in this case, the word fields and topics mostly capture the differences that the centrality features are able to cover.

It can also be instructive to look at single characters and the features that were decisive for their classification. The Shapley graph in Fig. 4 shows the feature importance for classifying two characters from Lessing's *Miß Sara Sampson*: Sara and Waitwell.

Both characters have been classified correctly. One rather informative observation in this regard is that the model has learned from the data to more likely classify characters as male when they are frequently talking about reason-related things and as female when they are talking more about love-related things.

4.3 Experiment 2: Age classification

To classify age, we distinguished between the classes *young*, *middle-aged*, and *old*, and used the same feature set as before. Fig. 5 shows the model's performance.

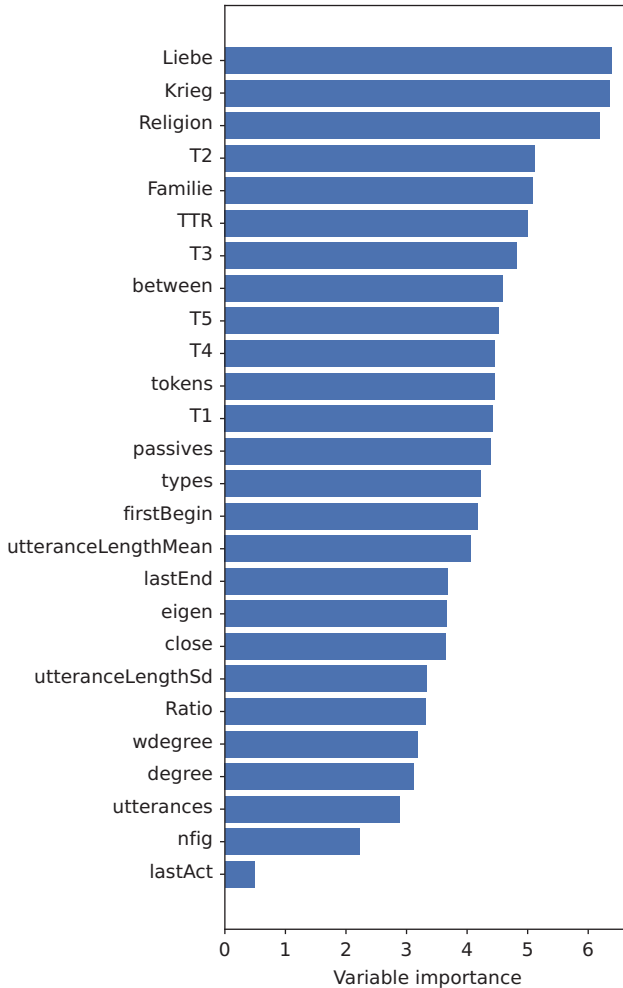


Fig. 3: Feature importance.

Once again, using all features yielded the best results. For all three classes, centrality features, topics, and the word fields performed best when using only a single feature set. As for the gender classification, this demonstrates that features capturing the semantic content of character speech, in one way or the other, are best suited to classifying these character properties.

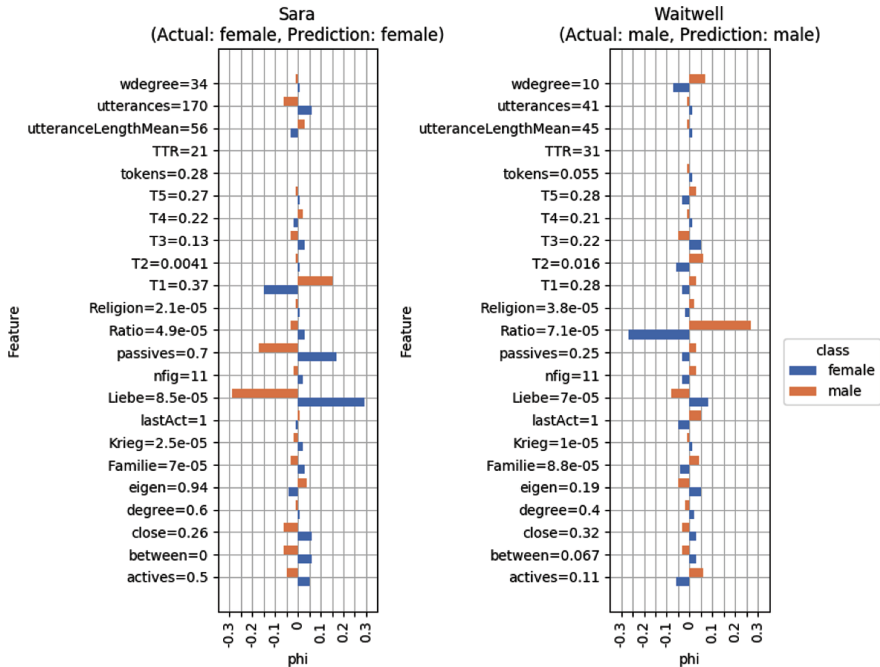


Fig. 4: Contribution of feature values for two individual predictions.

4.4 Experiment 3: Status classification

To classify characters according to their social status, we distinguished between low, middle, and high status, again using the same feature set as above. Class distribution is skewed, with characters of middle status the most frequent (126 of 247 characters), followed by high (79) and low status (42).

The general classification performance for the different feature sets can be found in Fig. 6. As before, a combination of all feature sets performed best. On the one hand, the low-status class achieved rather high precision for all features. But on the other hand, the recall for low-status characters was rather low. This means that the model was able to recognize a few low-status characters very reliably but assumes that most of them are either of middle or high status. The features that represent the characters' speech content (word fields and topics) and centrality fared comparably well in this classification. This suggests that the characters' social status is indeed reflected in the character speech itself as well as in their relationships to other characters.

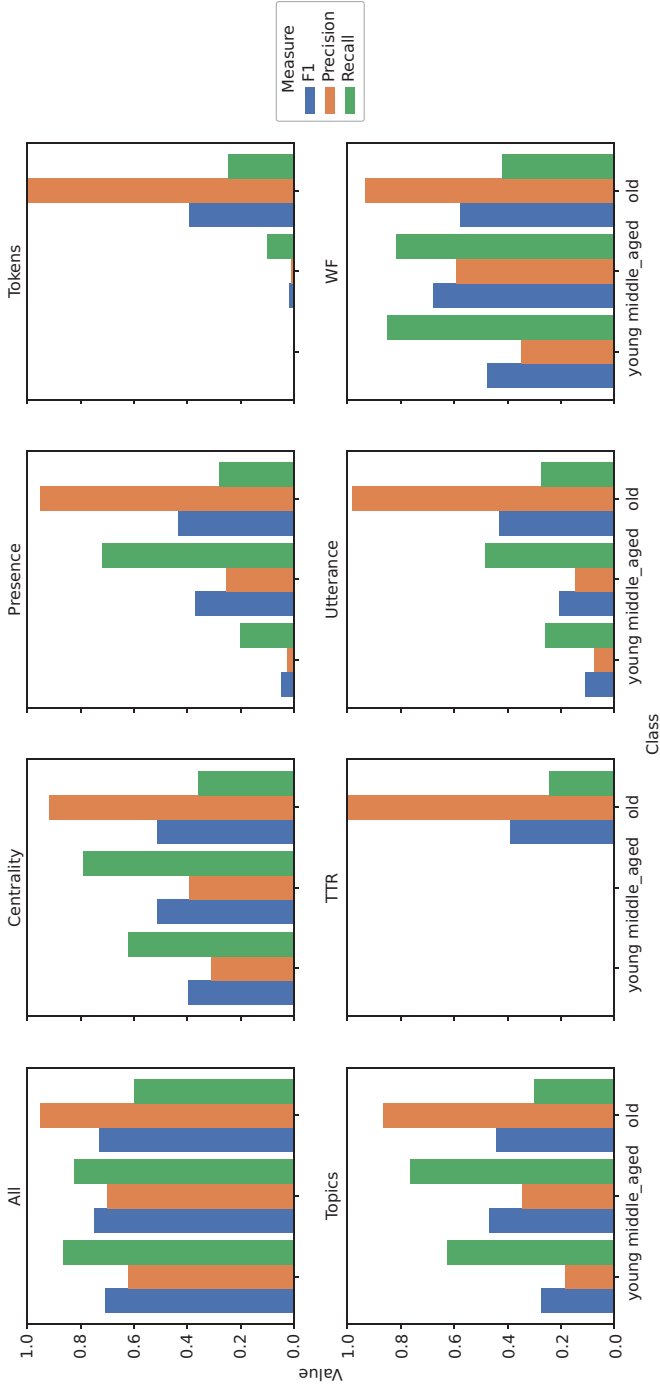


Fig. 5: Classification performance for age.

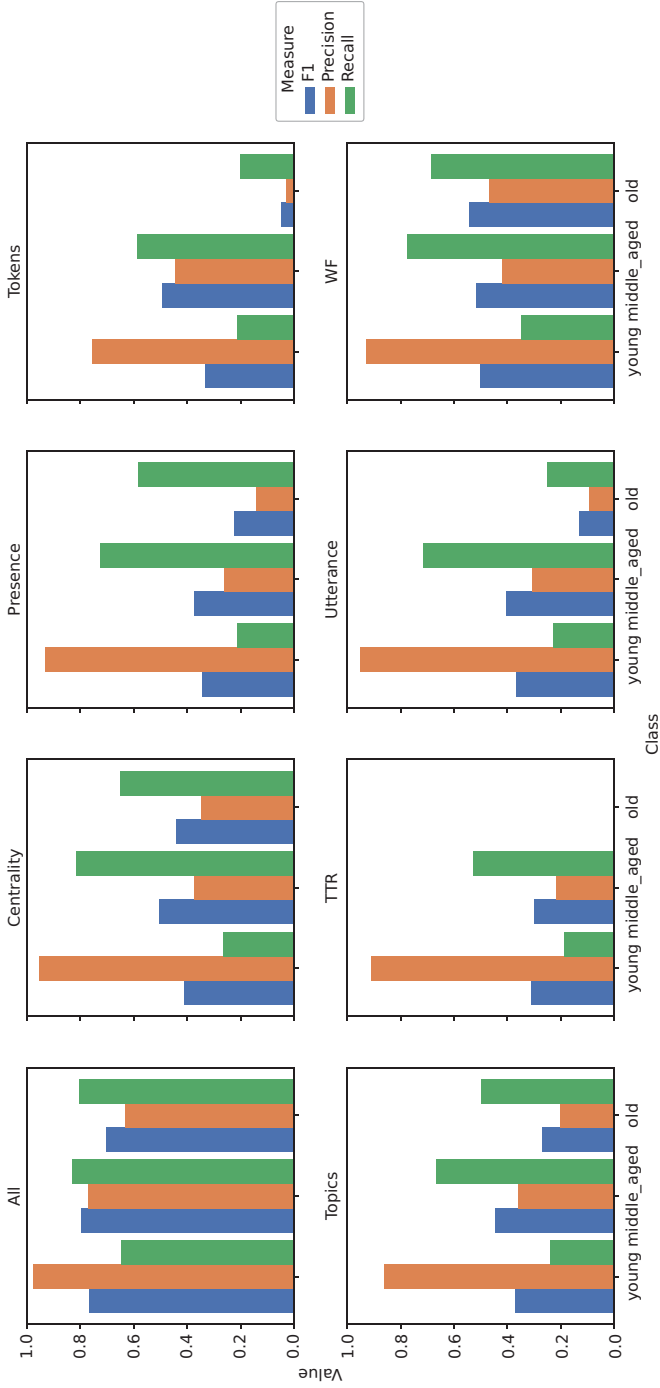


Fig. 6: Classification performance for social status.

5 Conclusion

Detecting character types is a crucial milestone in character analysis across individual plays, as this step allows us to meaningfully compare larger numbers of characters. We can expect the operationalization of character types, i.e., the development of procedures to detect and distinguish between them (cf. Pichler, Reiter, 2021: 4), to be quite complex as it involves interpretative decisions that may rely on text-external context. It is, nevertheless, quite clear that the character properties we have discussed here – gender, age, and social status – are relevant features for the subsequent detection of character types: male and female characters have a clearly separated character type inventory, some types can only be implemented by old characters, and others are related to their social status.

Operationalizing these – presumably simpler – character properties, however, reveals complexities that might be underestimated at first sight: social status can only be annotated as a relative category, potentially accounting for a character in one play as a low-status character (compared to the king) but as a high-status character in another (compared to non-noble characters). This, in turn, will require a reliable automatic classification approach to incorporate information that is much more play-specific than the information discussed in our model above. For the property of age, another challenge emerges as it is underspecified for many characters. Given the textual information, one property value might be excluded (e.g., a character with children is unlikely to be very young), but it still leaves multiple possible values open. Hence, it is difficult for annotators not to rely on stereotypes or interpretations of the play when making these decisions.

Another methodological challenge is that finding historical data on theater performance is a major task in itself. The concept of the role type has its origin in performance practice, and the actors' performances would certainly be very relevant for any kind of classification. As machine-readable data on historical performances is not available, we have used the characters' written speech from the dramatic texts.

We have already pointed out the next steps on our research agenda. While this chapter focuses on the automatic identification of (relatively) simple character properties, our long-term objective is to a) operationalize character types and thus b) contribute to drama history research. We aim to explore how different character types are established, how their popularity and their qualities change over time, and how they relate to and interconnect with other types.

References

- Baldick C. *The Oxford Dictionary of Literary Terms*. Oxford: Oxford University Press, 2015.
- Barner W. Lessing als Dramatiker. In: Hinck W, editor. *Handbuch des deutschen Dramas*. Düsseldorf: Bagel, 1980: 106–119.
- Barner W, Grimm G, Kiesel H. *Lessing: Epoche – Werk – Wirkung*. Munich: CH Beck, 1998.
- Begemann C. Einleitung. In: Begemann C, editor. *Realismus: Epoche – Autoren – Werke*. Darmstadt: Wissenschaftliche Buchgesellschaft, 2007: 7–10.
- Bullard J, Ovesdotter Alm C. Computational Analysis to Explore Authors' Depiction of Characters. In: *Proceedings of the 3rd Workshop on Computational Linguistics for Literature*. Gothenburg: Association for Computational Linguistics, 2014: 11–16.
- Chawla N, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. In: *Journal of Artificial Intelligence Research* 2002; 16: 321–357.
- Detken A, Schonlau A. *Das Rollenfach – Definition, Theorie, Geschichte*. In: Detken A, Schonlau A, editors. *Rollenfach und Drama*. Tübingen: Narr Verlag, 2014: 7–30.
- Diebold B. *Das Rollenfach im deutschen Theaterbetrieb des 18. Jahrhunderts*. Nendeln: Kraus Reprint, 1978 [1913].
- Doerry H. *Das Rollenfach im deutschen Theaterbetrieb des 19. Jahrhunderts*. Berlin: Selbstverlag der Gesellschaft für Theatergeschichte, 1926.
- Eder J, Jannidis F, Schneider R. Characters in Fictional Worlds: An Introduction. In: Eder J, Jannidis F, Schneider R, editors. *Characters in Fictional Worlds: Understanding Imaginary Beings in Literature, Film, and Other Media*. Berlin, New York: De Gruyter, 2010: 3–64.
- Fischer F, Börner I, Göbel M, Hechtl A, Kittel C, Milling C, Trilcke P. Programmable Corpora: Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor. In: Sahle P, editor. *DHD 2019. Digital Humanities: multimedial & multimodal. Conference abstracts*. Frankfurt a.M., Mainz, 2010: 194–197. <https://doi.org/10.5281/zenodo.2596094>.
- Fischer-Lichte E. *Geschichte des Dramas, vol. 1: Von der Antike bis zur deutschen Klassik*. Tübingen, Basel: A. Francke Verlag, 1999.
- Fishelov D. Types of Character, Characteristics of Types. In: *Style* 1990; 24(3): 422–439.
- Forster EM. *Aspects of the Novel*. New York: Harcourt, Brace and Company, 1927.
- Harris EP. Lessing und das Rollenfachsystem: Überlegungen zur praktischen Charakterologie im 18. Jahrhundert. In: Bender WF, editor. *Schauspielkunst im 18. Jahrhundert: Grundlagen, Praxis, Autoren*. Stuttgart: Franz Steiner Verlag, 1992: 221–235.
- Hillebrand J. *Die deutsche Nationalliteratur im XVIII. und XIX. Jahrhundert, vol. 2: Die deutsche Nationalliteratur im letzten Viertel des XVIII. Jahrhunderts*. Hamburg, Gotha: Perthes, 1875.
- Hinck W. *Das deutsche Lustspiel des 17. und 18. Jahrhunderts und die italienische Komödie: Commedia dell'Arte und Théâtre Italien*. Stuttgart: JB Metzler, 1965.
- Ho, TK. Random Decision Forests. In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montréal: Institute of Electrical and Electronics Engineers (IEEE)*, 1995: 278–282. <https://doi.org/10.1109/ICDAR.1995.598929>.
- Hochman, B. *Character in Literature*. Ithaca: Cornell University Press, 1985.
- Jannidis F. Typologie3. In: Braungart G, Fricke H, Grubmüller K, Müller J-D, Vollhardt F, Weimar K, editors. *Reallexikon der deutschen Literaturwissenschaft: Neubearbeitung des Reallexikons der deutschen Literaturgeschichte, vol. 3: P–Z*. Berlin, New York: De Gruyter, 2007: 712–713.

- Jannidis F. Character. In: Hühn P et al., editor. *Handbook of Narratology*, vol. 1. Berlin, Boston: De Gruyter, 2014: 30–45.
- Krautter B, Pagel J, Reiter N, Willand M. Titelhelden und Protagonisten – Interpretierbare Figurenklassifikation in deutschsprachigen Dramen. In: *Litlab Pamphlet* 2018; 7: 1–56.
- Krautter B, Pagel J, Reiter N, Willand M. “[E]in Vater, dächte ich, ist doch immer ein Vater”: Figurentypen und ihre Operationalisierung. In: *ZfdG* 2020; 5. https://doi.org/10.17175/2020_007.
- Kretz N. Bausteine des Dramas (Figur, Handlung, Dialog). In: Marx PW, editor. *Handbuch Drama: Theorie, Analyse, Geschichte*. Stuttgart, Weimar: JB Metzler, 2012: 105–121.
- Kurz H. *Geschichte der deutschen Literatur: Mit ausgewählten Stücken aus den Werken der vorzüglichsten Schriftsteller*, vol. 3. Leipzig: BG Teubner, 1861.
- Lessing GE. *Werke und Briefe in zwölf Bänden*, vol. 11/2: *Briefe von und an Lessing 1770–1776*. Kiesel H, editor. Braungart G, Fischer K, Wahl U, collaborating editors. Frankfurt a.M.: Deutscher Klassiker Verlag, 1988.
- Mann O. *Geschichte des deutschen Dramas*. Stuttgart: Kröner, 1969.
- Maurer-Schmook S. *Deutsches Theater im 18. Jahrhundert*. Tübingen: Niemeyer, 1982.
- Mehlin UH. *Die Fachsprache des Theaters: Eine Untersuchung der Terminologie von Bühnentechnik, Schauspielkunst und Theaterorganisation*. Düsseldorf: Pädagogischer Verlag Schwann, 1969.
- Moretti F. Conjectures on World Literature. In: *New Left Review* 2000a; 1: 54–68.
- Moretti F. The Slaughterhouse of Literature. In: *Modern Language Quarterly* 2000b; 61(1): 207–227.
- Moretti F. “Operationalizing”: Or, the Function of Measurement in Modern Literary Theory. In: *Pamphlets of the Stanford Literary Lab* 2013; 6: 1–13.
- Pagel J, Reiter N, Rösiger I, Schulz S. A Unified Text Annotation Workflow for Diverse Goals. In: Kübler S, Zinsmeister H, editors. *Proceedings of the Workshop for Annotation in Digital Humanities*. Sofia: CEUR Workshop Proceedings, 2018: 31–36. <http://ceur-ws.org/Vol-2155/> (accessed May 5, 2022).
- Pfister M. *Das Drama: Theorie und Analyse*. Munich: Wilhelm Fink, 2001.
- Pichler A, Reiter N. Zur Operationalisierung literaturwissenschaftlicher Begriffe in der algorithmischen Textanalyse: Eine Annäherung über Norbert Altenhofers hermeneutische Modellinterpretation von Kleists *Das Erdbeben in Chili*. In: *Journal of Literary Theory* 2021; 15(1–2): 1–29.
- Reiter N, Willand M. Poetologischer Anspruch und dramatische Wirklichkeit: Indirekte Operationalisierung in der digitalen Dramenanalyse; Shakespeares *natürliche* Figuren im deutschen Drama des 18. Jahrhunderts. In: Bernhart T, Willand M, Richter S, Albrecht A, editors. *Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven*. Berlin, Boston: De Gruyter, 2018: 45–76.
- Rilla P. *Lessing und sein Zeitalter*. Munich: Beck, 1977.
- Rimmon-Kenan S. *Narrative Fiction: Contemporary Poetics*. London, New York: Routledge, 2002.
- Sautermeister G. Maria Stuart: Ein Trauerspiel. In: Arnold HL, editor. *Kindlers Literaturlexikon*, vol. 14: *Ror-Sez*. Stuttgart, Weimar: JB Metzler, 2009: 520–521.
- Schiller F. Maria Stuart: Ein Trauerspiel. In: *Schillers Werke: Nationalausgabe*, vol. 9, ed. by Petersen J, Schneider H. Weimar: Hermann Böhlaus Nachfolger, 1948: 1–164.
- Schonlau A. Rollenfach und Geschlecht im Drama des 18. Jahrhunderts. In: *Der Deutschunterricht* 2010; 62(6): 82–87.

- Shapley, LS. A Value for n-Person Games. In: Kuhn HW, Tucker AW, editors. Contributions to the Theory of Games, vol. 2. Princeton, NJ: Princeton University Press, 1953: 307–317.
- Steinwart I, Christmann A. Support Vector Machines. New York: Springer, 2008.
- Trilcke P. Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft. In: Ajouri P, Mellmann K, Rauen C, editors. Empirie in der Literaturwissenschaft. Münster: Mentis, 2013: 201–247.
- Propp V. Theory and History of Folklore. Minneapolis: University of Minnesota Press, 1968 [1928].
- Vonhoff G. *Maria Stuart*: Trauerspiel in fünf Aufzügen (1801). In: Luserke-Jaqui M, editor. Schiller-Handbuch: Leben – Werk – Wirkung. Stuttgart, Weimar: JB Metzler, 2011: 153–168.
- Willand R, Reiter N. Geschlecht und Gattung: Digitale Analysen von Kleists “Familie Schroffenstein.” In: Kleist-Jahrbuch 2017: 177–195.
- Winter S. Von der Maske zur Rolle, vom *Magnifico* zum Familienvater: Die Fächerrezeption der Commedia dell’arte am Beispiel des Pantalonos. In: Detken A, Schonlau A, editors. Rollenfach und Drama. Tübingen: Narr Verlag, 2014: 33–47.

Pablo Gervás

N-Gram-Driven Word Level Recombination: Exploring a Search Space of Metrically Valid Verse

Abstract: Procedures for generating poetry by recombining textual fragments give rise to very large spaces for solutions. While this is part of their charm, it also poses a significant challenge for readers trying to appraise their aesthetic merit. The present chapter proposes a computational solution to the exploration of such large spaces. The solution is based on sampling the search space in an informed manner driven by simple quantitative metrics about elementary features of poems. The quantitative metrics are selected empirically and validated in terms of their potential value for discriminating between different corpora of human-generated poems, poems generated by recombining human-written verse, and poems generated entirely by computational methods. This chapter postulates a mechanism for extracting valuable samples from the search space based on the establishment of a target vector of values for the identified features and applying a distance measure to the representations of this target and the features for each candidate poem in the corresponding vector space.

1 Introduction

Recombining textual fragments at random to generate interesting texts has been a known literary method since the work of the Dadaists in the 1920s (Zubrug, 1979), and it was popularized in the 1950s by Gysin and William Burroughs (Cran, 2013). These approaches involved repurposing existing texts by cutting them up into fragments that were then recombined into new texts. Such processes in practice define a search space of possible texts determined by the set of

Acknowledgements: This project has been partially supported by project CANTOR (Ref PID2019-108927RB-I00), funded by the Spanish Ministry of Science and Innovation. The work reported on in this chapter was heavily influenced by discussions with Thierry Poibeau and Pablo Ruiz Fabo during my visit to the Lattice Laboratory at École Normale Supérieure, Paris, in January 2020.

Pablo Gervás, Institute of Knowledge Technology, Faculty of Computer Science, Complutense University of Madrid, e-mail: pgervas@ucm.es

source texts used and the procedure used to recombine them. In his book *Cent mille milliards de poèmes*, Raymond Queneau (1961) proposed an approach to poetry that involved not so much particular poems produced in such a fashion but rather a special format of book designed to allow the reader to explore a search space of poems built by recombining a given set of poetry lines. His book is essentially a collection of ten sonnets written using the same rhyme scheme and the same rhyme at the end of each line position. Each sonnet is set on a different page, and each page is cut up into horizontal strips, each holding a different line. By flipping the strips for the different lines, the reader can obtain a different sonnet at each view. The number of combinations is 1,014 (= 100,000,000,000,000). Queneau's book is a search space of poems and provides the reader with the means to explore it by browsing the movable lines of poetry to establish a given combination and then reading the resulting poem. The book becomes a machine for exploring the search space of potential poems. Computers make this task even easier. A few lines of code applied to a set of source poems can generate millions of potential poems – so many that it becomes almost impossible to read enough of them to make a fair appraisal of the creative effort involved. Computers can also help to address this challenge.

The size of the search space constructed by combining a set of basic units is usually very large. Queneau's book had the potential to create 1,014 sonnets. It is impossible for a person to read so many. In a way, there is also no need. The poems thus generated will not be of comparable quality; many will be of poor quality; some may be good. In a way, having so many of them available makes it possible to refine the concepts of quality for the genre. If we were to select some of the better poems from the set, how would we choose them? This is a task traditionally faced by editors trying to compile anthologies of the works of human poets. However, when the set of poems to be considered becomes as large as those created by these generators, the task is infeasible by traditional means. If we wanted to select ten sonnets out of the set of 1,014 potential sonnets, there would be no possible way to do so by traditional means. Under such circumstances, only a computational solution would be possible: an automated editor that would consider this huge search space of potential poems and select just a few to present to an interested reader. One may object that part of the value of this type of work is that it does not produce a fixed set of poems and that overlaying a machine anthologizer onto it would reduce it to a flat set of poems in the traditional sense. However, the computational solutions envisaged can be designed to be dynamic. Rather than a procedure to be run once to produce a single anthology of good poems from the set, we can hope for a procedure that can be run at any time, each time producing a different anthology while ensuring that each time the set of poems selected satisfies

some basic quality criteria. This would provide the computational means for exploring the search space, sampling from the regions where poem quality is better than elsewhere.

The present chapter describes an effort to computationally explore a search space of verse generated by recombining fragments of existing sources. By defining a set of simple metrics that can be applied to the set of generated poems, the computer can explore a huge number of poems and identify particular ones that may be worthy of more detailed perusal by an interested reader.

2 Previous work

To better understand the material presented in this chapter, it is important to review some highlights of existing work on the construction of new poems by means of recombination and on the automatic evaluation of poem quality.

2.1 Building new poems by means of recombination

The consideration of a search space of verse constructed by recombining other verse was pioneered by Queneau (1961). This approach involves building a set of sonnets, specifically designed so that their lines can be interchanged. Such an endeavor was carried out with the help of ten Spanish poets in the book *Cien mil millones de poemas* (Doce et al., 2011). Another classic approach to recombining poetry was the *Rimbaudelaires* (Oulipo, 1981), which created sonnets by deleting a substantial percentage of the words (nouns, adjectives, adverbs, verbs) from a poem by Rimbaud and replaced them by metrically matching words from poems by Baudelaire. Again, in computational terms, this generated a search space of possible poems.

Computer-driven solutions have explored the recombination of text fragments into poems in different ways, mostly in search of valuable methods for the automatic generation of poems. The *Generador de sonetos* system relies on a super-set of 15 sets of four sonnets to which it applies the computational equivalent to that of Queneau's book (Asen, 2003). Each set of sonnets provides for different options for each verse within the same rhyme scheme. This provides an even larger set of potential sonnets (allowing, as it does, for 15 different choices of rhyme scheme and four choices of verse for each position in the rhyme scheme). The approach taken for the *Rimbaudelaires* (essentially extracting a template from a given poem to be filled in with words obtained from a different source) has been applied to

generate poetry in Finnish (Toivanen, 2012) and in English (Colton, 2012). The *Po-eTryMe* system combines these two approaches (Oliveira, 2012): it breaks down poems from a corpus into lines, identifies replaceable words in the resulting lines, and then recombines lines into poems, filling in the gaps using words from other sources. The ASPERA system goes a step further, extracting lines from a corpus, stripping them down to the corresponding sequence of part-of-speech tags, and then using them to build a syntactic structure to be replaced by new words (Gervás, 2001).

A different approach arose out of considerations of n-gram models as a tool for abstracting information from a corpus of text. An n-gram model compiles how frequently sequences of words of length n appear in a given corpus. N-gram models can be used to estimate the probability of a given sequence occurring and, by sampling, to generate probable sequences of words based on that corpus. This generates a search space of poems obtained by recombining individual words according to their probability of co-occurrence. Several automated poetry generators have been built based on this approach. The RKCP cybernetic poet¹ developed by Ray Kurzweil is trained on a selection of poems by an author or authors, and it uses them to create a language model of the work of those authors (Riordan, 2003). The WASP (*Wishful Automatic Spanish Poet*) system uses evolutionary algorithms to explore a population of drafts generated from an n-gram model (Gervás, 2013). The SPAR (*Small Poem Automatic Rhymers*) system explores a similar n-gram-based search space by randomly sampling word combinations into lines, and lines into stanzas, and recombining stanzas into poems to match approved rhyme schemes (Gervás, 2017). A more refined approach involving Markov chains has been applied to the generation of lyrics (Barbieri et al., 2012).

2.2 Computational approaches to poem quality

In general terms, all these poetry generators are built upon the idea of creating new poems by means of recombination, but they do not necessarily discuss the exploration of the corresponding search space. Whereas the production of poems is based on such exploration, the published reports on the systems tend to focus on particular samples selected from the set of possible solutions. This has been identified as problematic in assessments of such systems because comparison between them would involve considering the *curation coefficient* (Colton, Wiggins,

¹ http://www.kurzweilcyberart.com/poetry/rkcp_overview.php (accessed March 6, 2020).

2012): the ratio between the number of output poems considered good enough to show as samples of system output and the total number of poems produced.

As a result of considerations of this type (Jordanous, 2011), the field of computational creativity has seen a shift in focus in recent years, from procedures for generating content to procedures for evaluating content. This has resulted in the appearance of computational solutions that do not just generate texts but actually involve procedures for the automatic appraisal of these texts (Gervás, León, 2016). In the domain of poetry, the WASP system includes a set of evaluation modules that act as a fitness function in order to judge the individuals from the population during the evolutionary process (Gervás, 2013). The SPAR system introduced specific metrics for thematic consistency and enjambment, and discussed the issue of whether they might be useful as discriminators between human-produced poems and computer-generated poems (Gervás, 2017).

2.3 Search-based creative systems

The exploration of a search space of possibilities in search of creative outputs has been characterized mathematically as a formal model of creative systems (Wiggins, 2006). The simplest possible version of this approach involves describing three different elements: the rules that *define the search space* (what is considered a potential output), the procedure for *traversing the search space* (how does the system arrive at particular samples), and the *evaluation function* (how the system attributes value to particular samples). This theoretical formulation suffers from a certain vagueness as to how the search space is defined, precisely because our experience as consumers of art, music, and literature tends to be based on either the techniques we are used to seeing employed – traversal procedures – or the established quality criteria – evaluation functions. When artists diverge from these, they move to parts of the search space not usually covered by them, and they usually encounter problems (at least in the initial stages) until the social perception of the “accepted” techniques and quality criteria expands. Examples that spring to mind are the advent of Cubism or dodecaphonic music.

3 Exploring a search space of sonnets in Spanish

The search space that is explored in this chapter is one of automatically constructed sonnets in Spanish. To take advantage of the insights provided by Wiggins’s analysis, I define the search space as the set of all combinations of Spanish

words that have a non-zero probability of occurring as a sequence in human discourse and that scan sonnets as metrically valid according to traditional rules. Considerations of grammaticality (whether the sequences can be parsed into sentences) or semantics (whether what they say makes sense) would come under the heading of evaluation functions that might be applied to particular elements in this search space. The goal of this chapter is to come up with particular traversal procedures for constructing such sonnets and to devise evaluation functions to score them such that they maximize the likelihood that human readers will consider the outputs interesting. It is important to keep in mind that such an approach might lead to results that would never have been written by a human poet. This is not a negative outcome, but rather a positive outcome of setting out to explore parts of the search space that have not yet been explored.

3.1 Traversal procedures for the search space

The obvious traversal procedure for obtaining a sonnet in Spanish is getting a Spanish poet to write one. Computationally describing the procedure that such a poet might use is beyond the current state of the art in the field of computational creativity. However, samples of their output are indeed available and may be used as evidence of where the traversal procedures in question lead when applied.

Established combinatorial procedures such as those applied by Queneau are one distinct possibility. For Spanish sonnets, two instances are available (Asen, 2003; Doce et al., 2011). Procedures based on recombining *n*-grams for Spanish have also been employed (Gervás, 2017).

In general terms, the search space of metrically valid sonnets in any given language is a comparatively very small subset of the search space of all possible text. Traversing the larger space in search of metrically valid options would be a major endeavor, which would obscure the more focused search for quality solutions within the smaller subspace. For this reason, most computational approaches rely on procedures based on recombining metrically valid lines of verse. In the hope of addressing a problem that can somehow be compared to existing solutions, I propose a traversal procedure in two stages, which first generates a set of lines of valid verse and then explores their recombination to produce sonnets.

To ensure that the outputs considered for a given traversal procedure are representative of its abilities (and not the result of a lucky run), the traversal procedure is run repeatedly to generate a set of system outputs and then sample these at random to obtain a subset of the results to examine.

3.1.1 Generating valid 11-syllable lines

Lines of verse of 11 syllables in Spanish are generated by exploring probable combinations of words. Given an n-gram model constructed from a corpus of Spanish texts, candidate lines are constructed by starting from a seed word and generating a sequence of words by successively extending it with additional words (at either end of the sequence) that have a non-zero probability of occurrence² next to the neighboring word. This procedure generates streams of plausible discourse, but it has two shortcomings for potential lines of verse: a sequence of words produced in this way does not in general satisfy the constraints on the placement of stressed syllables required of Spanish *endecasílabos*, and it does not necessarily end in a word with a potentially useful rhyme. To address these shortcomings, line candidates are generated starting from words with potentially useful rhyme (words for which there are rhyming words in the corpus) and building sequences toward the start of the verse, imposing constraints on the placement of stressed syllables to rule out invalid combinations. This produced a set of valid *endecasílabos* with rhyming potential.

As a reference corpus, a set of adventure novels in Spanish was used.³ The decision not to include any poetry in the corpus was taken to avoid the risk that any poetic quality appearing in the resulting poems was directly attributed to a loan from poems appearing in the reference texts used as seed.

The line generation procedure driven by the n-gram model obtained from this corpus was run for a set time to produce a set of 5,420 valid *endecasílabos*, corresponding to 25 different rhymes.

3.1.2 Generating sonnets by recombining lines

The objective was to establish sonnets in Spanish comprising 14 lines of 11 syllables with the rhyme scheme ABBA-ABBA-CDC-DCD. Sonnets were generated using four different procedures.

² Although it would be ideal to consider this probability absolute, it is usually computed with respect to a given corpus used as reference.

³ The set of novels includes (Spanish versions of) *Tarzan of the Apes* by Edgar R. Burroughs, *Sandokan* by Emilio Salgari, *The Jungle Book* and *The Second Jungle Book* by Rudyard Kipling, *Peter Pan* by JM Barrie, *Alice in Wonderland* and *Through the Looking Glass* by Lewis Carroll, *The Prince and the Pauper* by Mark Twain, and *The Hound of the Baskervilles* and *Study in Scarlet* by Conan Doyle.

The first one is a baseline procedure based on exhaustively exploring combinations of the available material into valid sonnets driven by the order in which the lines appear in the resources employed. The first attempt to run the baseline procedure ran for a whole night and, when it was forcefully stopped in the morning, it had produced 1,236,593 sonnets. The last sonnet has the same line as the first. This means that 12 hours of computation were insufficient to explore the full set of possible combinations for one single choice of rhymes. The set of possible permutations (because the relative order makes for different sonnets) of 25 rhymes into sets of four is 303,600. To obtain a reasonable set of poems to test, two sampling procedures were applied: random sampling over the whole set of results (*Exhaustive Explorer RS corpus*) and random sampling applied to equally sized partitions of the result set (*Exhaustive Explorer SS corpus*).

The second one applies exhaustive exploration as well but starts from a different choice of rhymes selected at random each time. This allows for a broader exploration of the search space. The results for each rhyme are sampled at random from different sectors of the search space (*Exhaustive Explorer MR corpus*). This solution also tries to combine lines into longer sentences.

The third one applies random exploration so that it jumps to a completely different section of the search space each time and applies the solution for combining lines into sentences. The results for each rhyme are sampled at random (*Random Explorer MR corpus*).

3.2 Designing discrimination functions for the search space

Given the sheer size of the set of outputs being considered, it is clear that some automated means for sorting through them are required. This is made difficult by the undeniable fact that the computational procedures available to us are to date incapable of modeling the finer sensibilities of human readers. But not all hope is lost. Any solution that rules out samples of output that may have less chances of being found interesting is a help.

To identify samples within the output that had a higher chance of being interesting, it was assumed that “interesting” samples would share basic, fundamental characteristics with poems already declared to be interesting by human judges. A set of sonnets generated by human poets was compiled to use as a reference, and a set of metrics about them was postulated that might be relevant to their interest. The metrics in question needed to be easy to compute automatically, as they had to be applied to the large sets of outputs under consideration.

3.2.1 Reference corpora of human, combinatorial, and machine generated sonnets

To ensure that the metrics in question would be significant for the kind of interest I wanted to achieve, I compared the results of the metrics when applied to a set of different sonnet corpora. The corpora include some sonnets produced by human poets, some generated in combinatorial fashion from sonnets generated by human poets, and some generated by computer poets. The hope was that this would aid in the identification of metrics that can help to discriminate sonnets created with full human intentionality from those generated in other ways.

The sets of sonnets employed to represent poems created with full human intentionality are: a corpus of 84 classic Spanish sonnets by well-known poets, a more exhaustive corpus of 727 Spanish sonnets compiled to test metrical annotation (Navarro et al., 2016), and a corpus of 122 sonnets downloaded from an Internet website.⁴ The idea behind having three different corpora is that the first one may include examples of high-quality sonnets, that the second (being more exhaustive) includes more examples of lesser quality, and that the third one is mostly made up of amateur efforts. All these corpora include sonnets generated by human poets but that vary widely in terms of quality. The working hypothesis is that, if the metrics in any way capture a semblance of quality in poetry, the averages of the values of the metric metrics over these corpora should show some significant difference.

The sets of sonnets used to represent partial human intentionality (built by randomly recombining lines written by humans for this purpose) are a set of 100 sonnets built by randomly sampling the search space generated by the sonnets in the book *Cien mil millones de poemas* (Doce et al., 2011) and a set of sonnets built by randomly sampling the search space generated by the *Generador de sonetos* system (Asen, 2003). In both of these corpora, individual verses have been written by humans, but they have been combined computationally.

The sets of sonnets used to represent computer-generated poems were the collection of 18 sonnets generated by the SPAR system for the *Festival Poetas 2017* poetry festival⁵ and the set of outputs for the baseline system described in Section 3.1.

Because the metrics needed to be quantitative, they might have been confused by significant differences in size between the poems. For this reason, all the corpora have been trimmed to ensure that they only include sonnets of 14 lines of

⁴ <https://poemas.yavendras.com/sonetos.php> (accessed March 7, 2020).

⁵ Celebrated in Matadero Madrid from May 27–29, 2017.

11 syllables. An exception has been made for the corpus of sonnets arising from the book *Cien mil millones de poemas* because the poems in question are all built from lines of 14 syllables (*alejandrinos*). This corpus has been retained to inform the differences between sonnets written by humans (the reference set for the book) and the versions of them obtained by means of recombination (the samples in the corpora). But results from these particular corpora should not be taken into account when defining discrimination functions to avoid the confusion induced by the 14-syllable line format.

3.2.2 Potential discriminative metrics

A large set of metrics was postulated and applied to the set of corpora. Metrics were considered that intuitively might reflect features that were noticed to be significantly different between human-generated poems and all the others. In each case, the mean value of the metric throughout the set of samples as well as its standard deviation were computed. Tables 1, 2, and 3 present the resulting values for the metrics throughout the various corpora.

Tab. 1: Results for content metrics for the corpora of Spanish sonnets created in human and combinatorial procedures. The metrics shown are: % OCW, percentage of open class words; % SBW, percentage of stress-bearing words; #PP1s number of first-person singular personal pronouns; #PP2s number of first-person singular personal pronouns. For each corpus, the first line shows the metric mean throughout the samples and the second line shows the standard deviation.

	% OCW	% SBW	#PP1s	#PP2s
H1 Classic sonnets	54.9	86.2	1.7	0.7
	5.9	3.5	2.1	1.2
H2 Published sonnets	55.1	86.2	1.6	0.7
	4.7	3.7	2.1	1.4
H3 Amateur sonnets	54.6	84.9	1.3	1.1
	5.2	3.7	1.7	2.0
C1 Asén sonnet generator	61.0	85.7	0.0	0.2
	2.9	2.7	0.1	0.5
C2 H. to Queneau originals	64.0	86.6	0.6	0.8

Tab. 1 (continued)

	% OCW	% SBW	#PP1s	#PP2s
	3.9	2.0	1.3	1.2
C3 H. to Queneau sampled	63.5	86.9	0.6	0.7
	3.0	2.7	0.8	0.9
M1 Matadero collection	42.7	77.7	0.3	0.1
	2.1	3.6	0.4	0.2
M2 Exhaustive Explorer RS	60.7	82.8	0.0	0.0
	1.9	2.4	0.0	0.0
M3 Exhaust Explorer SS	60.8	82.7	0.0	0.0
	2.0	2.2	0.0	0.0
M4 Exhaustive Explorer MR	67.1	90.1	0.1	0.1
	4.1	3.0	0.2	0.4
M5 Exhaustive Explorer LF	62.3	89.7	0.0	0.0
	1.1	1.1	0.0	0.0
M6 Random Explorer	61.2	88.8	0.3	0.3
	3.0	2.1	0.7	0.5

Table 1 shows values for metrics that are focused on the types of words included in the poem. It is interesting to observe that three clusters emerge with respect to the percentage of stress-bearing words in poems. Efforts involving humans group around 85%, whereas machine-generated efforts separate into two efforts around 82% (M2 and M3), three efforts around 89% (M4, M5 and M6), and one effort standing lonely at around 78% (M1). This is related to the construction procedure used for lines in most of the machine efforts. Because lines are built incrementally, subject to constraints on the placement of stressed syllables (to restrict outputs to valid *endecasílabos*) and no constraints on syntax or semantics, the procedure produces outputs that diverge from human performance. Standard deviations indicate that these differences may not be very significant. The values on the use of personal pronouns require some analysis. The values for human-produced efforts indicate that these poems oscillate between some with around three or four uses of first-person singular pronouns and some with no personal pronouns at all. This makes it difficult to use metrics of these types to discriminate. However, it is clear that machine-produced efforts tend not to

use first-person personal pronouns at all. This is understandable but indicates that there is a challenge for machine-generated poetry to address here. Values for second-person pronouns are lower for human efforts, but, again, values for machine-generated efforts are always close to zero.

The values observed in Tab. 2 indicate a number of significant differences between the various corpora. It must be said that the values for C2 and C3 are not very relevant because the format of the book avoids punctuation completely, which makes these metrics uninformative for these corpora. I will nevertheless show them here for the sake of completeness.

Tab. 2: Results for metrics on sentence distribution throughout poems for the corpora of Spanish sonnets created in human and combinatorial procedures. Metrics shown are: #sent, total number of sentences per poem (as indicated by the use of full stops); V/Sent, number of verbs per sentence; PpSent, number of punctuation signs per sentence. For each corpus, the first line shows the metric mean throughout the samples, and the second line shows the standard deviation.

	#sent	V/Sent	PpSent
H1 Classic sonnets	4.0	4.1	4.2
	1.9	3.2	3.3
H2 Published sonnets	4.2	4.2	4.1
	2.3	2.8	2.1
H3 Amateur sonnets	4.6	3.1	3.4
	2.3	2.9	2.2
C1 Asén sonnet generator	3.9	3.6	3.5
	0.6	1.4	0.7
C2 H. to Queneau originals	1.1	12.8	3.9
	0.3	4.2	3.2
C3 H. to Queneau sampled	1.2	12.9	3.7
	0.4	3.4	1.6
M1 Matadero collection	8.6	0.6	1.0
	2.5	0.5	0.0
M2 Exhaustive Explorer RS	14.0	1.0	1.0
	0.0	0.1	0.0

Tab. 2 (continued)

	#sent	V/Sent	PpSent
M3 Exhaustive Explorer SS	14.0	1.0	1.0
	0.0	0.1	0.0
M4 Exhaustive Explorer MR	14.0	0.8	1.0
	0.0	0.4	0.0
M5 Exhaustive Explorer LF	9.8	1.5	1.0
	0.4	0.5	0.0
M6 RandomExplorer	9.3	1.1	1.0
	1.3	0.3	0.0

With respect to the number of sentences (#sent), the values for M2, M3, and M4 stand at an immovable 14 because the generation procedures add a full stop at the end of each line. For the rest, a clear distinction appears between corpora of poems generated with human intervention (which stand at around three sentences per poem) and machine generated ones (which stand closer to nine). The value for the number of verbs per sentence (V/s) also shows a significant difference, with human-produced efforts at around 3.5 and machine produced efforts at around one. The value of number of punctuation signs per sentence also seems to be discriminative, with human efforts at around four and machine efforts stuck very rigidly at one (the single final full stop).

Table 3 shows values that address issues of cohesion across the lines in a poem. Metrics of this type present significant challenges because the human perception of cohesion relies on complex abstractions such as grammaticality, co-reference, and semantics. The metrics discussed here are simple first approximations intended to point out obvious differences. Further work would be required to consider the application of existing tools of syntactic analysis, co-reference resolution, semantic similarity, or topic modeling to these data. Four different aspects are considered in this first approximation: acceptable conditions at extremes of word sequences bordering on end-stopped lines, the existence of valid enjambments across lines that are not end-stopped, the repetition of words, and the existence in the poem of words that co-refer. The issues related to the perceived continuity (or not) of words across line break boundaries were addressed by considering whether the n-gram model in use for the construction processes under consideration predicted a non-zero probability for each of the cases considered (sentence end and sentence beginning for end-stopped lines, sentence

continuation for enjambed lines). The metric proved to be inadequate, predicting invalid values for the human-produced poems. This is probably due to poor coverage of the n-gram model in use (obtained from a corpus of adventure novels) for human poetry. Values on enjambment for C2 and C3 are, again, uninformative because the format in which the poems are presented does not distinguish between end-stopped lines and enjambed ones. Much the same applies to the values on enjambment for M2, M3, and M4, the procedure that directly disallows enjambment. The values for the percentage of repeated words show that this is another practice that some human poems have and others do not (oscillations between 0% and 10%). The M1 corpus stands out with a very high value (28%), probably due to the fact that the second stage of recombining stanzas into poems does not consider word repetition as a possible constraint. The co-reference metric is, again, not entirely appropriate as observed differences in value seem to overlap within the standard deviations observed. It should also be noted that the metric used (size of longest potential reference chain observed) will in most cases not be significant due to reference chains of the same gender and number being merged by the metric.

Tab. 3: Results for cohesion metrics for the corpora of Spanish sonnets created in human and combinatorial procedures. Metrics shown are: %NGvESL, the percentage of end-stopped lines that the available n-gram model recognizes as valid sentence ends; %NGvEnj, the percentage of enjambed lines that the available n-gram model recognizes as valid continuations; %RepW, the percentage of repeated words; LRefC, the length of the potential reference chain (identified as a set of words of the same number and same gender). For each corpus, the first line shows the metric mean throughout the samples, and the second line shows standard deviation.

	%NGvESL	%NGvEnj	%RepW	LRefC
H1 Classis sonnets	62.7	29.8	5.8	23.2
	27.0	21.0	4.4	6.6
H2 Published sonnets	64.1	25.6	6.4	24.0
	27.0	25.0	5.0	6.2
H3 Amateur sonnets	61.0	24.8	5.6	24.3
	27.8	21.1	5.4	6.0
C1 Asén sonnet generator	49.8	6.2	5.7	25.4
	26.1	12.3	3.5	5.3

Tab. 3 (continued)

	%NGvESL	%NGvEnj	%RepW	LRefC
C2 H. to Queneau originals	80.0	0.0	4.1	35.0
	40.0	0.0	3.1	4.5
C3 H. to Queneau sampled	76.5	0.0	3.8	32.5
	37.7	0.0	2.2	3.5
M1 Matadero collection	81.8	92.9	28.1	30.6
	12.8	23.0	6.9	6.1
M2 Exhaustive Explorer RS	57.0	0.0	4.8	39.0
	4.9	0.0	2.4	3.3
M3 Exhaust Explorer SS	57.7	0.0	4.7	38.8
	4.4	0.0	2.2	3.5
M4 Exhaustive Explorer MR	63.4	0.0	3.3	31.9
	10.6	0.0	2.1	4.6
M5 Exhaustive Explorer LF	79.5	96.4	8.8	34.9
	1.2	7.7	1.2	1.3
M6 Random Explorer	95.1	1.3	9.2	30.1
	6.8	5.3	3.3	3.7

4 Discussion

If human poets were described in terms of Wiggins's formal model, the parts of the search space of sonnets that each of them produces as output would be a function of their traversal procedures (the parts of the search space that they reach) and their evaluation function (the criterion they use to decide whether a given draft is worth showing to the world). A different issue is that different poets may use evaluation functions on different levels. Amateur poets may be content with poems that barely satisfy formal metric restrictions, seasoned poets may frown upon rhymes that are too easy, emerging poets may embrace particular rhetorical devices with a passion, traditional poets may object to tropes they are not familiar with, etc.

An important evaluation function that has not been addressed but that is fundamental to the evaluation of poetry is the measure of originality. Poems that too closely resemble poems written previously by other authors should be valued significantly lower than more original poems. This may be particularly relevant in the present context in the sense that combinations that are too similar to the original poems written by human authors (and indeed the original poems themselves) should not be considered part of the search space explored by the computational procedures.

In view of these considerations, the traversal procedures outlined above would not in themselves be comparable to human poets unless some kind of sampling driven by a good evaluation function were applied to their results.

Tab. 4: Metrics chosen to act as evaluation functions for the automated poetry generation process. For each one, the third column indicates the value of the feature chosen as a target, and the fourth column indicates the value of the feature obtained by the best scoring sample.

	Description	Target	Best sample
#sent	Number of sentences	4	7
V/Sent	Number of verbs per sentence	4	3
% OCW	Percentage of open class words	55	65
% SBW	Percentage of stress-bearing words	85	93
#PP1s	Number of first-person singular pronouns	2	1
#PP2s	Number of second-person singular pronouns	1	0
LRefC	Size of longest potential reference chain	24	39
%RepW	Percentage of repeated words	5	6
%NGvESL	Percentage of end-stopped lines validated by n-gram model	100	57
%NGvEnj	Percentage of end-stopped lines validated by n-gram model	100	0

The metrics proposed in Section 3.2.2 are only very simple approximations of the kind of evaluation that would be required to achieve this. Nevertheless, they already provide the means for selecting particular outputs from the sets of results. An example of system output has been chosen by applying a selection of the metrics that seemed to best discriminate among the studied corpora. With the values for a selection of these features expressed as a vector, the results produced by the best scoring procedure (M6 RandomExplorer) are scored by computing the cosine distance between the vector of features provided for them by the metrics and the target feature vector.

The metrics that seemed most promising for driving the automatically generated poems toward comparability with human produced poems are shown in Tab. 4. The example poem presenting the shortest distance to the target feature vector is presented in Tab. 5.

5 Conclusions

I have proposed a set of basic metrics that have the potential to discriminate between human-generated and machine-generated sonnets in Spanish. The metrics have been calibrated over a set of reference corpora including: human-generated sonnets at three different levels of proficiency, sonnets obtained by combinatorial means from source material generated by human poets, and sonnets built entirely by software by recombining words based on n-gram models extracted from a non-poetry corpus.

The corpora of human-produced sonnets have been used as a source to establish a vector of target values for the various features captured by the metrics, and an example selection procedure based on computing similarity between the vector of values obtained by the metrics for new poems and the vector of target values.

Tab. 5: Sonnet selected for having the values on the selected metrics closest to the target values abstracted from the average results for the reference corpora (see Tab. 4).

Misma hubiera sido una conjetura dime cómo es el preso es mi marido. Las hojas habían aparecido. Erguido imponente e inmóvil figura	The same would have been a conjecture tell me how the prisoner is my husband. The leaves had appeared. Upright imposing and unmoving figure.
tomado dos días pues su estatura eso es un vulgar pirata salido. Quedar pues era él el tiempo perdido yo cuando era una pequeña hendidura	Taken two days as his height that is an obsessed pirate. Staying as he was the lost time me when I was a small slit
era todo su cuerpo sudoroso. Tan sólo un instante el preso escapado. Delfines dio un mordisco cariñoso	it was all his sweaty body. Just an instant the escaped prisoner. Dolphins gave a tender bite
sabido tantas cosas ocupado sucedió una vez qué es un buen esposo. Luego el otro el primer puesto avanzado.	known so many things busy it happened once he was a good husband. Then the other one the first post advanced.

References

- de Asen M. *Generador de sonetos*. Alire, Docks, France, 2003. L.A.I.R.E. (Lecture, Art, Innovation, Recherche, Écriture). <http://motsvoir.free.fr/index.htm> (accessed May 4, 2022).
- Baldick C. *The Oxford Dictionary of Literary Terms*. Oxford: Oxford University Press, 2008.
- Barbieri G, Pachet F, Roy P, and Esposti MD. Markov Constraints for Generating Lyrics with Style. In: Raedt LD, Bessire C, Dubois D, Doherty P, Frasconi P, Heintz F, Lucas PJF, editors. *ECAI Proceedings of the 20th European Conference on Artificial Intelligence*, vol. 242 of *Frontiers in Artificial Intelligence and Applications*. Amsterdam: IOS Press, 2012: 115–120.
- Colton S, Wiggins G. Computational Creativity: The Final Frontier? In: Raedt LD, Bessire C, Dubois D, Doherty P, Frasconi P, Heintz F, Lucas PJF, editors. *ECAI Proceedings of the 20th European Conference on Artificial Intelligence*, vol. 242 of *Frontiers in Artificial Intelligence and Applications*. Amsterdam: IOS Press, 2012: 21–26.
- Cran R. “Everything is permitted”: William Burroughs’ Cut-up Novels and European Art. *Comparative American Studies: An International Journal* 2013; 11(3): 300–313.
- Doce J, Reig R, Aramburu F, IRazoki FJ, Auserón S, Adón P, Azpeitia J, Aguado M, Valero M and Molina Foix V. *Cien mil millones de poemas*. Madrid: Demipage, 2011.
- Gervás P. An Expert System for the Composition of Formal Spanish Poetry. *Journal of Knowledge-Based Systems* 2001; 14: 181–188.
- Gervás P. Evolutionary Elaboration of Daily News as a Poetic Stanza. In: IX Congreso Español de Metaheurísticas, Algoritmos Evolutivos y Bioinspirados- MAEB 2013. Madrid: Universidad Complutense de Madrid, 2013.
- Gervás P, León C. Integrating Purpose and Revision into a Computational Model of Literary Generation. In: Esposti M, Altmann E, Pachet F, editors. *Creativity and Universality in Language*. Berlin: Springer, 2016: 105–121.
- Gervás P. Template-Free Construction of Poems with Thematic Cohesion and Enjambment. In: *Computational Creativity in Language Generation (CC-NLG 2017) workshop*, International Natural Language Generation Conference, Santiago de Compostela, Spain. 2017. <https://www.doi.org/10.18653/v1/W17-3900>.
- Jordanous A. Evaluating Evaluation: Assessing Progress in Computational Creativity. In: Ventura D, Gervás P, Harrell DF, Maher ML, Pease A, Wiggins G, editors. *Proceedings of the Second International Conference on Computational Creativity*. México City: Autonomous Metropolitan University, 2011.
- Korpel MCA, Moor JC. *The Structure of Classical Hebrew Poetry: Isaiah 40–55*. Leiden: Brill, 1998.
- McDonald, R. *Shakespeare’s Late Style*. Cambridge: Cambridge University Press, 2006.
- Moreh, S. *Studies in Modern Arabic Prose and Poetry*. Leiden: Brill, 1988.
- Navarro Colorado B, Ribes Lafoz M, Sánchez N. Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation. In: Calzolari N, Choukri K, Declerck T, Moreno A, editors. *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, 23–28 May 2016. Portorož: European Language Resources Association, 2016: 4360–4364.
- Oliveira, HG. *PoeTryMe: A Versatile Platform for Poetry Generation*. In: Besold T, Kuehnberger K-U, Schorlemmer M, Smaill A, editors. *Proceedings of the ECAI 2012 Workshop on*

- Computational Creativity, Concept Invention, and General Intelligence. 2012. Osnabrück: Institute of Cognitive Science, 2012.
- Oulipo. *Atlas de Littérature Potentielle*, Collection Idées, vol. 1. Paris: Gallimard, 1981.
- Queneau R. *Cent mille milliards de poèmes*. Paris: Gallimard, 1961.
- Riordan T. The Muse is in the Software. *The New York Times*, November 24, 2003. <http://www.writing.upenn.edu/~afilreis/88v/kurzweil.html> (accessed May 4, 2022).
- Toivanen JM, Toivonen H, Valitutti A, Gross O. Corpus-based Generation of Content and Form in Poetry. In: *Proceedings of the International Conference on Computational Creativity*. Dublin, 2012: 175–179. <https://www.computationalcreativity.net/proceedings/ICCC-2012-Proceedings.pdf> (accessed May 9, 2022).
- Wiggins G. A Preliminary Framework for Description, Analysis and Comparison of Creative Systems. *Knowledge-Based Systems* 2006; 19: 449–458.
- Zurbrugg N. Dada and the Poetry of the Contemporary Avant-Garde. *Journal of European Studies* 1979; 9(33–34): 121–143.

Anne-Sophie Bories, Pablo Ruiz Fabo, and Petr Plecháč

Closing Remarks: What Was This All About?

Abstract: Centered on the main question of poetics and poeticity, this volume provides a broad overview of computational methods including motif analysis, network analysis, machine learning, and natural language processing. Without limiting ourselves to poetry, we explore the poetics of various literary productions in verse or in prose, as well as experiments towards the computational generation of poems. The volume is meant to gather a representative set of such approaches, and to offer a space for sharing perspectives, practices, and inspiring insights into the issues, old and new, being addressed by digital literary studies.

As we reach the end of this volume, one can see, or so we hope, how the various tools, approaches, and goals presented here come together to form a larger picture of what computational stylistics can bring to the literary scholar, and to inform a relevant set of methods and applications for such a purpose. Centered around the main question of poetics and poeticity, yet not limited to poetry, the contributions gathered here explore a variety of methods and techniques, ranging from motif analysis, network analysis, machine learning, and natural language processing, with the purpose of exploring the poetics of various forms of literary production, whether written or spoken, in verse or in prose, and the possibility of computationally generating literary texts, in this case poems.

There is a dizzying array of methods, approaches, and tool combinations, mirroring the diversity of hermeneutical goals. The way the contributing authors interact with each other's works is testament to how such an abundant diversity of traditions, cultures, methods, and goals offers opportunities for mutual enrichment. This is why our objective in putting together this volume has been not just to gather a representative set of such approaches – although it has been precisely that as well – but also to bring them together in one place to create an active dialogue, to share perspectives, tools, and good practices, so that we can benefit from each other.

Anne-Sophie Bories, University of Basel, Department of Languages and Literatures, French Seminar, e-mail: a.bories@unibas.ch

Pablo Ruiz Fabo, University of Strasbourg, LiLPa UR 1339, Linguistics, Language, and Speech, e-mail: ruizfabo@unistra.fr

Petr Plecháč, Institute of Czech Literature, Czech Academy of Sciences, e-mail: plechac@ucl.cas.cz

The authors in this volume present and discuss several automatic detection efforts, geared toward, e.g., sonnets (Bermúdez et al.), sonic genre (Mustazza), and enjambments (Delente), and providing an overview of this rich and dynamic field, which encompasses areas such as natural language processing, distant listening methods, and user interfaces. Semantic issues, because of their intrinsic link to questions of poetics, play a central role in this volume, with quantitative methods serving reflections on rhyming words and their influence on meanings in Victorian poetry (Houston), Melville's patching and mixing of characters (Armoza), poeticisms in Russian poetry (Vekshin et al.), and the detection of character types in German drama (Krautter et al.). Finally, the last contribution (Gervás) presents the automatic generation of poetic texts by means of computational efforts, which is something of a poetic quest in its own right, and provides insights into what poetry is.

These collected articles, besides forming a solid and up-to-date primer on computational stylistics and poetics, are also usefully representative of the various stages of research activity, from the early development of new devices (Bermúdez et al.) to the results of long-term efforts (Meister, Armoza, Delente) and reflections upon years of evolving practice (Bandry-Scubbi). Not only did it seem enlightening to both appreciate the evolution of known endeavors and to learn about new initiatives, but we also thought it useful to provide a sense of how such undertakings progressively shift their focus as computational challenges are progressively resolved, as preliminary results influence ongoing questions, and as both original and emerging hermeneutical goals are addressed.

Several chapters (Meister, Bandry-Scubbi, Delente) raise one, same important issue, albeit from different angles: the focus on slow, cautious interpretation. This angle is worth stressing, for, at a crucial time for the humanities, when our scholarly community is taking a fast methodological turn toward new quantitative and statistical analyses, we need to keep in mind the qualitative, hermeneutical benefits to be gained from these novel approaches. And the chapters presented here do just that: they put in focused efforts to apply and interpret data analyses in order to advance our understanding of texts, literary history, and genre boundaries.

The diversity of the approaches, methods, and goals described in this volume is representative of a thriving research community within the growing, dynamic field of digital literary studies and will be useful to both students and scholars looking for an overview of current trends, relevant methods, and possible results. Although the volume gives more room to written poetry, chapters addressing spoken productions (Mustazza), narrative analysis (Armoza), and drama (Krautter et al.) share the spotlight too. Moreover, there are discussions of not only poetic analysis but also the possibilities of generating poems computationally (Gervás). The chapters consider various methods, such as motif analysis (Bandry-Scubbi),

machine learning, and NLP (Bermúdez et al., Houston, Krautter et al., Vekshin et al.). The volume pays particular attention to annotation, one of the most fundamental practices in computational stylistics, and Jan Christoph Meister very usefully problematizes its hermeneutical value.

Strikingly, both Anne Bandry-Scubbi and Jan Christoph Meister, who delivered the two keynote lectures at the Plotting Poetry conference held in Nancy in 2019, reach a similar conclusion – one we fully agree with – centered on the “pleasure” and “human emotion” of our contact with literature, on the importance of a certain something that escapes quantification and operationalization, and on the fruitfulness and relevance of combining this intuitive, exclusively human approach with the different strengths of computer-aided, systematized treatments in order to achieve a truly multifocal, augmented apprehension of literature.

Anne Bandry-Scubbi writes about zooming in and zooming out, presenting us with a genuine and most relevant exploration of the question of focal length. Like us, she advocates for a focus on texts that brings together close and distant reading practices in a dialogical, back-and-forth movement to sort out canonical and non-canonical features, norm, and typicality.

As she traces her own steps back through 30 years of “corpus stylistics,” describing its evolving methods and goals, how it has navigated its growth as a young field, as well as exciting and disappointing results, and her many collaborations, she affords us a unique glimpse into how an evolutionary approach has allowed her to explore several overlapping corpora and to delineate a system within eighteenth-century British fiction. Aiming for a common definition of style, she emphasizes how computational literary scholars have a duty to use the novel possibilities of hypothesis testing to genuinely check and possibly disprove prior constructs, how they should be careful to avoid moving around numbers without interpreting them, and how, when exploring literary works, we should always keep in mind the effect that such objects always aim to have on their readers, to treat them as the devices with a purpose that they are and not as aimless, accidental configurations.

Jan Christoph Meister chooses to address poetry, phenomenon, and phenomenology together, combining theoretical reflections on epistemological issues raised by the field with a more practical exposition of the tool he and his team built to make standoff markup available barrier-free: CATMA.

He starts from the premise that the phenomenology of the aesthetic artefact on the one hand and the methodology of digital analysis and modeling of empirically observed phenomena on the other appear to be epistemological opposites: the subjectively experienced “world” that is the subject of the humanities and the arts vs. the quantifiable and objective domain that is of relevance to

the sciences. His central argument is a rebuttal of this dualistic distinction. In this spirit, he argues that, from a philosophical-historical perspective, the so-called methodological divide turns out to be a discursive trope rather than a logical necessity; that attempts to bridge the gap between hermeneutic and quantitative, formalistic methods can be traced back to well before the “digital turn,” namely to the late eighteenth century; and that it is indeed possible to conceptualize and build tools for the digital analysis of aesthetic artefacts – notably for that of literary texts – in which hermeneutic and formal, analytical methods can be productively combined. As for the annotation software CATMA, it too exemplifies how a new, *scalable* practice of digital text analysis can bring together “close” and “distant” reading as two parts of one continuum and bring an empirical perspective on poetic texts and poetry into fruitful contact with their conceptualization. This call to bring together qualitative and quantitative approaches is nothing new, but it makes our time a very exciting one to be a part of and our endeavor to combine computation and hermeneutics a laudable challenge. Indeed, and Meister stresses this, we need to address textual objects at both a holistic and an analytical level, blend systematic and intuitive reasoning routines, navigate linear and iterative processes, and, of course, integrate distant and close reading.

Helena Bermúdez Sabel, Pablo Ruiz Fabo, and Clara Martínez Cantón present something that sounds like a little rover exploring some faraway Disco planet: DISCOver. This is a well-thought-out web interface that builds on their previous work, DISCO, the Diachronic Spanish Sonnet CORpus. DISCO is not a planet but a large dataset, comprising 4,085 sonnets written between the fifteenth and nineteenth centuries, including canonical and lesser-studied authors from both Spain and Latin America, with detailed author metadata, metrics, rhyme scheme, and enjambment annotations. DISCO had already been made available on public repositories in plain text and TEI, enriched with the linked data format RDFa. Yet some audiences who could have been interested in using it had been prevented from doing so for lack of an easier navigation system. So, the authors constructed DISCOver, for which the user needs no prior knowledge of XML or linked data. The impressive technical apparatus of DISCO is both accessible and hidden in DISCOver. Users with limited technical capabilities can use it to easily view literary annotations, define sub-corpora, and discover quantitative data. They can also start with aggregated data and travel back from these to the texts from which they came. This interface has been precisely designed to bring together, not just for the team’s own benefit but for that of a larger, non-specialist audience, various levels in the understanding of a text, packing a lot of expertise into a portable virtual device, one that feels like an extension of, but certainly not a replacement for, our reading capabilities.

It is an astute tool, affording the luxury of a circular reading process: from distant to close and back, seeing the text, the rhymes, the meters, and the text again. We can spot trends and outliers, contributing to our knowledge of the sonnet in Spanish. The authors argue for similar implementations being appended to other poetry corpora interfaces and for further exploration features to be added, such as sentiment analysis or imperfect rhymes, for instance, in their desire to bring elaborate tools into the hands of anybody who is interested – literary scholars of course, but also teachers – or as a means to increase interest in poetry and its wealth of forms.

Chris Mustazza takes us to a seemingly very different area, to the machine listening of sermons, a genre with which we poetry scholars are not very familiar. His main argument is that, alongside the features of what is generally recognized and expected as a “literary” reading, which often shapes the way poets perform their own work, we must also acknowledge other sonic genres, some of them outside the conventional definition of poetry. Indeed, Mustazza shows how poetic performances can feature sonic aspects of political radio speeches, vaudeville monologues, or sermons, placing a special focus on the influence of the latter genre in his chapter. The machine listening method presented and applied to sermon-poems by James Weldon Johnson and sermons by Rev. AW Nix, is still largely a prospective one. Far from just letting an algorithm loose on a heterogeneous corpus, the author carefully studies whether, where, and how exactly the sermonic genre can be understood as a dimension of form within passages of individual poems, and how this sonic reference may interact with the lines’ content. Ultimately, Mustazza’s interest is truly in poetics as he subtly and firmly focusses his gaze on the precise interactions between the meanings brought about by a recorded reading’s sonic references and the ones to be found in the written text itself. What does it mean for a poem’s reading to stick closely to or stray from its expected “poem reading” voice? It is thus a productive dialogue between an algorithm that is *incapable* of understanding content and a perceptive reader and scholar who has been trained to decipher it but is *incapable* of abstracting his reading from the meaning of texts. This collision creates a new view of the text, one combining two aspects that the human mind alone would have been unable to dissociate.

Stepping away from the technical perspective on computational devices to question their hermeneutical uses, Éliane Delente discusses the extent to which the relationship between rhythm and meaning in versified poetry can be automated. More precisely, she examines the limitations of aligning syntactic and metrical structures, a common practice when it comes to detecting enjambments and one that the author finds unsatisfactory in several respects. Instead, she suggests analyzing the metrical expressions themselves, taking into account their

beginning, their end, and their internal consistency, as well as the processing time of such successive expressions, then linking all these observations to specific reader expectations, and the time periods and individual poets being considered. By analyzing the metrical expressions themselves, Delente instructively shows how versified poetry develops and how meaning is processed, metrical expression by metrical expression, in a dynamic of interpretative constructions and readjustments. While she is herself an advocate of corpus-based research and observations, she adds a contrasting voice to the discussion of how enjambments are automatically detected and supports giving careful thought to the constitution of homogeneous subcorpora and improving the alignment of the type and level of metrical expressions, time periods, or even individual poets so that enjambment detection can work within a more stable frame.

Working within a reasonably homogeneous corpus of nineteenth-century English poetry, Natalie Houston offers us a distant reading of rhyme. She rightly argues that the effect of rhyme should be taken into account when analyzing word frequencies in poetry, and that it is thus essential to distinguish between word frequencies at and outside the end-of-line position. The author herself explores rhyme word frequencies and discusses three methods for analyzing them: rhyme frequency ranking, effect size metrics, and the rhyme frequency ratio. With a view to better understand the conventions that shaped nineteenth-century English verse, Houston offers us a glimpse into how historical readers might have experienced rhyme's structuring force within poetic discourse. Readers might respond positively or negatively to conventional language and poetic style, depending on their aesthetic preferences. But their reading of poetry will inevitably have been marked by those conventions of poetic discourse. This examination of which words or phrases would have been perceived as typical or as unique can only be achieved by computational means, and Houston's informed analysis makes it a valuable hermeneutical journey through the question of canon and style.

As Jonathan Armoza cobbles together a model of Melville's *Moby-Dick*, he revisits Harrison Hayford's previous stylistic analysis of duplicate and vestigial settings, events, and characters in the novel. He tests this initial set of hypotheses regarding Melville's drafts to explain its apparently unnecessary redundancies by submitting the scant and heterogenous data to nonnegative matrix factorization (NMF), an efficient filtering method for patching together data from multiple, insufficient sources. Rather elegantly, the author warns us not to give blind, undue credit to his statistical results, as his entire experiment is based on the assumption that these duplicates indeed originated in the blending of different drafts. Armoza thus advocates for the forming and testing of adversarial hypotheses while conceding to himself, so to speak, that the similar findings obtained by applying

two different methods do seem convincing when addressing the characterization of Queequeg in the novel.

In their article on poeticisms and common poetic discourse, Georgy Vekshin, Egor Maximov, and Marina Lemesheva present the development of the *Russian Live Stylistic Dictionary*. In order to identify poets' self-positioning strategies, the poeticisms that say "I am a poet" in naive Russian literature, they focus on writers' social stylistic positioning. They do this by examining the contextual role determinant of the social coloring of the word and phrase, and are developing a web application to automatically determine sociocultural variations of linguistic units. From there, they speculate that there is a link between a high concentration of poeticisms and the "low artistic value" of popular poetry texts. They base their diagnosis on a poem's lack of stylistic originality, measuring the degree of uniqueness in the features of a text as signs of its quality and originality. This method also enables them to monitor changes in the use of the word and to trace the dynamics of its stylistic meaning.

In many theater traditions, characters are types, and the audience is well-used to identifying them. Benjamin Krautter, Janis Pagel, Nils Reiter, and Marcus Willand are trying to automatically detect them by linking their interpretative, literarily understood types, such as intriguer, with better operationalizable properties, such as gender, age, and social status. Their larger goal is to produce a quantitative and diachronic analysis of character types over large corpora, which would allow for the rational comparison of characters. Determining exact character types without any human interpretative effort seems out of reach, and turning to the simpler, seemingly more detectable features of gender, age, and social status is intended to help select a list of possible types within an inventory. But these apparently reasonably objective features, in turn, pose their own set of difficulties. Annotating a character's social status is relative at best, the same character being placed in different contexts depending on the play, and his or her relative status changing accordingly. Additionally, age is usually underspecified. Actually, some level of interpretation is required just to annotate the three desired features for characters. An optimal solution might be found in the use of machine learning, and the authors' efforts aim to make this possible. The authors here give us valuable insights into the interpretative efforts that go into operationalizing any material and explain how this immediate objective is linked to their broader hermeneutic goal of making a nuanced contribution to drama history.

Pablo Gervás produces poetry, but not as a poet. Rather, he produces robots who produce poems. The robot presented here works with n-gram-driven word level recombination. To address the challenge of exploring a search space of metrically valid verse – such spaces tend to be very large – his program samples it in an informed manner, driven by simple quantitative metrics about elementary

features of the poems. Capable of discriminating between human-generated poems, the recombination of human-written verse, and purely computationally generated poems, the program presented here calculates vectors for the samples being examined from the large search space of metrically valid verse. Samples can thus be selected based on a target distance to human- or machine-generated poems. This somewhat vertiginous article – there is after all no great shortage of poetry, no masses of readers hoping to obtain poetry rations by industrial means – raises one very exciting contradiction: while human poets try to set themselves apart from other poets, a robotic poet typically tries to emulate the real human thing and fails because of this, as its imitation prevents the emergence of a voice of its own. Gervás's efforts address this quirk, absolutely justifying the computer generation of poetry.

Some of the chapters share similar standpoints or methodological bases. Meister describes the “hermeneutic circle,” whereby individual data points are understood within the larger corpus context, and our understanding of the larger context shapes our grasp of individual data points. This underlies the design philosophy behind the CATMA annotation and analysis tool that he also describes in his chapter. The same promotion of a back-and-forth movement between the aggregated data and the individual observations upon which the data are based underpins the DISCOVER corpus exploration interface described in Bermúdez et al.'s chapter. Bandry-Scubbi also refers to efforts to go “back and forth between book and data” in her chapter titled “zooming in, zooming out,” echoing the same idea. As another example of shared methodologies, word frequencies are one of the raw materials on which the methods in several articles rely. Bandry-Scubbi uses measures of word overrepresentation in subcorpora, extracting specificities to tease out how female characters are depicted by female vs. male authors, as well as other topics in eighteenth-century British novels. In Houston's chapter, various measures of frequency difference for words in rhyme position vs. the rest of the poem (rhyme frequency rank and ratio, and effect size) allow her to assess poems' typicality vs. uniqueness in nineteenth-century English verse. In a related manner, Vekshin et al.'s chapter utilizes supervised classification methods that also rely on a term's frequency and discriminativeness in order to detect conventional expressions typical of naïve writing in Russian.

The reader will have understood the argument underlying the whole volume, expressed collectively by the editors and individually by the contributing authors. This argument is a refusal to set computational and traditional approaches in opposition to each other as contradictory, incompatible, or otherwise exclusive of one another. It is a plea to increase the visibility of an intrinsic feature of this novel proliferation of techniques and possibilities: all these new methods and goals are following in the footsteps of previous and ongoing methods and goals.

This is not a revolution but a leap in technology. It has not come about as a reaction to or in disagreement with traditional hermeneutics but as a welcome enhancement of our capabilities. Getting all the help we can get from computers is intended to palliate the limitations of our own human minds, not to invalidate our ability to think, analyze, or interpret. Popular fears about artificial intelligences taking over our own critical thinking do not actually reflect a real threat looming over literary studies when the computational literary studies community is so skillfully bringing together exciting hermeneutical questions and innovative processing solutions.

We hope that this book will provide a concrete overview, serve as an introduction, and allow readers to discover current trends and recent works in the field. We also hope that readers will enjoy the multifaceted vision put forward by these scholars, as their many personal approaches together provide inspiring insights into the issues, both old and new, being raised and addressed by digital literary studies.

About the Editors

Anne-Sophie Bories is an Assistant Professor at the University of Basel. She received her PhD in French Literature from University Paris 3 Sorbonne Nouvelle after being a visiting scholar at the University of California at Berkeley and at the University of Leeds. Her focus on versification and quantitative analysis is geared toward interpreting texts from the perspective of poetics and stylistics. She founded the Plotting Poetry group in Switzerland in 2017.

Pablo Ruiz Fabo has been an Associate Professor (*Maître de conférences*) in Computational Linguistics at the University of Strasbourg since 2018, after obtaining his PhD from PSL Research University Paris. He examines the contribution that language technology can make to corpus annotation and exploration, as well as related issues of evaluation. He joined Plotting Poetry's steering committee in 2019.

Petr Plecháč is head of the Versification Research Group at the Institute of Czech Literature, part of the Czech Academy of Sciences. He received PhDs in Literary Theory and Mathematical Linguistics from Palacký University Olomouc and Charles University in Prague respectively. His main areas of interest are the quantitative analysis of poetic texts and the problems of authorship recognition. He joined Plotting Poetry's steering committee in 2019.

