A Computational Stylistics of Poetry:

Distant Reading and Modeling of German and English Verse

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde eines Doktor der Philosophie (Dr. phil.) genehmigte Abhandlung

Vorgelegt von

Thomas Nikolaus Haider

aus Landshut

Hauptberichter: **Prof. Dr. Jonas Kuhn**

Mitberichterin: **Prof. Dr. Stefanie Dipper**

Mitberichter: **Prof. Dr. Winfried Menninghaus**

Tag der mündlichen Prüfung: 08.03.2022

Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart

2023

Erklärung zur Autorschaft / Statement of Authorship

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet.

I hereby declare that I have written this thesis independently and that I have not used any literature other than stated. All quotations and borrowings are marked as such and the source is clearly indicated.

Acknowledgements

I am privileged that I had the opportunity to work on the computational modeling of poetry. The following pages document my research to conceptualize and then computationally model the aesthetic, the historical, and the formal linguistic dimensions of one of the arguably highest forms of language art: poetic writing.

This endeavor would not have been possible without the continued trust and genuine interest of Winfried Menninghaus, who generously made me a member of the fairly young Max Planck Institute for Empirical Aesthetics in Frankfurt am Main, Germany. Likewise, this thesis would not be what it is without my thesis supervisor Jonas Kuhn, who supported me with well measured professional advice, methodological oversight and insightful questions. Thanks are also due to Steffen Eger for our fruitful collaborations, resulting in multiple papers, the inception of a new seminar on 'Deep Learning and Digital Humanities', and countless emails at ridiculous times of the day. I also wish to extend my gratitude to Christine Knoop for her inviting expertise regarding anything poetry, to Stefan Blohm for challenging me psycho-linguistically, to Gerrit Kentner for opening the world of phonology and syntax to me, and for being the best office mate one could wish for. Thanks go to Evgeny Kim, Roman Klinger, and Ines Schindler for their support on all things modeling emotions. Also, I'd like to thank to Dominik Schlechtweg for his enthusiasm and encouragement, Sascha Rothbart for his obsession with rhythm and poetry, and to my student assistants Debby Treziack and Gesine Fuhrmann for their annotation efforts. Also, the members of CRETA, the faculty staff at the Max Planck Institute, and the members of the IMS Stuttgart.

Finally, I am deeply grateful to my partner, Emma Galan, who stood by me through thick and thin, for her continued and impassioned endurance, and to our daughter Alma for her cuteness, and to my parents who made all this possible.

Thomas Nikolaus Haider

Frankfurt am Main

October 2022

Abstract

This doctoral thesis is about the computational modeling of stylistic variation in poetry. As 'a computational stylistics' it examines the forms, social embedding, and the aesthetic potential of literary texts by means of computational and statistical methods, ranging from simple counting over information theoretic measures to neural network models, including experiments with representation learning, transfer learning, and multi-task learning.

We built small corpora to manually annotate a number of phenomena that are relevant for poetry, such as meter, rhythm, rhyme, and also emotions and aesthetic judgements that are elicited in the reader. A strict annotation workflow allows us to better understand these phenomena, from how to conceptualize them and which problems arise when trying to annotate them on a larger scale.

Furthermore, we built large corpora to discover patterns in a wide historical, aesthetic and linguistic range, with a focus on German and English writing, encompassing public domain texts from the late 16th century up into the early 20th century. These corpora are published with metadata and reliable automatic annotation of part-of-speech tags, syllable boundaries, meter and verse measures.

This thesis contains chapters on diachronic variation, aesthetic emotions, and modeling prosody, including experiments that also investigate the interaction between them. We look at how the diction of poets in different languages changed over time, which topics and metaphors were and became popular, both as a reaction to aesthetic considerations and also the political climate of the time. We investigate which emotions are elicited in readers when they read poetry, how that relates to aesthetic judgements, how we can annotate such emotions, and

then train models to learn them. Also, we present experiments on how to annotate prosodic devices on a large scale, how well we can train computational models to predict the prosody from text, and how informative those devices are for each other.

Zusammenfassung

In dieser Dissertation geht es um die computergestützte Modellierung stilistischer Variation in der Lyrik. Als 'Computerstilistik' untersucht sie die Formen, die soziale Einbettung und das ästhetische Potential literarischer Texte mit Hilfe computergestützter und statistischer Methoden, von einfachem Zählen über informationstheoretische Maße bis hin bis zu neuronalen Netzwerken, einschließlich Experimenten mit Repräsentationslernen, Transferlernen, und Multi-Tasklernen.

Wir haben kleine Korpora erstellt um eine Reihe von Phänomenen, die für Lyrik relevant sind, manuell zu annotieren, wie Metrum, Rhythmus, Reim, aber auch Emotionen und ästhetische Urteile, die bei Leser.innen ausgelöst werden. Ein strikter Annotations-Workflow ermöglicht es uns, diese Phänomene besser zu verstehen, von der Konzeptualisierung bis hin zu den Problemen, die auftreten, wenn man versucht, sie in größerem Umfang zu annotieren.

Darüber hinaus haben wir große Korpora erstellt, um Muster in einem breiten historischen, ästhetischen und sprachlichen Spektrum zu entdecken, mit einem Schwerpunkt auf deutscher und englischer Literatur, was gemeinfreie Texte vom späten 16. bis ins frühe 20. Jahrhundert beinhaltet. Diese Korpora sind publiziert mit Metadaten und zuverlässiger automatischer Annotation zu Wortarten, Silbengrenzen, Metrum und Versmaßen.

Diese Dissertation enthält Kapitel über diachrone Variation, ästhetische Emotionen, und die Modellierung von Prosodie, welche Experimente beinhalten die ihr Zusammenspiel untersuchen. Wir untersuchen wie sich die Wortwahl der Dichter in verschiedenen Sprachen im Laufe der Zeit verändert hat, welche Themen und Metaphern populär

waren und wurden, sowohl in Reaktion auf ästhetische Erwägungen als auch auf das politische Klima der jeweiligen Zeit. Wir untersuchen welche Emotionen bei Lesern ausgelöst werden wenn sie Gedichte lesen, wie das mit ästhetischen Urteilen zusammenhängt, wie wir solche Emotionen annotieren können, und lernen dann diese Annotation mit Computermodellen. Außerdem stellen wir Experimente vor, wie man prosodische Merkmale (Metrum, Rhythmus, Reim) in großem Maßstab annotieren kann, wie gut wir computergestützte Modelle trainieren können um die Prosodie von Texten zu bestimmen, und wie informativ diese Merkmale füreinander sind.

Contents

List of Figures

List of Tables

Introduction 1

1.1 Modeling Poetry

This thesis is concerned with the computational analysis of poetry, with a focus on New High German and Modern English texts. It aims to contribute to *a computational stylistics of poetry* that examines the *forms*, *social embedding*, and the *aesthetic potential* of poetic texts by means of computational and statistical methods (see Herrmann et al. (2021) for an overview of the field).

Poetry is one of the oldest and most universal vehicles of human expression. From the earliest oral traditions to post-modern language art, poetry has captivated humans across many cultures and languages (Fabb and Halle, 2010), and continues to be recognized as one of the most refined and artistic forms of writing, known for its intricate language, imagery, and beauty (Lamping, 2016; Knoop et al., 2016). The chapters of this thesis are thus centered around a handful of questions regarding literary reading and literary history, with a spotlight on poetry: How is poetry located in a system of literary genres? How did poetry evolve over the last centuries? What do people feel when they read poetry? How does prosody work in poetry? How is prosody in poetry related to its syntax and the (aesthetic) emotions it elicits?

To address these questions, we employ a framework with the following components: (i) Manual annotation workflows that follow a scientific protocol, (ii) descriptive and predictive modeling of data (poetry is understood as data), (iii) extensively curated corpora of poetry in varying sizes, and (iv) ideas from distant reading and stylistics, which often emphasize exploration and discovery, yet are familiar with notions and hypotheses derived from literary scholarship.

Manual annotation is the process of adding labels or categories to data (e.g., annotating emotions or rhyme schemes in poetry) based on a set of predefined rules and guidelines. The manual annotation process should follow a scientific protocol to ensure consistency, accuracy, and reliability. The protocol includes defining the annotation task, the annotation guidelines, choosing an annotation tool, training annotators, annotating data, and then iteratively evaluating and improving the guidelines and the protocol itself. Inter-annotator agreement and gold standard evaluation are used to measure the quality of annotations. A manual annotation process should generate high-quality annotated data for training and evaluating machine learning models or conducting experiments more generally. Furthermore, the iterative process of improving the guidelines offers insight on the phenomena under question and how they can be formalized and operationalized (cf. Gaidys et al. (2017); Gius and Jacke (2017)).

Such an approach can be beneficial for **literary scholarship**, because as Winko (2009) notes: 'If we had features that provide distinct and accepted literature criteria, it would support literary studies in the determination of their object, their procedures, and ultimately, with their subject identity'. Having clear and accepted criteria for literary studies can help establish the parameters for literary analysis and evaluation, providing a framework for literary criticism and fostering a common understanding of literary concepts and terminology. This can aid in shaping the discipline's identity and direction, and ensure more consistent and meaningful evaluations of literary works. Manual annotation is extensively employed in chapters 5 (emotion) and 6 (prosody).

Underwood (2019, p.x,xii ff.) argues that the advances that have made large historical patterns visible have less to do with computers than with new ideas about **modeling and interpretation**. Instead of relying on isolated facts (like the length of words used by different authors (Mendenhall, 1887)), quantitative literary research now starts with social evidence related to things that interest literary readers, such as audience, genre, character, and gender (or reception, style, and period in our case). The literary meaning of those phenomena comes, in a familiar way, from historically grounded interpretive communities. Numbers (or models) are used not as an objective basis for meaning, but as a way to compare different aspects of historical records.

The systematic study and **analysis of literature** dates back to the beginnings of written literature (cf. the poetics of Aristoteles (2012, orig. 4th cent. BCE)), and traditionally has been carried out via close examination of individual texts and interpreting their cultural or historical relevance (Rommel, 2004). In fact, the analysis of poetic verse is still widely carried out by example- and theory-driven manual annotation and interpretation of experts, through so-called **close reading** (see e.g., Carper and Attridge (2020); Kiparsky (2020); Attridge (2014); Menninghaus et al. (2017); Brummett (2018)). At the same time, different schools of literary criticism have highlighted varying aspects of literature, and the history of movements in literary criticism is typically seen as a succession of conflicting ideas (Goldstone and Underwood, 2014). Such a conflict could be seen in the move from close reading to distant reading practices. However, we argue in the spirit of Ted Underwood (2019, p.xvii ff.): The discourse surrounding distant reading is not a debate like the struggle between structuralism and poststructuralism, where one approach had to be abandoned in favor of the other. Instead, it represents a novel mode of examination. Distant reading is not in opposition to close reading, but rather an extension of it at a new scale of description. Critical tradition is not to be pitted against a new technological initiative called 'digital humanities' (Underwood, 2019, p. x).

The work here draws its **interpretive framework** from structuralism,[1] such as the Russian formalists. While not necessarily agreeing about what specific elements make a literary work good or bad (or beautiful), authors like Jakobson (1960), Shlovsky (1965), or Propp (2010, orig. 1928) were united on the assumption that a work of literature (poetry) contains certain linguistic features that make it identifiable as such (features that have the 'function' of literariness or poeticness). Jakobson (1960) prominently developed a theory of communication, featuring a dedicated 'poetic function', as discussed in chapter 2, and championed so-called 'parallelisms' as the main drivers of poetic language (which are primarily investigated in chapter 6). On that discourse, also see Genette (1992) and Winko (2009). Structuralism points us to relevant features and allows us to find meaning in the inter-relatedness of features. Furthermore, in the respective places, the thesis draws from selected secondary literature, such as the literary history of Gigl (2008), or the notes of Heyse (1827) or Knörrich (2005) on meter. Finally, method criticism generally (e.g., from the discourse in computational stylistics), and balanced corpora in particular help to guide the analysis of data and the interpretation of results.

Annotations derived through manual annotation (and also raw texts) can be used in **computational models**, to allow us to reflect on the representation of knowledge and ideas, and give us 'an important new form of mediation in reading and interpretation' (Piper, 2017). A model (of data) defines a relationship between variables, and provides a mode of inquiry to study relationships rather than isolated facts (for instance, it is one thing to figure out whether two words rhyme, but it is a another thing to train a model that allows us to reflect about the underlying structure of rhyming).

Once we extract linguistic or stylistic features from a given text, our analysis shifts from the text proper to the development of a model. Moretti (2011, p.4) maintains that through the process of reduction and abstraction, the text

[1] For a discussion whether computational literary studies are in a structuralist tradition, see Gius and Jacke (2022).

is transformed into a representation consisting of its essential features and their interactions. This process results in a model that, while representing a reduced version of the original text, provides insights into the complex underlying structures (for an argument regarding reductionism in computational literary studies see Gius and Jacke (2022)). However, a pure 'data science' of literature will have a hard time to give a literary interpretation to (isolated) facts (Fish, 1980, cf.). When building models from literary data these should be adequate for the subject and the associated research questions, and they should be grounded in an interpretive framework or 'domain knowledge' (Pichler and Reiter, 2020) that ideally allows us insight both for literary scholarship, but also regarding the computational models, which may yet pose open research questions themselves (Kuhn, 2020).

For example, models are used in this work to study the boundaries and characterization of literary genres through the co-variance of lexicon-based features. Unsupervised methods from distributional semantics help us to find themes that are associated with literary periods, and for tracking the emergence of poetic tropes. A transfer-learning model can measure the relationship between aesthetic emotions and meter. Another model learns a representation of rhyming words. Recurrent neural network models learn metrical tagging and the inter-relatedness of prosodic devices through multi-task-learning.

Kuhn (2020) points out that in the digital humanities, a (computational) model typically refers to a concept within a methodology of the data-driven computer sciences (such as computational linguistics). A model is then understood as a (usually complex) algorithmic system that implements an input/output function on the computer that approximates an empirical process, or, in the case of 'data modeling', the systematic characterization and standardization of the relation of data elements to each other and their relation to represented entities. A model is a simplified representation of an aspect of reality, where the process of 'modeling' boils down to a search for a best model that most adequately maps to reality according to some qualifying criteria, on

the basis of some representative domain knowledge, which in the text sciences is most often supplied by reference corpora and expert knowledge (e.g., in form of annotations according to guidelines that define the operationalization of such domain/expert knowledge, and corpora that are representative of the underlying strata).

The **representativeness of a corpus** refers to how well it reflects the larger population and the strata of the underlying variables (w.r.t. a research question). A representative corpus should contain a sufficient amount of data that represents variables of interest and provides a comprehensive view of the phenomena being studied (Gray et al., 2017). It aims to be representative of a specific variable, while a reference corpus aims to be balanced and comprehensive. Both are important in distant reading research because they help ensure that the findings are based on a representative sample and not on a biased subset, so researchers can reduce the risk of drawing incorrect conclusions, and increase the validity and reliability of their findings. This thesis occasionally quietly assumes that a used corpus is representative, while typically operating on a canon (however large) that has survived and was digitized. Some resources that form the basis for our research, such as the large German poetry corpus, or the annotated German poetry corpus were not available at the outset, and had to be built, allowing us to take considerations about representativeness into account. Corpus representativeness is considered in this thesis, e.g., by aiming at a considerable size of the large reference corpora and ensuring that every time period is adequately represented, or through sampling and balancing of features in the smaller corpora (for example, the rhyming corpus is sampled in a way to represent a diachronically balanced cross-section of a larger corpus, or the smaller corpora for prosody and emotion are balanced in a way that highlights a variety of forms and contents across a considerable time-frame). See chapter 3 for details.

Research in **distant reading** typically makes heavy use of language corpora, and the work in this thesis readily aligns with work of that persuasion.

Distant reading was popularized by scholars such as Franco Moretti (2013) and Matthew Jockers (2013), who used the increasingly digitized literary record as departure to ask a new set of questions about literature and the literary record, in an attempt to gain a broader understanding of literature and its evolution over time. For example, Jockers and Mimno (2013) extracted themes from a corpus of 3,346 works of 19th century fiction, finding some differences in how female and male authors wrote about certain themes such as religion, war and fashion. Studying British novels between 1740 and 1850, Moretti (2009) highlighted the relationship between the length, syntax, and semantics of book titles and changes in the economic and cultural environment. He found that with a growing book market, titles became much shorter and also that titles from the beginning of the nineteenth-century reflect nineteenth-century ethics. Thus, titling practices follow the book market, but also presumably the cultural and aesthetic preferences of readers.

This research in distant reading is arguably concerned with **stylistics**. And while approaches to style and stylistic analysis have changed over time, literary traditions, and fields of study (Herrmann et al., 2015), in basic terms, style refers to the perceived distinctive manner of expression in writing or speaking. We might talk of someone writing in an 'ornate style', or speaking in a 'comic style'. For some people, as for Aristotle (or the formalists mentioned earlier), style has evaluative connotations: style can be 'good' or 'bad' (Wales, 2014). Furthermore, Wales (2014) points out that the goal of most stylistics is not only to describe the formal features of texts for their own sake, but to show their functional significance for the interpretation of the text, or to relate literary effects to linguistic 'causes' where these are felt to be relevant (see chapter 2 for background on stylistics).

Stylistic study sharpens the understanding and appreciation of literary works, including poetry, but also fiction (Milli and Bamman, 2016; Mahlberg, 2013; Sims et al., 2019; Bamman et al., 2019) and drama (Fischer and Skorinkin, 2021; Trilcke et al., 2020; Blessing et al., 2016; Reiter et al., 2018).

Furthermore, style can help (forensic linguistics) to identify authors, their personality, their gender, their demographic and social obligations (Koppel et al., 2013; Grieve et al., 2019b; Purschke and Hovy, 2019; Underwood and Sellers, 2012; Labov, 2019), style is vital for the distinction of different genres and registers (Biber and Conrad, 2019; Argamon, 2019; Haider and Palmer, 2017) and to determine the use of rhetorical features in fiction (Leech and Short, 2007) and poetry (Leech, 2014; Attridge, 2014).

Modern methods in **computational stylistics** allow us to look more closely at questions of authorship attribution and verification (Koppel et al., 2009; Grieve et al., 2019a; Grieve, 2007; Evert et al., 2017), e.g., to shed light on the disputed authorship of certain Shakespearean works (Plecháč, 2020). But also different voices in a work of literature can be identified (Lee et al., 2021; Brooke et al., 2015a), or we may analyze networks of literary characters (Agarwal et al., 2012; Elson et al., 2010; Bamman et al., 2014c). Methods have been developed for the detection of literary events (Sims et al., 2019). Futhermore, stylistics allows us to investigate the emergence of literary diction (Underwood and Sellers, 2012), or the representation of gender in fiction (Underwood et al., 2018).

This list is by no means exhaustive, and some of these papers don't necessarily start from a research question derived from literary scholarship. However, being able to e.g., automatically and reliably detect names (Bamman et al., 2019) or events (Sims et al., 2019) in literary works, builds the basis for further research that is interested in specific names or events. Furthermore, patterns that have been found through distant reading, particularly those that run counter to expectation, give incentive to go back to the text and closely examine the issue, a technique that is also employed in this thesis.

The **computational analysis of poetry** has been concerned with formal style devices such as meter (Greene et al., 2010; Agirrezabal et al., 2019; Estes and Hench, 2016; Haider, 2021), rhyme (Reddy and Knight, 2011; Haider and Kuhn, 2018), and enjambement (Ruiz et al., 2017; Baumann et al., 2018).

More recently, higher-level phenomena, including semantic coherence (Herbelot, 2014), poetic metaphors (Reinig and Rehbein, 2019; Kesarwani et al., 2017) and tracking the rise of tropes over time (Haider and Eger, 2019) (chapter 4) have been in the focus of poetry analysis. Haider et al. (2020) annotate poetry for fine-grained aesthetic emotions that poetry elicits in readers (chapter 5).

The stylistic approach to **literary quality** was continued in work in computational linguistics (Ganjigunte Ashok et al., 2013; Kao and Jurafsky, 2015, 2012), digital humanities (Koolen et al., 2020), and empirical aesthetics (Menninghaus et al., 2017), where stylistic features are used in statistical models to determine the conditions of 'quality', of success, or the emotional impact of texts. At the same time, statistical language models have found their way into literary production (Bajohr, 2021), as research on the automatic generation of poetry and fiction is already looking back at a long tradition (Bense, 1969; Gonçalo Oliveira, 2017; Manjavacas et al., 2017), and showing increasing aptitude to learn stylistic features (Wöckener et al., 2021; Belouadi and Eger, 2022). Increasingly, analytical computational tools also find their way into literary markets. Automatic spell checking and testing the thematic adequacy of book submissions are obvious use-cases for publishers (Bläsi, 2020), as well as improving consumer facing marketing with the help of recommender systems ('you like X, so you might like Y'), which are already widely used in online market places (Amazon), video streaming platforms (Netflix) and social media (Facebook). Another important application includes the optimization of book distribution logistics, i.e., how a publisher (with the help of AI) estimates the sales of a book in certain markets (Stamper-Halpin, 2019). It seems increasingly likely that publishers (followed by critics) will, or already are making use of computational models that help them assess the quality of potential publications (Althoff, 2016; Phillips, 2016), possibly leading to a form of 'digital criticism'. This will require research that examines the computational models that can be used to judge the quality of a work of literature.

The end of this introduction and the beginning of the thesis proper is marked with **a poem**. In Figure 1.1 you'll find the first stanza of Emily Dickinson's poem 'I felt a funeral in may brain', a poem that contains many of the stylistic elements that are discussed in this work. What does the poem mean? What are you feeling when you read it? How does it sound? How is it embedded in the history of poetry? What stylistic devices are employed?

```
-    +  | -      + | -      +    | -    +   |
I   felt  a     funeral     in     my brain ,

-    +  | -      + | -      +    |
And  mourners ,  to  and    fro   ,

-    +  | -      + | -      +    | -    +   |
Kept treading ,  treading , till    it seemed

-    +  | -      + | -      +    |
That sense was  breaking  through .
```

Figure 1.1: The first stanza of 'I felt a funeral in my brain' by Emily Dickinson. It was written in the 19th century, and it carries an emotional undertone of sadness and serenity (the overall lexical sentiment is negative however). Examples of alliteration (red), a (near) rhyme (blue), and meter (black/grey), with stressed syllables (+), unstressed syllables (-), and foot boundaries (|). The rhyme scheme is 'abcb'.

This text is a prime example to show how intertwined the semantic and the formal stylistic features of a poem can be.[2] We find a regular meter, rhyme, and alliteration. The meter of a poem can shape the reader's experience of the poem, creating a particular rhythm and movement that can draw the reader into the poem and make the reading experience more engaging (Obermeier et al., 2013; Menninghaus et al., 2017, 2015). In this poem, we find iambic tetrameter and iambic trimeter,[3] which should be familiar to English speakers,

[2] See Šeļa et al. (2022) for the 'semantic halo of meter'.

[3] Iamb: The recurrence of the pattern of an unstressed syllable followed by a stressed syllable (-+), either four times (tetra-) or three times (tri-). See chapter 6.

creating a sense of natural rhythm.[4] Similarly, rhyming and alliteration can create emphasis, but also unity and coherence, tying together the different parts of the poem. We find a slant rhyme[5] (fro, through), putting emphasis on spatial motion.[6] We also have the alliterations (felt, funeral) and (treading, treading, till), where the latter can be interpreted to support the sense of marching mourners.

The poem carries a strong emotional undertone, perhaps of sadness and serenity. It is about the letting go of something (using a funeral as metaphorical device) and the onset of understanding (sense was breaking through), and thus relates to feelings one can have when finishing a thesis. Arguably, this poem is about a spiritual transformation (Balacarcel, 2013, cf.): The poem illuminates the idea that death, often seen as a negative occurrence, can also serve as a transformative event that leads to growth and renewal. The solemn and serious tone of the ceremony, the rhythmic treading of footsteps, and the melodic meter evoke a sense of reverence and awe. The ambiguity and slant rhyme present in the poem serve to further emphasize the complexities of the transformative journey and the poetic truth that it reveals about the human spirit.

[4]And while Dickinson used traditional meters, fitting to the middle of the 19th century, her work is generally regarded as pioneering modernism (Dickie, 1990).

[5]The rhyming pair (fro, through) shows an important aspect of artistic license: For the sake of artistic freedom and expression, certain conventions (of rhyming) can be adjusted (e.g., that the pronunciation of words can be fit to the needs of the text).

[6]The rhyme scheme (abcb) is kept throughout the poem.

1.2 Chapter Summaries

The **core findings** of this thesis include:

We found that there are specific topics (themes/semantic fields) associated with particular literary periods/movements. For example, the topic 'world, power, lust, time' is very specific to baroque poems, or the topic 'virtue, arts' is mainly found in poems of the enlightenment age. Furthermore, certain metaphorical compositions (here called tropes) such as 'love is magic' gained traction over time (were increasingly used), and other words and aspects of their meaning changed (e.g., the German word 'billig' changed its meaning from 'appropriate' to 'cheap' since 1600 BC). These trends in the semantic evolution of poetry are shown with the help of unsupervised methods from distributional semantics, such as statistical topic models and diachronic word embeddings (cf. chapter 4).[7]

We found that the emotions that are felt when reading poetry are more subtle and nuanced than the emotions that help us navigate (everyday) life. We found that trained human expert annotators can consistently annotate feelings of Beauty, but also Sadness, Uneasiness, Energy/Vitality, Suspense, Awe/Sublime, Humor, Annoyance, and Nostalgia.[8] Furthermore, we found that crowdsourcing does not work that well in this scenario, but that a large language model (BERT) can learn these emotion from text to a certain degree, and in a transfer-learning setup we found that there appears to be a systematic relationship between poetic meter and aesthetic emotions (cf. chapter 5).

[7]Diachronic/variational aspects of formal linguistic and reception aesthetic features (such as rhythm, verse measures, and emotions) can be found in the chapters on emotions (Chapter 5) and prosody (Chapter 6), since we need a considerable amount of preparatory work to model such variables first.

[8]The concepts *Beauty* and *Awe/Sublime* primarily define object-based aesthetic virtues. Kant (2001, orig. 1790) emphasized that such virtues are typically intuitively felt rather than rationally computed. Such *feelings* of *Beauty* and *Sublime* have therefore come to be subsumed under the rubrique of *aesthetic emotions* in recent psychological research (Menninghaus et al., 2019). For this reason, we refer to the whole set of category labels as *emotions* throughout this thesis. Also, we found that a combined category Beauty/Joy is more consistent than the individual terms. Emotion terms are capitalized to highlight their categorical character.

We found that prosodic features in poetry (such as rhyme, meter, rhythm) can be reliably annotated through silent reading, when also considering the schematic and intentional structure of poems.[9] We found that it is possible to build predictive models on top of this annotation that bring the state-of-the-art to a level that allows us to robustly annotate (some of) these features large scale and look e.g., at their distribution over time. We showed that modeling explicit similarity measures between phonetically similar (or dissimilar) words allows us to gauge the extent of imperfect rhyme, and which sound deviations are responsible for these imperfections. Furthermore, in a multi-task learning setup (predicting different poetic features jointly), particular beneficial task relations illustrate the inter-dependence of poetic features. For example, we found hat caesuras are quite dependent on syntax and also integral to shaping the overall measure of the line, and that jointly learning the tasks of predicting aesthetic emotions and verse measures benefit from each other (tentatively confirming their systematic relationship) (cf. chapter 6).

The remainder of the thesis begins with a background chapter on stylistics and poetry, elaborating on what is meant with with the concepts 'style' and 'stylistics' in rhetorics, poetics, but also sociolinguistics and corpus- and computational stylistics. The chapter provides a quick introduction to the terms 'literature' and 'poetry', to finally illustrate these concepts with an experiment in genre stylistics, showing how texts vary across literary genres regarding their style. For example, we found that through modeling the variation of lexical features across the main literary genres that the two most pervasive communicative functions in literature are laid out in the dimensions of interpersonal vs. expository[10] and narration vs. non-narration (cf. chapter 2).

Chapter 3 documents the poetry corpora that were compiled for the purpose of this thesis, and that are used throughout it. A major contribution of

[9] In cases where this is more challenging, we use dis-agreements as departure to further closely investigate the texts in question, encountered e.g., in the annotation of foot boundaries, where closer examination revealed that there are ambiguous patterns at work.

[10] Drama is more interpersonal and poetry is more expository.

the thesis is the creation of large poetry corpora for German and English, and augmenting them with reliable automatic annotation, such as part-of-speech tags, syllabification, metrical syllable stress, and verse measures. We discuss how we built large poetry collections for German and English, which allow insight on 350 years of poetic writing, and small corpora that were manually annotated, in accordance to scientific protocols that aim to make this work transparent and sustainable (the protocols are described in the respective chapters where we describe the annotation process). Besides describing the corpora and giving some basic statistics and overview plots, we also illustrate the formats in which these corpora were published for the scientific community to use, and lastly also the development and evaluation of part-of-speech taggers and robust automatic syllabification systems that are both used to tag and segment the large corpora. It should be noted that we focus on (New High) German and English poetry corpora by default, and these are developed in this chapter. However, the thesis in general also incorporates resources from other languages (e.g., French, Russian, Czech) where possible.

In the end, Chapter 7 will wrap up the thesis and provide last remarks with a discussion of shortcomings of the used methods and directions for future avenues.

Publications

This thesis is mainly based on the following publications, which are grouped thematically here according to the respective chapters. The chapter on corpora (Chapter 3) does not have a dedicated publications list, as the corpora described there were used in different versions in the respective papers. The background chapter (Chapter 2) and the introduction (Chapter 1) were written exclusively for this thesis.

Diachronic Variation (Chapter 4):

Haider, T.N., 2019. `Diachronic Topics in New High German Poetry`. In Proceedings of DH2019, Utrecht. arXiv:1909.11189.
Contribution: This paper uses topic modeling to extract popular themes in German poetry, and how they align with literary periods. This is a single author paper. The research was carried out by myself, from the initial idea to writing up the paper.

Plechac, P. and Haider, T. 2020. `Mapping Topic Evolution Across Poetic Traditions`. In Proceedings of DH2020, Ottawa. arXiv:2006.15732
Contribution: This paper is an extension of the work on German, trying to compare the results to other languages. Petr Plechac was interested in reproducing the results of the above paper for other languages. I was curious if we would find the same topics in a replication study for German and whether it was possible to compare this to other languages. I provided the German and English corpora, and assisted in implementing the method that I had developed in the earlier paper. Since I do not speak Czech or Russian (the latter corpus also not being accessible to me), I could not help in translating from these languages. However, I had substantial involvement in framing and writing the paper.

Haider, T. and Eger, S., 2019. `Semantic Change and Emerging Tropes In a Large Corpus of New High German Poetry`. In Proceedings of the Historical Language Change Workshop at ACL, Florence.

Contribution: This paper was born from the idea to investigate how certain semantic fields (like topics) correlate over time, so that they would form pervasive poetic metaphors (like 'love is fire'). We propose a method for the discovery of emerging tropes in poetry (alongside a change point analysis and testing the law of linearity), using a joint model for learning diachronic embeddings, and PCA to extract the most apparent trajectories of word similarities. Steffen Eger acted in an advisory capacity, helping to iron out some methodological issues (like how to plot self-similarity, or how to solve alignment of embeddings over time steps), while other problems (like PCA for emerging tropes) were developed by myself. Writing was done mainly by myself.

Aesthetic Emotions (Chapter 5):

Haider, T., Eger, S., Kim, E., Klinger, R. and Menninghaus, W., 2020. `PO-EMO: Conceptualization, Annotation, and Modeling of Aesthetic Emotions in German and English Poetry`. In Proceedings of LREC 2020, Marseille. arXiv:2003.07723

Contribution: This paper introduces a new annotation framework for emotion in art (poetry), under a reception aesthetic premise. We evaluated different annotation techniques (experts vs. crowds), and tested how well a Transformer Language Model can learn our categories. The idea for the paper was originally mine. It won the interest of Steffen Eger, Evgeny Kim and Roman Klinger, to develop the concept and the principal narrative of the paper together. At a later point, Winfried Menninghaus assisted in writing the introduction. I was in charge of principal writing and coordinating the project.

Modeling Prosody (Chapter 6):

Haider, T. 2021, April. `Metrical Tagging in the Wild: Building and Annotating Poetry Corpora with Rhythmic Features`. In 16th Conference of the European Chapter of the Association for Computational Linguistics (pp. 3715-3725).

Contribution: This paper documents the annotation of various rhythmic features in poetry, presenting the first approach to reliably predict verse measures to enable large scale analysis, and showing the interdependence of poetic features with multi-task-learning. This paper had the longest gestation period of all mentioned here. I am the single author, but the paper is built on the annotation effort of students under my guidance.

Haider, T. and Kuhn, J., 2018. `Supervised Rhyme Detection with Siamese Recurrent Networks`. In Proceedings of the SIGHUM Workshop at COLING Santa Fe, NM 2018.

Contribution: Jonas Kuhn helped in understanding and conceptualizing the problem and the principal idea of representation learning. Writing was done by myself. Since publication of the paper, the respective section in this paper received some more attention, particularly wrt. the error analysis.

Haider, T., Trzeciak, D. and Kentner, G. 2020. *Speech Rhythm and Syntax in Poetry and Prose.* In Proceedings of the International Digital Humanities Conference in Ottawa, 2020.

Contribution: This extended conference abstract developed the initial ideas for measuring the interaction of prosody and syntax, where Gerrit Kentner proved to be a great partner for discussing the problems, and Debby Trzeciak assisted the work in training and evaluating part-of-speech taggers under my guidance. Relevant sections of his work built a basis for the 'metrical tagging in the wild' paper.

The following list includes publications that I co-authored while working on the thesis, but that are not immediately related to it (besides selected text snippets of my own contributions for the introductory chapters).

Wöckener, J., Haider, T, Miller, T., Nguyen, T., Nguyen, T., Pham, M., Belouadi, J., Eger, S., 2021. `End-to-end style-conditioned poetry generation: What does it take to learn from examples alone?`. In Proceedings of the SIGHUM Workshop at EMNLP Punta Cana, 2021.

Birte A. K. Thissen, Schlotz, W., Abel, C., Scharinger, M., Merrill, J., Haider, T., Menninghaus, W., 2021. `At the Heart of Optimal Reading Experiences: Cardiovascular Activity and Flow Experiences in Fiction Reading`. In Journal for Reading Research Quarterly.

Rother, D., Haider, T. and Eger, S. 2020. `CMCE at SemEval-2020 Task 1: Clustering on Manifolds of Contextualized Embeddings to Detect Historical Meaning Shifts`. In Proceedings of SemEval at COLING 2020, Barcelona. **Contribution**: Our contribution to the SemEval task on semantic change detection. The results for English were among the best, using mBERT and UMAP.

Haider, T., van Dyk-Hemming, A. and Eberhardt, J. 2020. `Extracting a Social Network of Musicologists`. In Proceedings of DH2020, Ottawa.

Birnbaum, D.J. Bories, A.S., Haider, T.N., and Sarv, M., 2019. `Plotting Poetry 3. Conference report`. Studia Metrica et Poetica Volume 6.2, 2019

Haider, T. and Palmer, A., 2017. `Modeling Communicative Purpose with Functional Style: Corpus and Features for German Genre`

`and Register Analysis`. In Proceedings of the First Workshop on Stylistic Variation at EMNLP, Copenhagen (also accepted at RANLP).
Contribution: This paper documents the main results from my Master's thesis, presenting a linguistically informed approach to register variation, and an evaluation of topic bias of stylistic features under unstable corpus conditions.

Background: Literature and Stylistics 2

2.1 The Foundations of Stylistics

The foundations of modern stylistics lie in the **rhetorics and poetics** of the classical world. Rhetoric requires understanding a fundamental division between *what* is communicated through language and *how* this is communicated. Aristoteles (1877, orig. 4th cent. BCE) distinguished between 'content' (what) and 'style' (how) through the terms *'logos'* (the logical content of a speech) and *'lexis'*[1] (the style and delivery of a speech). Roman authors such as Quintilian (1924, orig. 1st cent. CE) or Cicero (2013, orig. 1st cent. BCE) used the term *'elocutio'* (similarly to lexis) to refer to a manner of oral delivery which is clear and appropriate.

	English term	Meaning	Latin name	Greek name
1	discovery	finding material for arguments	Inventio	heúrisis
2	arrangement	ordering your discourse	Dispositio	taxis
3	stylisation	saying things well/persuasively	Elocutio	léxis/phrases
4	memorization	strategic remembering	Memoria	mnémē
5	delivery	presenting your ideas	Pronunciatio/Actio	hupókrisis

Table 2.1: The five canons of rhetoric. After Burke (2017).

[1]'lexis', (λεχις, 'diction', 'word'), from λεγ- (leg-, 'to speak').
Also see the term 'phrasis' (Śnieżewski, 2014, p.208)

See Table 2.1 for an overview of the five canons of rhetoric, specifying the components of persuasive acts of communication. According to these canons, an argument should be constructed through first finding material for your argument (discovery), ordering your discourse (arrangement), figure out the proper style (saying things well and convincingly) and how your argument can be best remembered (strategic memorization), to finally consider how to present and deliver your ideas (delivery).

In **rhetorics**, style (elocutio, lexis) pertains to the aspects of language—such as syntax, diction, pronunciation, or figuration—that can be used to adapt a message to be persuasive. Basically, 'good style' furthers the cause of the argument, by promoting language that is clear and better understood by the addressee. For example, Quintilian (1924, orig. 1st cent. CE) (also see Lausberg et al. (1998)) found that (the style device of) repetitions can be beneficial for the persuasive character of a message. Quintilian believed that repetition is a powerful tool in persuasion, as it can be used to reinforce a message and make it more memorable (see also the 4th canon, memorization, in Table 2.1), to create a sense of emphasis and to fix ideas more firmly in the mind, but he also warned that too much repetition can be detrimental to persuasion, as it can lead to monotony and boredom, and that it should be used sparingly and strategically (Śnieżewski, 2014, p.217).

There is now ample evidence that repetitions already exist in infant speech (Falk, 2004), but also in ritual language, songs, slogans, and poetry (Menninghaus et al., 2017). So-called parallelisms, like rhyme (two words sound similar) or rhythm (regular recurrence of pitch/duration/loudness values) can act as mnemonic devices (Assmann, 2006), such that same-sounding words and phrases are remembered better in conjunction. Jakobson (1960) suggested that parallelisms may 'foreground/emphasize' (Leech and Short, 2007; Leech, 2008) certain meaningful units (words, phrases) to directly captivate the receiver's attention and enhance understanding of the text. This also directly translates to the study of poetry, where many of these features occur.

Such aspects of linguistic style are also relevant to **poetics**. The systematic study and analysis of poetry dates back to the beginnings of written literature (cf. the poetics of Aristoteles (2012, orig. 4th cent. BCE)). Poetics is the theoretical and practical study of poetry. It was, in its beginnings and into the 18th century, a classical subject either in the context of practical philosophy or in the context of rhetoric education, and it is, at the latest since the academic establishment of literary studies in the middle of the 19th century, a subfield of these disciplines (Jung, 1997). Poetics is concerned with the question "What makes a verbal message a work of art?", specifically how literature creates meaning and effects through its formal elements. It is the study of how literature is made, how it works and how it creates meaning.

Modern poetics can be regarded an integral part of linguistics, since "linguistics is the global science of verbal structure" (Jakobson, 1960)[2] In the past, the field of poetics was about broad theoretical issues in literature (Aristoteles, 2012, orig. 4th cent. BCE), or a name for poets' reflections on their practice (Opitz, 2020, orig. 1624). Recently however, the discipline has turned more explicitly toward historical and cross-cultural questions, with a focus on empirical semiotic methods.[3] More contemporary approaches to poetics include 'cognitive poetics' (Tsur, 2008; Jacobs, 2015), looking into the mental processes of reading, or 'computational poetics', which, since poetics traditionally has a focus on the production of poetry, is more concerned with the 'poetics of computation', as in making art with the computer (Tenen, 2017; Schwartz, 2017). For the computational analysis of literature, the term 'computational stylistics' is now generally accepted (Herrmann et al., 2021).

Approaches to style and stylistic analysis have changed over time as the field of literary studies has evolved and new methods and techniques have been developed. **Approaches to style** and stylistic analysis have changed over time, literary traditions, and fields of study (Herrmann et al., 2015). In

[2]It should be noted that linguistics as a field emerged far later than poetics.
[3]Opposed to positivist hermeneutics (Gadamer, 1960).

basic terms, style refers to the perceived distinctive manner of expression in writing or speaking. We might talk of someone writing in an 'ornate style', or speaking in a 'comic style'. For some people, as for Aristotle, style has evaluative connotations: style can be 'good' or 'bad' (Wales, 2014). Furthermore, Wales (2014) points out that the goal of most stylistics is not only to describe the formal features of texts for their own sake, but to show their functional significance for the interpretation of the text, or to relate literary effects to linguistic 'causes' where these are felt to be relevant. Thus, stylistics tries to identify the linguistic devices that signal the genre, the medium, the effect, or more generally, the social embedding of a text.

A language user may choose certain sounds, words, or syntactic constructions depending on the situation in which communication is established (Jakobson, 1960; Biber and Conrad, 2019; Haider and Palmer, 2017; Simpson, 2004). The specific composition of syntactic phrases, words, and syllables may depend on whether they are written or spoken (medium), or who is speaking them to whom and to how many (participants). And also the purpose of communication (to persuade, inform, instruct, edify, etc.) may dictate a specific choice of words, just as much as a chosen literary form (poetry, fiction, drama).

In **sociolinguistics**, different conditions for communication are commonly referred to as 'variables'. Typical variables under scrutiny are demographic factors. Extra-linguistic variables such as 'age', 'gender', 'region', 'social class', or 'medium', 'purpose', 'genre' are determinable by intra-linguistic variables, like grammatical features or lexical choice. Hence, intra-linguistic variables have a certain variation across extra-linguistic variables, where the use of a combination of linguistic features facilitate and signal a particular mode of communication, or the demographic embedding of an utterance. Variables are represented by a set of shared regularities.

Consider the variables 'formality' and 'politeness': One can communicate in a polite or impolite, and also in a formal or informal manner. Typically, at

formal social occasions, in most languages, speakers will address listeners on a last-name basis, rather than addressing each other with their first names. In cases like these, in German the personal pronoun 'Sie' is preferred over 'du', or correspondingly in French the use of 'vous' instead of 'tu'. The choice of pronouns and name variants are thus a socially deterministic function of the variable 'formality'. Using the improper name or pronoun would be considered impolite and rude, and possibly lead to stigmatization.

Following the work of William Labov (1986, 2019), sociolinguists often study shifts between an informal conversational style and a (self-concious) formal "standard" way of speaking,[4], to assess dialect styles and in some cases to measure the degree of closeness of a variant to the standard/convention. In his seminal study, Labov (1986) showed that the (lack of) rhoticity in dialectal speakers may elicit **complex social biases** in listeners. He studied citizens of New York City, who variably pronounced /r/ in certain contexts, so that the words 'cart', 'pork', and 'bird' are sometimes pronounced 'r-less', like 'caht' (/kɑːt/), 'pohk' (/poək/), and 'boid' (/bɔɪd/).

When investigating diction (or linguistic style) in literature the focus is typically less on social biases, but more on considerations of **aesthetic value and/or historical context**.[5] So far, we have looked at **style as choice**, such that we choose words (wittingly or unwittingly) based on the affordance of extra-linguistic variables, given by our environment, or by a preferred option of expression. When studying literature, the focus is more often on **style as quality**, when a certain composition of words is aesthetically preferable over another choice. The extra-linguistic variable is then modeled around appraisal and appreciation, or 'liking' something, and the complex emotions, and sensible perceptions that arise when exposing oneself to the beauty of art

[4]For example w.r.t. to a agreed-upon wordlist

[5]The exception is of course research that investigates engaging literature, in the manner that socially relevant literature shapes the political bias of its readership (Pöhls, 2020).

or nature. Then, stylistic features may contribute to the aesthetic value of an object, artefact, or text, making it, in colloquial terms, 'stylish', or 'poetic'.

Roman Jakobson (1960) (also see Waugh (1980)) famously theorized about a **poetic mode of communication**, the so-called 'poetic function'. According to this *functional theory of communication*, distinctive linguistic features are not arbitrary,[6] but instead 'immediately tangible signs', based on a *markedness hierarchy of distinctive features* (some features are more marked for particular variables). He argues that a message which is modulated by the poetic function, 'directly lifts the relationship of signs and signified into conciousness'. Thus, encoding language (in poetry) is not primarily about the use of particular features, but the cognitive processes that underly those features. These marked features include any linguistic features that can be identified to carry meaning (here: reference to the world), especially beyond lexical semantics: For example, the sound of an utterance can have a certain meaning (enhancing quality).

In order to establish communication, a number of vital elements must be present. Jakobson calls these elements *factors*, and they include an *addresser* (speaker, encoder), an *adressee* (hearer, decoder), a *code* (system, langue), a *message* (semelfactive parole, discourse, text), a *context* (topic, or referent: what is talked about), and a *contact*, i.e., 'a physical channel and psychological connection between speaker and addressee. A message in the communication channel is modulated to different degrees by a set of the following *functions*:

[6] Jakobson's theory breaks with a central aspect of Ferdinand de Saussure's theory of linguistics. Saussure claimed that linguistic signs are arbitrary (meaning is only determined by linguistic context), and thus that there would be no objective relationship between signifier and signified. The meaning of a word only arises through (syntagmatic and paradigmatic) contrast with other words and not in reference to the world.

1. emotive (expressive)
2. conative (appellative)
3. metalingual (metalinguistic, 'glossing')
4. poetic (aesthetic)
5. referential (cognitive, denotative, ideational)
6. phatic (socio-pragmatic, politeness)

FIGURE 2.1: Jakobson's Communicative Functions.

Note that 'poeticness' is understood as a function here. Thus, any text may be modulated by a poetic function. While not necessarily agreeing about what specific elements make a literary work "good" or "bad", Russian Formalists like Jakobson (1960), Shlovsky (1965), or Propp (2010, orig. 1928) were (more or less) united on the assumption that a work of literature contains certain linguistic features that make it identifiable as such (features that have the 'function' of literariness), and also that we can identify these features since they are 'marked' against the background. At the same time, there are markers for 'politeness' (phatic), or emotive language.

This work emphasized what linguistics could bring to literary studies and gave rise to some of the earliest theoretical foundations of stylistics, such as foregrounding and parallelism, as well as key concepts of Russian Formalism such as "defamiliarization" (Shlovsky, 1965) or "de-automatization" (Jakobson, 1960), where it is posited that focusing on the message for its own sake shifts the attention away from everyday language towards an aesthetic appreciation of strange and unfamiliar language and fictional worlds. These concepts remain central to many contemporary stylistic studies (Wales, 2014).

2.2 Computational and Corpus Stylistics

Corpus linguistics allows us to investigate not only the stylistic features that stand out (marked features), but also the underlying (latent) features. Most notably, stylistic methods have found their way into computational literary studies, computational social science, corpus based work in psychology, work on register and genre, and on automatic style transfer, which will be sketched below. The scientific and computational approach to stylistics was already proposed in the late 19th century, when the **pioneers of quantitative text analysis** counted the length of sentences, the frequency of words, or other surface features of texts to determine the evolution of these features over time (Sherman, 1892), or their variation across authors (Mendenhall, 1887). The key idea of Mendenhall (1887) was that the writing of any author could be characterized by a unique curve expressing the relationship between word length and its relative frequency of occurrence. These characteristic curves would thus provide a basis for author attribution of anonymous texts. This early work was put on a firmer statistical basis in the early 20th century with the search for invariant properties of textual statistics (Zipf, 2016, orig. 1953). The existence of such invariants suggested the possibility that some related feature might be found that was at least invariant for any given extra-linguistic variable (such as authorship or time period), though possibly varying among different variables (Koppel et al., 2009). But for the most part, such univariant methods (that rely on the informativity of a single variable), have proved problematic, as they may not be stable, e.g., over different textual domains. Similarly, information theoretic measures of aesthetics (Birkhoff, 2013; Bense, 1969) have not found wider application.[7] In consequence of the shortcomings

[7]However, Kreuzer and Gunzenhäuser (1965), gathering some of the most influential thinkers on the topic at that time, already pointed out that "developments in the fields of [...] [computer science], statistics, modern linguistics, and other disciplines have created a situation in which the question of the possibilities and limitations of an exact literary science can be raised and pursued scientifically more decisively than before". However, they could not look back at a developed field of computer science. Instead, they used the terms

of such early methods, multivariate methods have been proposed (Burrows, 2002; Koppel et al., 2009). See Grieve (2007) for a large scale comparison of textual measurements for authorship attribution, or Burrows (1992) and Sichel (1975) for first attempts to put stylometric (word frequency) methods on a sound statistical basis.

Modern methods in stylistics allow us to look more closely at questions of **authorship attribution and verification** (Koppel et al., 2009; Grieve et al., 2019a; Grieve, 2007; Evert et al., 2017), e.g., to shed light on disputed authorship of certain Shakespearean works (Plecháč, 2020). But also different voices in a work of literature can be identified (Lee et al., 2021; Brooke et al., 2015a), or we may analyze networks of literary characters (Agarwal et al., 2012; Elson et al., 2010; Bamman et al., 2014c), or methods have been developed for the detection of literary events (Sims et al., 2019). Futhermore, stylistics allows us to investigate the emergence of literary diction (Underwood and Sellers, 2012), or the representation of gender in fiction (Underwood et al., 2018).

The combination of stylistic analysis and corpus linguistic methods is relatively recent. The notion of style is fundamentally comparative, and **corpus stylistics** helps to put it on a firmer empirical basis. An early account of a corpus stylistics is found in Semino and Short (2004). This work was mainly based on looking at contexts of words in corpus samples and counting i.a., parts-of-speech, and was further developed (mainly with frequentist methods) e.g., by McIntyre and Walker (2010, 2019) or Mahlberg (2013).

Research in psychology also has influenced corpus linguistic methods, e.g., to determine the **psychological meaning of words and texts**. This includes the association of words in terms of how concrete or abstract they are ('apples' are concrete and tangible, concepts like 'justice' or 'information' are abstract), if they evoke imagery, or at which age they are acquired by chil-

'information theory' and 'cybernetics', in the tradition of Shannon and Wiener, consequently mostly using methods grounded in entropy measures.

dren (Gilhooly and Logie, 1980; Coltheart, 1981; Köper and Im Walde, 2016). The work of Tausczik and Pennebaker (2010); Pennebaker et al. (2001) with the creation of the 'Linguistic Inquiry and Word Count' (LIWC) dictionary was influential i.a., to detect depression (Ramirez-Esparza et al., 2008; Rude et al., 2004), suicidality (Stirman and Pennebaker, 2001), dementia (Le et al., 2011), the personality of people (Schwartz et al., 2013; Plank and Hovy, 2015; Mobasher and Farzi, 2021), deception (Ott et al., 2011, 2013) or to analyse the language of dreams (Nadeau et al., 2006; Hawkins II and Boyd, 2017). Also see Boyd (2017) for a survey on psychological text analysis. Work in Natural Language processing has built on the basis of corpus work in psychology mainly from a perspective of detecting emotions from text (Bostan and Klinger, 2018; Buechel and Hahn, 2017a; Alm et al., 2005a).

The field of **computational sociolinguistics** uses large datasets e.g., from social media (Hovy et al., 2015; Plank and Hovy, 2015), to investigate linguistic characteristics of gender (Bamman et al., 2014b) and gender bias (Garimella et al., 2019), or racial bias (Kiritchenko and Mohammad, 2018; Waseem, 2016) and hate speech (Waseem and Hovy, 2016). Other research has approached linguistic differences in age groups (Schler et al., 2006; Hovy et al., 2020), dialects and geographically specific language use (Jørgensen et al., 2015; Bamman et al., 2014a; Hovy and Purschke, 2018; Shoemark et al., 2017), or prejudice against social groups (Vidgen et al., 2020).

Finally, research on **style transfer** has investigated the detection and re-writing or generation of text according to certain extra-linguistic variables such as formality and politeness (Rao and Tetreault, 2018; Fu et al., 2020), sentiment/emotion (Helbig et al., 2020), gender (Prabhumoye et al., 2018), the style of fine art (Elgammal et al., 2017), or for the obfuscation of authorship (Emmery et al., 2018, 2021). For a survey on style transfer see (Jin et al., 2020).

2.3 Literature and Poetry as Genres

Before the invention of writing systems, poetry in the form of versified stories and songs, was passed down orally over generations and later these oral literatures were written down (Beissinger, 2012; Finnegan, 2012; Höivik and Luger, 2009; Goody, 1987). Among the first written records of poetry we find Sumerian tombstones from the third millenium BCE (Schmandt-Besserat, 2015). However, writing systems (such as cuneiform) predate these tombstones by at least 1000–2000 years. Irving Finkel (2019, 6:43) claims that 'the motivation of our ancestors to develop writing systems was certainly not driven by the desire of lovelorn poets to record their low and lewd desires for posterity'. It was merchants who desired written records, and thus writing systems up to that point were used mostly for mundane tasks like accounting. The evolution of poetry and its analysis follows alongside the history of writing and the book, which can be roughly divided into three key revolutionary phases (Finkelstein, 2008): The first phase encompasses the movement from oral to written cultures, including the development of writing systems and the creation of writing tools such as clay tablets or ink and paper. Second, the shift from literacy to printing, starting with the first printing presses of Gutenberg, up to the technological advances in the industrial age that allow the distribution of books and newspapers to mass audiences. Third, the phase through which we are living right now, the move from the printed word to digital technology that allows the mass storage of artefacts in digital form and their distribution via the internet.

The meaning of the term 'literature' has changed considerably over the centuries. In fact, the modern definition of literature, restricted to imaginative writing or *belles lettres*, emerged only gradually between 1750 and 1850. Prior to that, the word 'literature' (from the Latin 'Littera' meaning 'letters') referred generally to writing or learning, as in 'being literate', the ability to

read and write (Underwood and Sellers, 2012), and the notion of 'imaginative writing' did not exist as such (Mark, 2009).

Moreover, the term 'lyric poetry'[8] (Ger. 'Lyrik') was not adopted from ancient poetics until the 16th century, and until the end of the 18th century it did not align with the collective term in use today (Knörrich, 2005). The triadic distinction of literary genres into 'Epic' (Epik), 'Drama' (Dramatik), and 'Lyric Poetry' (Lyrik) is largely attributed to Johann Wolfgang von Goethe (1749–1832) (Knörrich, 2005), who, both as an artist and scholar, substantially influenced the subsequent discourse on literary writing, shaping our modern notion of it. According to Steele (2012), for a considerable amount of time artistic writing was actually constrained exclusively to verse, and prose was reserved for writing of 'non-artistic' form (like chronicles). Consequently, it was debated if 'non-versified' language was even capable as a means of artistic expression. Moreover, rhetorics and poetics differentiated only between prose and verse.

Whether lyric poetry constitutes an independent genre and, if so, how it is systematically located in a system of text genres only arose under the influence of the new discovery of Aristotle's Poetics around the middle of the 16th century in Italy (Hempfer, 2008). From the Renaissance to the 18th century, all production of lyric poetry was still firmly bound to the traditions of ancient Greek and Roman lyric poetry[9] in terms of genres, themes and stylistic features. It was not until the mid-18th century that there was a growing tendency to orientate oneself towards genre and style models of non-ancient origin, such as folk poetry (Lamping, 2016, p.336).

[8] I will use the terms 'lyric poetry' and the German term 'Lyrik' interchangeably. The English term 'poetry' encompasses a wider concept, from ancient Greek 'poein' (to make), and most often only refers to 'verse', or 'versified language', and thus more often than not also means 'epic poetry' and 'dramatic poetry', whereas the term 'Lyrik' or 'lyric' originates most likely from the musical instrument 'lyra'. While this thesis has a focus on lyric poetry, it does not systematically exclude other verse forms. But when 'lyric poetry' is mentioned, 'Lyrik' is meant.

[9] And partly to Italian Renaissance lyrics, according to (Lamping, 2016, p.336).

2.4 An Experiment on Style across Literary Genres

This section is intended to give a first overview to corpus stylistic methodology with an experiment to characterize literary genres through stylistic features. We propose a simple but effective method to investigate topological linguistic style variation over prototypically prescribed genre categories. The co-variance of an assortment of lexicon-based text features enables us to study the boundaries and characterization of literary genres with an unsupervised method. On the one hand, our results support the view that literary genres are not monolithic with regard to their linguistic features, but that the prevalence of certain features in particular genres enables us to study genre edge cases, and also which linguistic features are prevalent in belles lettres versus other text genres, i.e., which features allow us to pinpoint aesthetic literature.

As we have seen in the previous section, a universal definition of 'genre' is elusive. Here, we have to rely on conventional genre labels, as they are tagged in a corpus. A Principal Component Analysis (PCA) allows us to locate documents within a topological feature space, illustrating more a notion of 'register', rather than a typology of genre (Sharoff, 2018, cf.). The terms register and genre have been central to previous investigations of discourse and textual variation. Both terms have been used to refer to language variety associated with particular situations of use and, lacking a clear differentiation between the two terms, many studies simply adopt one and disregard the other (Biber et al., 2007). However, register, according to Lee (2001) and Biber and Conrad (2019), is understood as (stylistic) variation according to use in broad societal situations. It describes a functional adaptation of language to the immediate situational parameters of contextual use, as different situations require appropriate configurations of language. On the other hand, genre views text by consensus within a culture, as artifacts categorized by purposive

goals, distinguished by conventionally recognized criteria and hence subject to change as conventions are challenged and revised over time. In short, genre is described by a conventional label, while register is described through its pervasive features (Biber and Conrad, 2019; Haider and Palmer, 2017).

For analyzing genre/register in categorical terms, text classification techniques (Sebastiani, 2002; Devlin et al., 2019), usually based on machine learning, are frequently used. The idea is straightforward: training texts are represented as numerical vectors, labeled by their genre categories, and machine learning methods are used to find a function that distinguishes between the categories that minimizes some loss function over the training set. Different algorithms will produce different results, with greater or lesser ability to generalize accurately to new data (not in the training set).

In multidimensional analysis, the goal is to find the 'natural' dimensions of variation among core grammatical features of the language. Principal Components Analysis (PCA) or Factor Analysis (FA) is typically used to compute the sets of linguistic features that most frequently co-occur in a corpus. These are called the dimensions of variation for the corpus. Numeric weights are computed for features in each dimension, enabling computation of a score for any text in a given dimension. Analysis of which features co-vary in each dimension and the relationships between the dimensional scores for different texts or registers enables a linguistic interpretation of how aspects of register variation are represented by the different dimensions.

The procedure of applying PCA to features of (literary) text genres has been used in previous investigations, such as by Passonneau et al. (2014), Laippala et al. (2021), or Schöch (2016), but never before has it been applied to the three principal genres of literature, which we will show is very straightforward. Also, the resulting visualization is better interpretable than these previous approaches.

Using this methodology, Biber (1989) identified seven factors of variation related to register, which will be listed below to put the results into con-

text. The generality of the result has been supported by the fact that similar studies on other corpora give substantially the same factors (Xiao, 2009; Passonneau et al., 2014; Clarke and Grieve, 2017), though factors differ somewhat in saliency across different corpora, depending on the exact mix of registers and genres present.

We are interested in two research questions: (1) Which features characterize the respective genres and (2) whether we can find modes of communication that are spanned by the genres of literature. Regarding the second research question: Within rhetorical theory, four basic 'modes' of discourse are traditionally distinguished: narration, description, exposition, and argumentation (Smith, 2003; Biber, 1989). García-Berrio (2016) finds that literary text is composed of narration and exposition. Biber (1989) studied textual variation with a factor analysis (which is similar to PCA) in the Lancaster-Oslo-Bergen Corpus, and found that texts vary along five register dimensions. In this study, he clusters features according to their covariance to find underlying dimensions. As can be seen in Figure 2.2, those dimensions are bipolar. In this research, we will look in particular at the first two dimensions, (1) 'Involved vs. Informational' (where 'informational' can be also understood as 'exposition'), and (2) a dimension that distinguishes between narration and non-narration.

1. Involved vs. Informational Production
2. Narrative vs. Non-Narrative Concerns
3. Elaborated vs. Situation-dependent Reference
4. Overt Expression of Persuasion
5. Abstract vs. Non-Abstract Style

FIGURE 2.2: Textual Register Dimensions After Biber (1989)

Considering the first research question, the obvious distinguishing factors between the literary genres is their overall visual form: Fiction is typically written in running prose, poetry in verse, and drama features speaker roles. However, these features are not considered here. Regarding their linguistic style, we have the following hypotheses: **Fiction** is dominated by a narrative

tone (mode of communication), where a narrator tells a story mainly in the third or first person (what's happening to the people in the story, possibly with the narrator being part of the story), describing their interactions and also their dialogue. **Drama** is centered around conversation (mode of communication), where protagonists address each other in the second person (but also referring to themselves and others). Drama is 'scripted', where (more or less) spontaneous speech is prescribed by a textual 'script', which actors memorize and perform on stage, typically in the form of dialogue (talking to other protagonists) or monologue (talking to oneself or the audience directly). Besides the actual words of dialogue, the text of a Drama also includes stage directions (instructional language that suggests a certain manner of conduct), and the proper names of the protagonists, to identify who's turn it is and what they should articulate. Since Drama is composed of spoken language, it likely entertains shorter sentences, more interjections, and more familiar and concrete language. A Drama may be typeset either in (modern) colloquial prosaic form, or in verse form, e.g., in blank verse or in alexandrines. **Poetry** is largely expository and typically dominated by a nominal and adjectival style. We hypothesize that the type-token-ratio (lexical density) is high, and that most poetry is written in the first person (lyrisches Ich, 'lyrical I'). Furthermore, poetic language should be more emotional and evoke imagery more frequently (Kao and Jurafsky, 2012, 2015).

2.4.1 Corpus

Genre corpora are faced with the problem of finding an operationalizable definition for each genre and avoiding meaningless miscellaneous categories, i.e., choosing the right granularity of classes. The multitude of possible genre categories makes it impractical to determine a fixed set of classes for a corpus that is representative for all genre.

# documents	Genre
1610	Non-Fiction
820	Poetry
550	Drama
405	Fiction
178	Biography
120	Oratory
56	Letters
28	Miscellany
10	Juvenilia
1	french

Table 2.2: Number of Documents per Genre in the TedJDH corpus.

For this experiment, we use a corpus that was introduced in Underwood and Sellers (2012). It is a collection of 4,275 documents of different text genres in the English language, covering a time period from ca. 1700–1900 CE. See Table 2.2 for an overview of the number of documents per genre in the corpus. Here, we focus on the literary genres 'Poetry', 'Drama' and 'Fiction'. Later, we also look at all genres, particularly non-fiction. Prose introductions and notes were removed from the poetry documents, leaving only the verse. This was done with a heuristic relying on the density of line-initial capitalization to identify verse.

2.4.2 Experiments

2.4.2.1 Method: Principal Component Analysis

The statistical method applied here is Principal component analysis (PCA). PCA is a dimensionality reduction technique often used to reduce the dimensionality of large data sets by transforming a large set of variables into a smaller set that contains most of the information contained in the larger set. Dimensionality is reduced through exploiting the co-variance of features. The number of principal components is less than or equal to the number of original variables. An orthogonal transformation is employed (the feature space

is rotated) that aligns correlated (co-variant) features, to identify principal components which are linear combinations of the original variables. Features that typically occur with each other e.g., in documents, will load into the same principal components. As there are as many principal components as there are variables in the data, principal components are constructed in such a manner that the first principal component accounts for the largest possible variance in the data set. The second principal component is calculated in the same way, with the condition that it is uncorrelated with (i.e., perpendicular to) the first principal component and that it accounts for the next highest variance.

However, PCA is very sensitive to the relative scaling of the original variables. We use the StandardScaler of sklearn that removes the mean for all features and then centers each feature individually on zero with unit variance.

Then, once the data is standardized, principal components are computed to yield normalized eigenvectors of the covariance matrix of features. Each component can then be interpreted as the variance of the samples when projected onto the component. These components map instances (documents) as well as the individual features. Through that, we can examine the feature loadings in tandem with the variation across documents.

2.4.3 Features

We extract the following features from the documents and count them to use in the PCA. Most features are based on lexicons, which are used to match words in documents to annotated categories. Not all features are used in all configurations, but we intend to give a broad overview which lexicons are typically used in computational stylistics studies. The core of the thesis will not use lexicons.

2.4.3.1 Surface Cues

This is a featuregroup of linguistic surface cues.

1. Avg. word length in # of characters.
2. Avg. sentence length in # of words.
3. Type-Token-ratio: The ratio of unique types and tokens. Always between 0 and 1. A high type/token ratio signals a high lexical density with few repetitions and vice versa.
4. Allcaps_ratio: The ratio of fully capitalized words (WORD) versus non-fully capitalized words (Word, word). Should show speaker roles in drama.
5. Alliteration: If two subsequent words start with the same character. Function words (from a stoplist) are skipped.
6. Assonance: If two subsequent words contain the same vowel (aeuoiäüö). Function words (from a stoplist) are skipped.

2.4.3.2 Personal Pronouns

Personal pronouns are grouped into the three grammatical persons.

First person pronouns: I, me, myself, my, mine, we, us, ourselves, our, ours

Second person pronouns: you, yourself, yourselves, your, yours

Third person pronouns: he, she, him, her, himself, herself, his, hers, they, them, themselves, theirs

2.4.3.3 Part-of-Speech Tags

We use the NLTK 'averaged perceptron' part-of-speech tagger. This might not be optimal for the poetry portion of the corpus, as the English tagger in section 3.8 was not of great quality. However, there should still be a sufficient signal for genre classification. The part-of-speech tagset of the Penn Treebank consists of 36 tags. See `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html` for an overview of the classes.

2.4.3.4 Verb Classes

Santini (2005) compiled lists for six verb classes in English. In total, these cover 296 unique verbs where four of those occur in two classes, totaling 300 tokens.

1. Activity_verbs
2. Communication_verbs
3. Mental_verbs
4. Causative_verbs
5. Existence_Verbs
6. Aspectual_verbs

2.4.3.5 MRC Dataset Dimensions

The MRC dataset, as described in Gilhooly and Logie (1980) offers word norms on four psycholinguistic dimensions.

1. Imagery
2. Concreteness
3. Familiarity
4. Age of Acquisition.

Gilhooly and Logie (1980) selected 1,934 words (nouns) that—as they assume—are evenly distributed across all texts. They put together a booklet where they shuffled the order of words (to avoid priming raters), and have raters judge these words on the above dimensions with a 1–7 Likert scale, except for age of acquisition, which uses a different scale. Imagery measures the ability of a word to evoke a picture in the recipient. Concreteness measures the tendency of a word to be a tangible object ('apple'), in opposition to abstract concepts ('justice'). Imagery and concreteness are highly correlated (pearson r = 80%)). Familiarity measures the degree to which a word is familiar to most users of the English language. And finally, age of acquisition indicates the age at which children learn this word.

2.4.3.6 NRC Color Associations

The NRC color dataset is described in Mohammad (2011a) and Mohammad (2011b). It contains 14182 unique word lemma. Per word it includes an associated color. The assigned colors are **black, blue, brown, green, grey, orange, pink, purple, red, white, yellow** or None. We filter None occurrences. Color terms in expressionist poetry are discussed in Reinig and Rehbein (2019).

2.4.3.7 NRC Emotions

The NRC emotions dataset includes 10 emotion types that are each assigned to 14182 tokens (a token may have multiple emotion labels). The dataset is described in Mohammad and Turney (2013a)]. For each token that we encounter in a text document, we count the associated emotions and normalize the final count for each emotion by the token length of the document. The ten emotion types are the following: **anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise, trust**.

2.4.3.8 LIWC - word norms

The English Linguistic Inquiry and Word Count (Tausczik and Pennebaker, 2010; Pennebaker et al., 2001) contains 6400 words and and word stems (to include all conjugations and declinations). The lexicon provides a hierarchical annotation of those words with 68 linguistic and psychological categories, e.g., the word 'cried' is part of five categories: sadness, negative emotion, overall affect, verbs and past focus. In this case, we count all five categories for a document. As seen in this example, many of the LIWC categories are arranged hierarchically. All sadness words, by definition, belong to the broader "negative emotion" category, as well as the "overall affect words" category. Note also that word stems can be captured with a wildcard (asterisk). For example, the English dictionary includes the stem hungr* which allows for any

target word that matches the first five letters to be counted (including hungry, hungrier, hungriest).

2.4.4 Results

In the following, we calculate and visualize the first two PCA components over different featuresets. Featuresets are normalized within themselves (emotions are normalized with emotions, verb types are normalized by all verb types, etc.). For each feature set, we first show a plot with the first two PCA components with the instances, and then the respective feature loadings. In Figure 2.3 we see the two first components of a PCA, spanned by part-of-speech features only, of which the loadings are presented in Figure 2.4. Also compare this to Figures 2.5 and 2.6, which includes an extended feature set but results in a similar co-variance space. Only using part-of-speech features already allows us to distinguish clusters of the three literary genres. Drama (yellow) loads to the positive dimensions of both components to the top right, poetry (red) loads to the negative dimension of component 1 and the positive dimension of component 2, while fiction (green) load to the negative dimension of component 2, but is evenly distributed across component 1. Still, some documents from the respective genres load into non-prototypical dimensions, where some poetry documents load into drama, and vice versa, or some fiction books are found in poetry, but only few fiction documents are found among drama.

We can identify two components as posited by Biber (1989) and García-Berrio (2016), (1) **Interpersonal** (Involved vs. Informational) and (2) **Narration** (Narrative vs. Non-Narrative). The characterization of these components is lined out on the following pages.

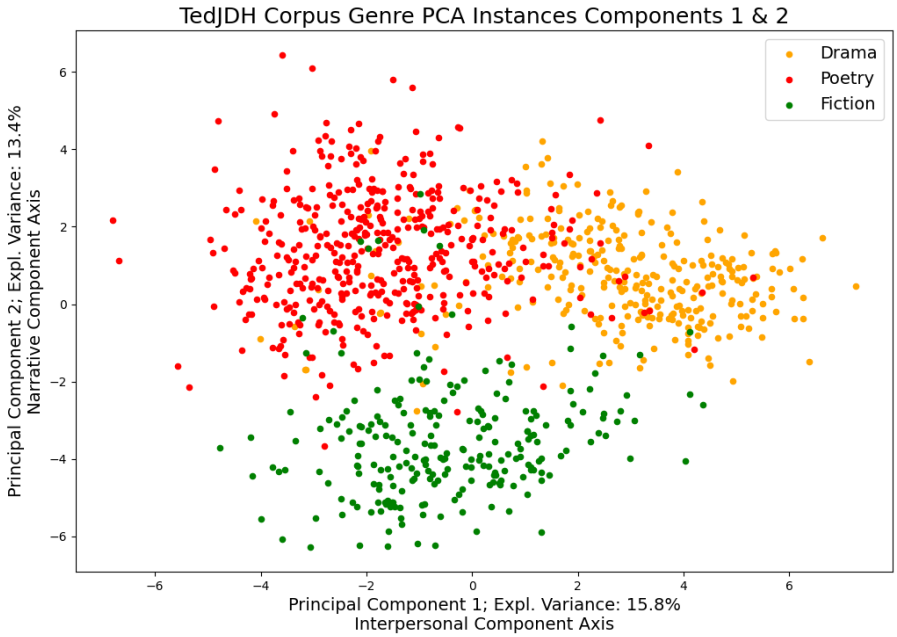

FIGURE 2.3: Instances of Three Genres of Literature in PCA, POS Features only

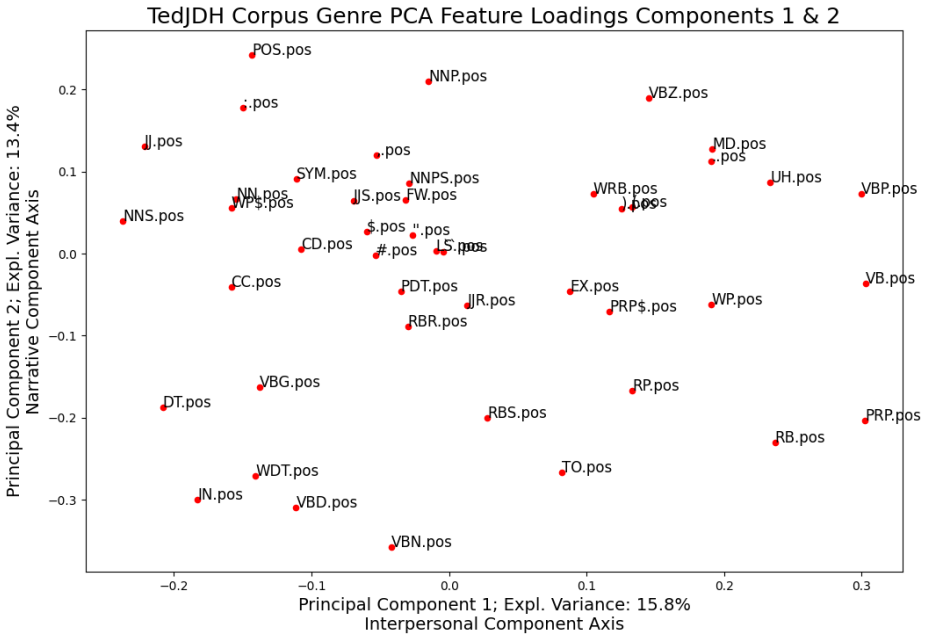

FIGURE 2.4: Feature Loadings of Three Genres of Literature in PCA, POS Features only

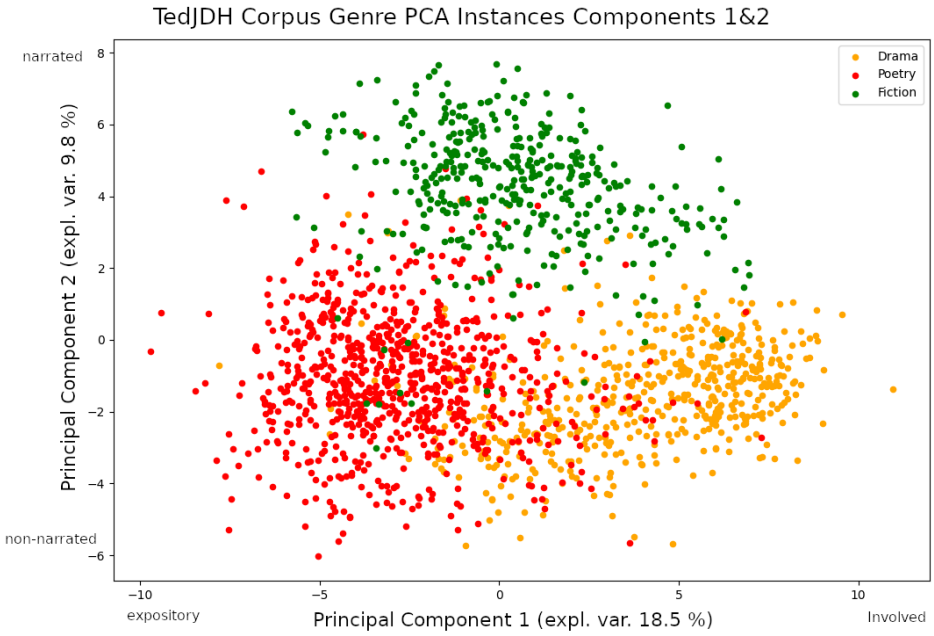

Figure 2.5: Instances of Three Genres of Literature in PCA, Selected Feature Set

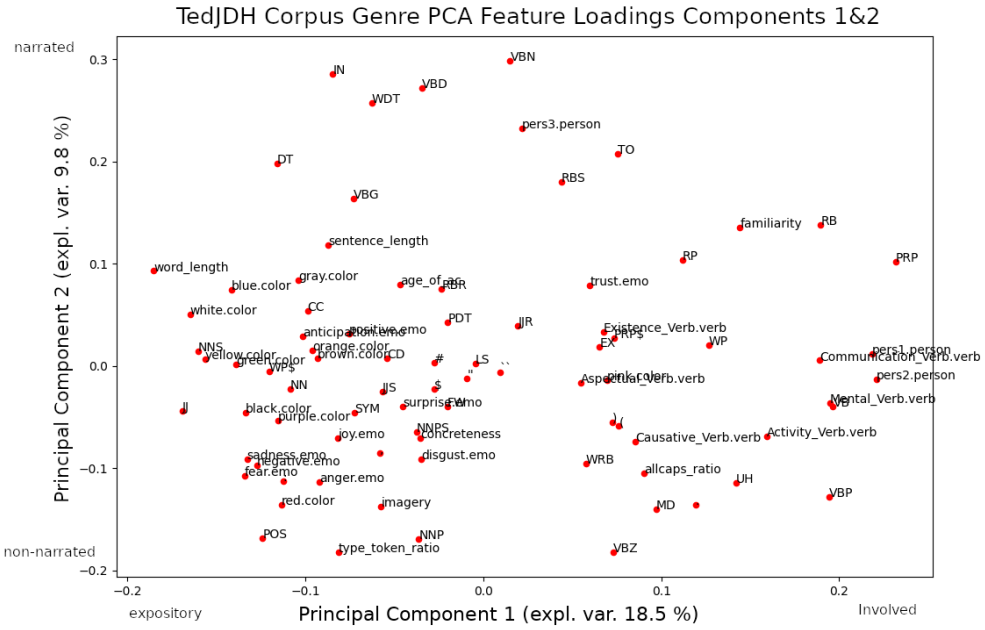

Figure 2.6: Feature Loadings of Three Genres of Literature in PCA, Selected Feature Set

2.4.4.1 Component 1 (Interpersonal)

Component 1 spans an 'interpersonal' dimension that distinguishes involved vs. informational (expository) language.

Involved language: Involved language is found on the side of drama, where the most salient features are:

1. VBP (Verb, non-3rd person singular present)
2. VB (Verb, base form)
3. PRP (Personal pronoun)
4. UH (Interjection)
5. MD (Modal verbs)
6. WP (Wh-Pronoun)
7. RB (Adverb)
8. PRP$ (Possessive pronoun)

FIGURE 2.7: POS Features for Involved (interpersonal) Language

1. 1st Person Personal Pronouns
2. 2nd Person Personal Pronouns
3. Communication Verbs, Mental Verbs, Activity Verbs
4. Familiarity
5. Allcaps Ratio
6. low word length (loads opposite)
7. low sentence length (loads opposite)
8. low age of acquisition (loads opposite)
9. low type-token ratio (loads opposite)

FIGURE 2.8: Other Features for Involved (interpersonal) Language

The Involved dimension lines up with Biber (1989) in adverbs, present tense verbs, modals, questions, 1st person pronouns, and 2nd person pronouns. Furthermore, the lexicon of involved language is familiar to the reader. All this points to language that directly addresses listeners or talks about other people in direct speech. The language is decidedly simple with short sentences and short words, and implements some redundancy (low type-token), and words are used that are understood by more people (low age of acquisition). Furthermore, we find verbs of communication, of mental processes and activity, all pointing to direct and involved language, focused on interpersonal

communication. We also see use of words that are associated with the Trust emotion, indicating language that encourages trust.

Expository (informational) language: On the other side of the spectrum, expository language is associated with poetry (and also operative text like instructions or advertising (Haider and Palmer, 2017)). We find long words, long sentences, a high type-token ratio, a high age of acquisition, many nouns, adjectives and also conjunctions. This points to elaborated language and extensive exposition of information. Thus, poetry is quite dense in transporting information and, as seen with the loading of imagery and concreteness, does this to evoke images with tangible language. The higher density of emotion and color association is likely more typical of poetry, and less of a bona fide expository register, since instructional language is also expository, but not emotional.

1. high word length
2. high sentence length
3. NNS (Noun, plural)
4. JJ (Adjective)
5. POS (Possessive ending)
6. WP$ (Possessive wh-pronoun)
7. NN (Noun, singular or mass)
8. CC (Coordination)
9. high age of acquisition
10. SYM (Symbol)
11. high type-token ratio
12. imagery, concreteness

Figure 2.9: Features for Expository Language

2.4.4.2 Component 2 (Narration)

Component 2 spans a 'narration' dimension that distinguishes narrated vs. non-narrated language.

Narrated language: Narrated language is found on the side of fiction, where the most salient features are found as listed in Figure List 2.10. Prototypical narration stands out through the use of the past tense and the third person (Verb past). We also find quite long sentences, and consequently many commas and prepositions/subordinating conjunctions.

1. VBN (Verb, past participle)
2. VBD (Verb, past tense)
3. WDT (Wh-determiner)
4. IN (Preposition or subord. conj.)
5. DT (Determiner)
6. 3rd Person Pronouns
7. TO (to)
8. VBG (Verb, gerund or past participle)
9. RBS (Adverb, superlative)
10. sentence length (long sentences)
11. familiarity (familiar words)

FIGURE 2.10: Features for Narrated Language

Non-Narrated language: Opposite we find features for non-narrated language, as seen in Figure 2.11. Non-narrated language stands out through 3rd person singular verbs, modal verbs, symbols, adjectives and interjections.

1. high type-token ratio
2. NNP (Proper noun, singular)
3. VBZ (Verb, 3rd person singular present)
4. POS (Possessive ending)
5. MD (Modal Verb)
6. WRB (Wh-adverb)
7. SYM (Symbol)
8. JJ (Adjective)
9. UH (Interjection)

FIGURE 2.11: Features for Non-Narrated Language

2.4.4.3 LIWC PCA

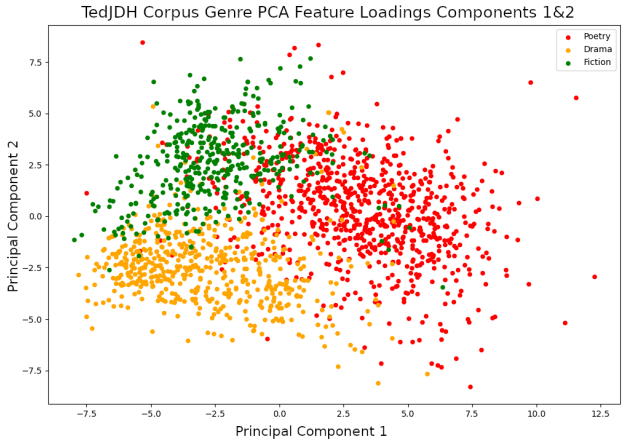

Figure 2.12: Instances of Three Genres of Literature in PCA, LIWC Feature Set

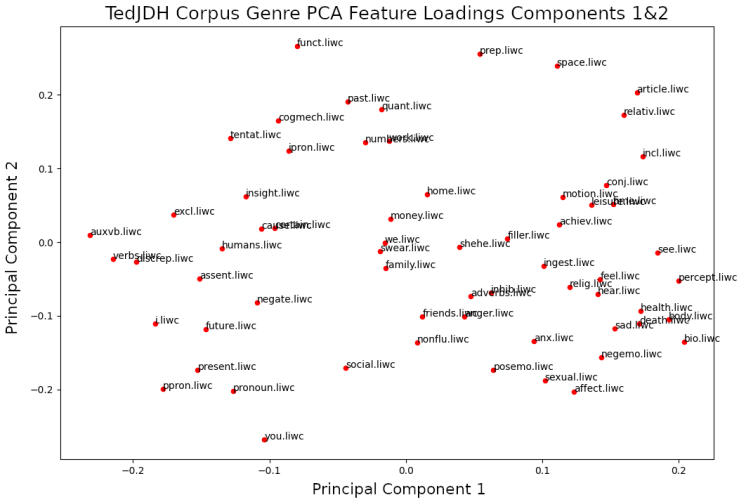

Figure 2.13: Feature Loadings of Three Genres of Literature in PCA, LIWC Feature Set

Figures 2.12 and 2.13 show PCA plots based on only LIWC features. We can see that these features also can separate the genres, but they are not as good in spanning register dimensions as we have analysed before with part-of-speech tags. However, with LIWC features we can see some topical trends,

e.g., that poetry is more explicitly sexual or about religion and friends, and more about perception with words in categories 'hear' and 'see'. Fiction uses more tentative words, and the language in drama is makes reference to the future.

2.4.4.4 Aesthetic Literature vs. 'Non-Fiction'

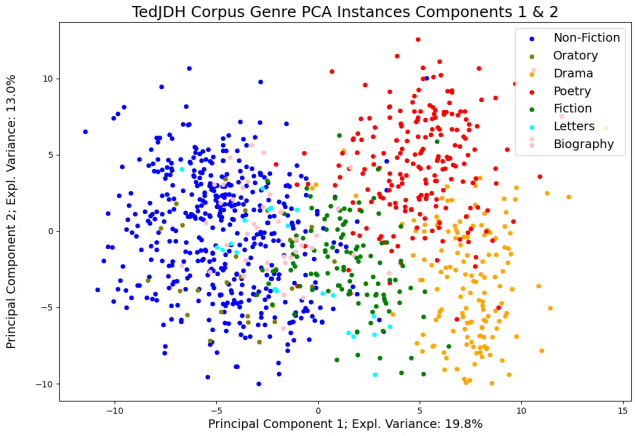

FIGURE 2.14: Instances of Three Genres of Literature vs. Non-Fiction in PCA, Full Feature Set

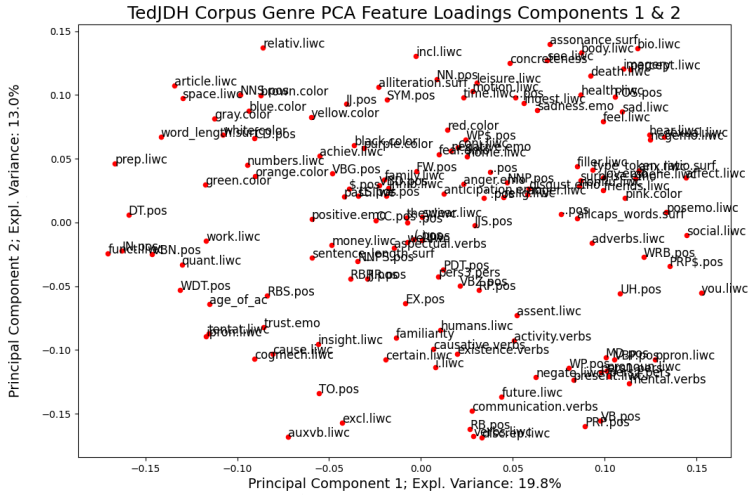

FIGURE 2.15: Feature Loadings of Three Genres of Literature vs. Non-Fiction in PCA, Full Feature Set

In Figures 2.14 and 2.15 we plotted PCA over all genres in the corpus with all features. This plot shows mainly the differences between literature and non-fiction, where poetry and drama are on the one side, fiction in the middle, and non-fiction on the other, far left side. We can clearly see that non-fiction does use a lot of prepositions and determiners and does talk about 'work' considerably, but it does not make use of interjections (UH), as this is clearly a feature of drama, and we also find that style devices like 'alliteration' and 'assonance', but also topics like 'death' and words associated with emotions all load to the literature side, indicating that these features are in fact markers for aesthetic literature, and specifically poetry.

2.4.5 Conclusion

We have shown that principal component analysis is a viable tool to show the variation of lexical features across literary genres, giving us insight regarding the communicative functions present in literature. Overall, we found characterizations of the genres, and also that there are two specific communicative functions present: An interpersonal dimension, and a narrative dimension. Furthermore, we have shown that we can identify an aesthetic dimension when comparing the literary genres with non-fiction.

Corpora 3

3.1 Introduction

A prerequisite for the computational study of literature is the availability of properly digitized texts, ideally with reliable meta-data and ground-truth or similarly reliable annotation. Several poetry corpora have been used in the Natural Language Processing community, both large and small, covering a number of languages. Smaller corpora allow the user to operate on a manageable number of poems that can be curated to be representative for a slice of poetic writing, and they can be annotated through manual labor with any stylistic devices of interest. Larger corpora on the other hand allow a far bigger picture, but to be useful for analysis, large computational models are needed to process the texts in order to see meaningful patterns. However, larger collections of poetry so far often lack consistency and are encoded in miscellaneous standards, while annotated corpora are typically small and constrained to particular text genres and/or were only designed with the analysis of certain linguistic features in mind (like rhyme or meter).

In this work, we compile large poetry corpora for German and English, and publish them with automatically detected syllable boundaries, part-of-

speech tags, and verse measure. The respective tools are developed in section 3.8 (part-of-speech tagging and syllabification) and chapter 6 (verse measure). Furthermore, we annotate poetic features in small corpora to then later train corpus-driven neural models that enable various experiments on linguistic variation and also robust large scale analysis. Smaller corpora allow us to annotate stylistic features to determine issues in disagreement, error or ambiguity. And then we may train corpus driven models that can be used to automatically annotate large corpora, i.a., with meter and verse measures.

We focus mainly on German and English, but also consider resources from other languages whenever they can be incorporated. Our corpora span roughly the time-frame from 1550 CE to 1936 CE, covering the era of New (High) German and a considerable amount of English poetry (from the 'Modern' era). This chapter discusses the corpora that we designed ourselves and some most annotation schemes that are implemented in them. Other corpora (from varying sources) are discussed at the place where they are used.

The remainder of this chapter is organized as follows: First of all, we give a quick overview on corpora that were built for the digital analysis of poetry. Second, we describe the large corpora that we built specifically for this thesis, alongside some surface statistics to get a first overview. Third, we describe the small corpora and give a first introduction to their annotation layers that are further elaborated on in Chapter 5 and Chapter 6, regarding specific annotation guidelines and their evaluation. Fourth, we describe the specific formats our corpora are stored in. And lastly, we document the development and evaluation of basic annotation tools for our corpora, in particular part-of-speech taggers and syllabifiers/hyphenators.

The corpora and code to process them can be found at

`https://github.com/tnhaider/metrical-tagging-in-the-wild`

`https://github.com/tnhaider/poetry-emotion`

`https://github.com/tnhaider/DLK`

3.2 Related Work

The main resources for New High German poetry were made available with the German Text Archive (Deutsches Textarchiv: DTA) and the Digital Library of Textgrid. Both of these corpora were curated within larger academic infrastructure projects (CLARIN, DARIAH, or BBAW). For English, most work is based on the collection of the Project Gutenberg, which grew and still grows on the shoulders of volunteers, or crowd-workers. Parrish (2018) previously published a dataset with the poetry from the English Gutenberg collection by filtering single lines with a heuristic (anything that could look like a line), but without considering the integrity of texts and their logical document structure. Jacobs (2018) scraped some poems from Project Gutenberg, but did not publish the resource.

Regarding larger collections in other languages, Ruiz Fabo et al. (2020) has published a diachronic corpus of Spanish sonnets, and Zhang and Lapata (2014) have compiled a corpus of classical Chinese Tang poetry which still finds wide application (around 280 citations in 2021), especially in research on (Chinese) poetry generation and work on language generation more generally. Smaller poetry corpora are also available for other languages and writing systems, such as Middle English (Zimmermann, 2015), Occitan (Wilson, 2012), Sanskrit (Krishna et al., 2019), or Old Greek (Tsagalis, 2009; Lamar and Chambers, 2019).

In regards to smaller corpora, there are resources for English, German and French that are annotated for rhyming patterns (Reddy and Knight, 2011; Sonderegger, 2011; Haider and Kuhn, 2018), alongside proposed methods to detect rhymes automatically. It should be noted that the English corpus of Sonderegger (2011) and Reddy and Knight (2011) puts a fairly strong focus on so-called perfect rhyme (which is untypical for English), and does not include stanzas that do not rhyme, consequently skewing the distribution of rhyming patterns. We on the other hand (Haider and Kuhn, 2018), have balanced a

German rhyming corpus over time that is closer to a real-world scenario to study the diversity of rhyming schema.

Regarding rhythmic patterns, Agirrezabal et al. (2016a,b, 2019) used a corpus of English poems, which was originally compiled by (Tucker, 2011), totalling around 1200 lines annotated for meter. Within a project of similar scope, Anttila and Heuser (2016) compiled a similarly sized collection of poems, also manually annotated for meter and metrical feet, according to the metrical constraints/theory proposed by Hanson and Kiparsky (1996).[1]. These two resources will be discussed in Section 3.8.2.1. The Spanish corpus of (Ruiz Fabo et al., 2020) is also annotated for rhythm/meter (Navarro et al., 2016; Navarro-Colorado, 2018a) and a form of enjambement (Ruiz et al., 2017). Estes and Hench (2016) compiled a corpus of Middle High German and annotated it for so-called hybrid meter (which is a hybrid between accent-based and length-based).

Poetry corpora, such as they are discussed here, have found application in work on English that has strongly focused on poetry that is written in iambic pentameter, either with a focus on Shakespeare (Greene et al., 2010; Jhamtani et al., 2017), or with a more general scope, by e.g., generating stanzas in iambic pentameter form (Hopkins and Kiela, 2017), or heuristically extracting sonnets (also in iambic pentameter) from the Project Gutenberg corpus (Lau et al., 2018). Other work has focused on specific genres like Spanish sonnets (Ruiz Fabo et al., 2020), limericks (Jhamtani et al., 2019), or Chinese Tang poetry (Zhang and Lapata, 2014).

Kao and Jurafsky (2012, 2015) built a corpus that distinguishes professional poets from amateurs by sampling from anthologies and web-forums respectively. Underwood and Sellers (2012) investigated the change in diction over time in poetry versus other genres of writing, providing a diachronic literature corpus that has since been included in a larger framework[2] and actively

[1] cf. `https://github.com/quadrismegistus/prosodic`
[2] See `https://github.com/quadrismegistus/lltk`

used. We use it in Chapter 2 for an experiment on genre stylistics. Lastly, a few poetry corpora are also annotated for emotions, notably our own corpus as discussed in Chapter 5 and in Haider et al. (2020), and some others mentioned in that paper.

Honorable mentions include the Chadwyck-Healey Poetry collections (for English), which are unfortunately not freely available to the public, and to the 'Freiburger Anthologie' that contained around 1800 German poems, but is currently only available in the context of `metricalizer.de`.

3.3 A German Poetry Corpus: Deutsches Lyrik Korpus (DLK)

As basis for the research of this thesis, we aimed to build a large, comprehensive, and easily searchable corpus of New High German poetry. We achieved this by collecting and parsing the bulk of digitized corpora that contain public domain German literature. As this newly compiled corpus contains the majority of digitized public domain poetry from the New High German period, we call this new corpus *German Poetry Corpus*, in German: *Deutsches Lyrik Korpus*, DLK for short. The corpus is available in a dedicated github repository: `https://github.com/tnhaider/DLK`

	TGRID 'Verse'	DTA: 'Lyrik'	DLK v5
#syllables	24,025,692	4,421,923	25,901,322
#words	16,049,526	2,986,912	17,335,638
#tokens	19,346,248	3,549,224	20,852,476
#lines	2,641,558	458,851	2,827,091
#stanzas	410,550	63,080	430,244
#poems	50,549	22,039	65,755
#authors	227	73	254

Table 3.1: Sub-Corpora of the German Poetry Corpus by Size

3.3.1 Contents and Size

This German Poetry Corpus corpus essentially contains the poetry from the German Text Archive (Deutsches Textarchiv: DTA)[3] and also the Digital Library of Textgrid.[4] DTA was originally mined from `wikimedia commons` and Textgrid was mined from `zeno.org`. Not all of the texts in these corpora were written in the German language (we found e.g., Latin or French poems), and across the corpora there is a considerable amount of duplicate poems. The language of poems was determined with the tool `langdetect`,[5] and the identification of duplicates is discussed below. Table 3.1 lists size statistics of these respective corpora and the final compiled DLK corpus.

As can be seen in Table 3.3, there are even 60,707 TEI texts with the genre label 'verse' in Textgrid, but not all of these contain line groups (`<lg type='poem'>`). The entire DTA corpus (not restricted to genre label 'Lyrik') contains a total of 40,077 line groups that look like poems, but without the proper genre labels (e.g., appearing in the genre 'science'), poems are likely embedded within other texts (by quotation, e.g., for criticism) and might not come with proper meta-data. Still, we kept DTA books that had no proper author name, but only author information 'N.A.'. In these cases, we annotated the author name 'Various', indicating anthologies with multiple contributing authors. These should be removed for authorship attribution studies. According to the number of `line group tags` with the `attribute-value pair type='poem'` (`<lg type='poem'>`), Textgrid contains 51,264 poems (poem line groups) with the genre label 'verse' (`<term>`), while DTA contains 23,877 poems with the genre label 'Lyrik' (`#dtasub`). However, after additional cleaning (e.g., removing duplicates, prose, and foreign language material like French or Latin), only the number of poems referenced in Table 3.1 remained.

[3] `http://deutschestextarchiv.de`
[4] `http://textgrid.de`
[5] https://pypi.org/project/langdetect/

DTA contains in total 128 documents, (complete) books that fall under the genre 'Lyrik', also with the possibility of multiple labels per document (such as 'Lyrik; Prosa'). In Table 3.2, we consider each label separately (thus in sum not counting the number of documents, but labels). The entire table can be found in the Appendix (Chapter 8).

# of Labels	Genre Label
155	Roman
128	Prosa
128	Lyrik
92	Leichenpredigt
78	Philosophie
76	Recht
70	Drama
46	Technik
44	Medizin
44	Historiographie
40	Geographie
40	Biologie
34	Psychologie
	[...]
12	Anstandsliteratur
11	Naturwissenschaft
10	Reiseliteratur
9	Chemie
7	Landwirtschaft
7	Kunstgeschichte
7	Handbuch
6	Sonstiges
5	Verslehre
5	Musik
	[...]

Table 3.2: Most Frequent Genre Labels in DTA per Document (File)

To get a better grasp on digitized German poetry, we also crawled the German version of Project Gutenberg (GUT-DE)[6]. However, we omit this corpus from our experiments, as it is wildly inconsistent (regarding its markup and document structure) and only offers metadata for less than 1/3 of its poems.

[6] `https://www.projekt-gutenberg.org/`

# of Documents	Genre Label
60,707	verse
29,624	other
3,290	prose
703	drama

Table 3.3: Genre Labels in Textgrid per TEI texts

This might not be very surprising, as Project Gutenberg developed mainly out of a crowd effort, rather than being curated by academic professionals. In total, GUT-DE contains 36,822 poems. Since this is a sizeable collection, it might still be useful for other work that does not depend on clean markup and metadata.

Unfortunately, it is not always clear from the Textgrid XML in which context a poem was published, as each poem comes with its own TEI P5 header, sometimes with adequate information, sometimes without it. Furthermore, titles (text headers) in Textgrid are not always correctly annotated (though DTA is not perfect here either), and there is no reference URN (of which DTA makes use to refer back to wikimedia). Additionally, it is not always clear if a Textgrid poem is actually just a stanza, since other poems with the same title exist (e.g., for Möricke). Additionally, despite considerable effort, there is no guarantee that there might still be (parts of) texts in the corpus that cannot be considered poetry, but are e.g., prose commentary with line breaks.

The version of DLK presented here is in Version 5 (v5). Version 2 is the full corpus of Textgrid and DTA amalgamted, but includes around 10,000 duplicate stanzas. Version 3 was cleaned up, but the integrity of certain poems was destroyed, because we removed duplicate stanzas without looking at the poem ids. Version 4 tried to reconstruct whole poems, but was still suffering from inconsistencies, broken Textgrid titles and sketchy duplication detection. Version 5 was completely rebuilt and now includes a number of automatic annotations like part-of-speech and syllable boundaries.

3.3.2 Temporal Distribution and Duplicates

An important factor to consider when compiling a diachronic corpus is the temporal distribution of poems by their publication date. Figure 3.1 shows a histogram of the number of poems over time, binned in 25 year increments. It is apparent that Textgrid (green) is considerably represented in most time slots, though it is a bit thin around the 1700 year mark. DTA is stronger in the pre-romantic period (pre 1750), but it is seriously lacking in substance in a majority of time slots (only containing a few hundred poems from 1850 to 1875). This illustrates that either corpus might not be considered representative for New High German poetry, due to significant underrepresentation in particular time slots. But together, we gain decent coverage over our time frame from 1600 to 1925 CE.

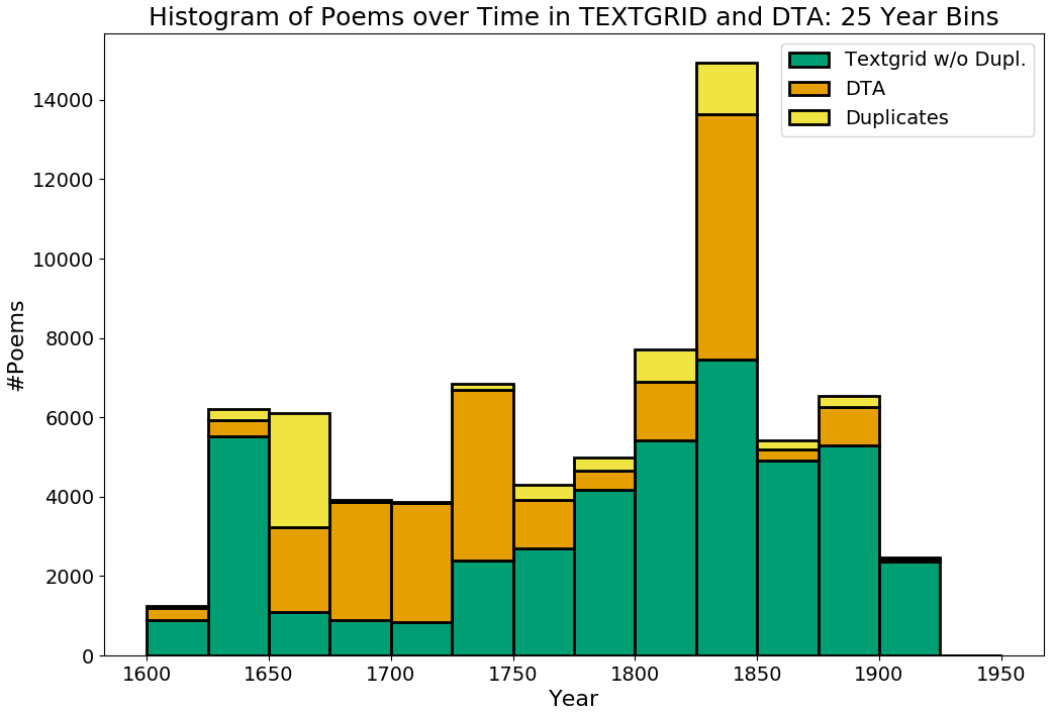

FIGURE 3.1: DTA and Textgrid Poems in 25 Year Bins. Identified duplicates are subtracted from Textgrid.

Since we aimed at a curated corpus, we removed duplicate poems. We identified duplicates by first grouping poems from both sub-corpora by authors

(after name standardization), and then calculated the Jaccard-Coefficient J (eq. 3.1) between the unigrams (word forms) of two poems A and B to measure their overlap.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.1}$$

We evaluated this metric by calculating J between all documents of the same author. We check J against titles and, if in doubt, by reading the actual texts. After manual inspection, we set a threshold for J to achieve high precision (to not identify false positives, i.e., saying that two texts are duplicates when in fact they are not). Optimizing for recall (not to miss too many actual duplicates) is hampered by not having a gold dataset, but set against precision, we could find a good balance.

Finally, if two poems exceeded the threshold $J = 0.5$, we considered these two poems duplicates (high J means more unigram overlap). It appears that in the time-frame 1650–1675 there are a number of duplicate poems within Textgrid itself already (which is not the case in DTA), even sharing the same title. Overall, DTA provides a cleaner resource, and if in doubt, we chose the DTA version of a poem to be included in DLK. In total, this method identifies more than 7600 poems as duplicates.

To get a better overview of the time stamps of poems in this corpus, see Figure 3.2, where we plotted each poem in DTA and Textgrid over time, from 1550 to 1950. Every dot represents a poem, where dots can lie on top of each other. Dots are partly transparent, so that fully saturated dots show poems that lie on top of each other. Red dots are Textgrid poems and blue dots are DTA poems.

This plot illustrates that the poetry in DTA is organized in whole books (editions), typically comprising collections of a single author with a proper book title. In contrast, Textgrid poems are each fairly independent and well distributed over time, where most poems contain their individual publication

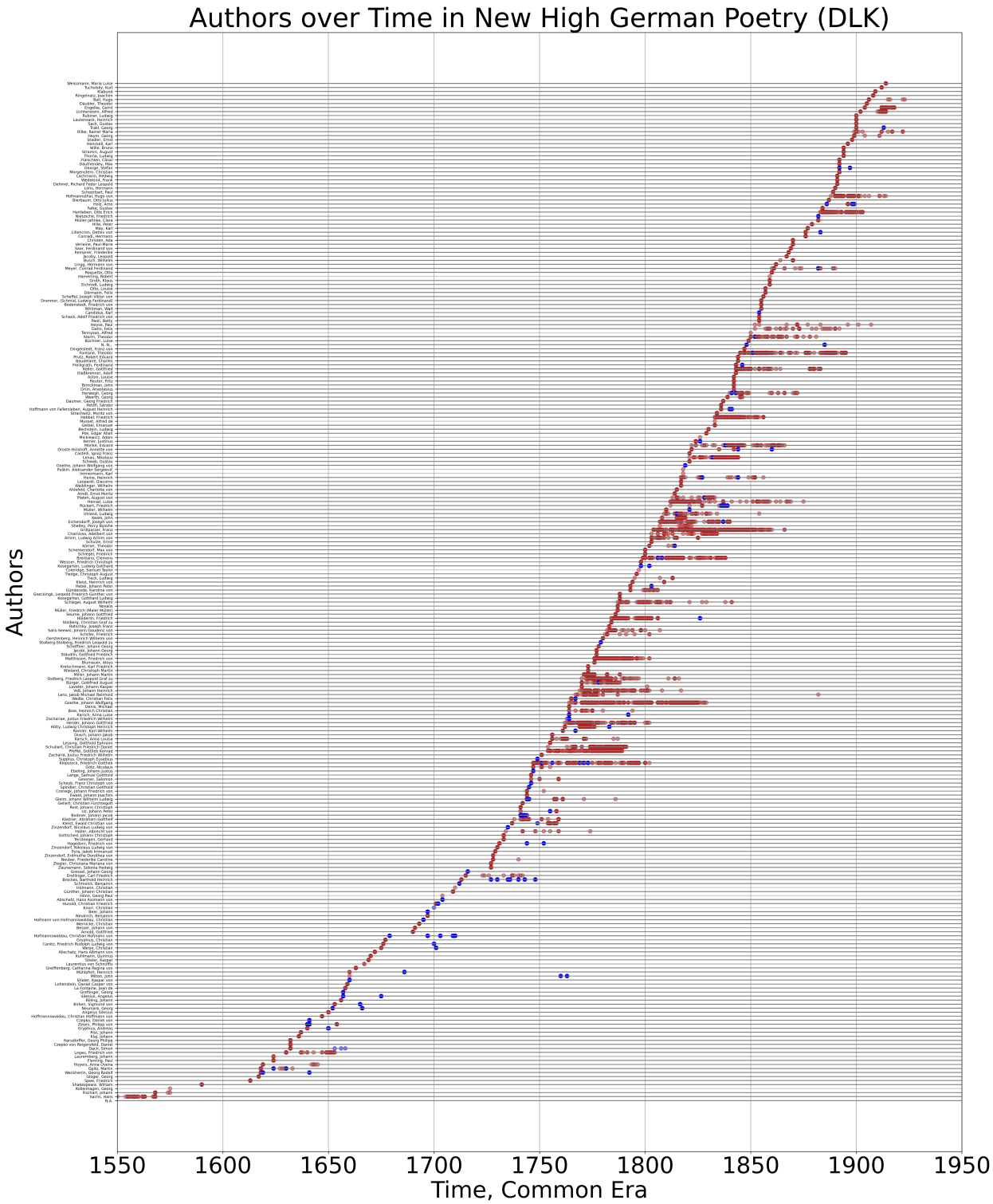

FIGURE 3.2: Poems over the years (1550–1950) from DTA and Textgrid. Each dot represents a poem of a German Author (y-axis) over Time (x-axis) from Textgrid (red dots) and DTA (blue dots).

date (rather than sharing their time stamps with all other poems from its publication).

The x-axis in Figure 3.2 shows the year of a poem, while the y-axis is populated by authors (readable when zoomed in), where both corpora are simply plotted on top of each other. One can see that DTA consists of full books that are organized by author, so that the datapoints for single poems get plotted on top of each other, while Textgrid has a time stamp for many individual poems (after 1750), outlining the productive periods of authors. When the publication date was not available in Textgrid, we used the `notBefore` and `notAfter` markers and took their mean.

3.3.3 Surface Statistics: Length of Poems and Lines

It is noteworthy that DTA poems are considerably shorter than Textgrid poems. As seen in Table 3.1, DTA contains about half the amount of poems than Textgrid (22k vs. 50k), but these amount to only a fifth (19%) in terms of total number of words (3k vs. 16k). Figure 3.3 shows a density plot of the length of poems in tokens for both corpora. Textgrid poems are overall longer with the highest density at 200 tokens, while DTA poems are shorter where most of them are around 100 tokens long. The length of poems gives a first indication of prevalent text genres in a corpus, as does the length of lines. Later in section 4.4.1.1, we will see the importance of these features to predict the publication year of a poem.

Consider this example: A typical sonnet contains 14 lines (4+4+3+3), and if these lines are set in iambic pentameter, each line is 10 or 11 syllables long. On average, a sonnet is then 147 syllables long. Thus, at an average word length of 1.5 syllables, a typical sonnet is around 100 words long (without punctuation tokens). This is not to say that DTA 'Lyrik' is mainly composed of sonnets, but it is fair to say that DTA is more dominated by short lyrical poems, while Textgrid contains comparatively longer forms.

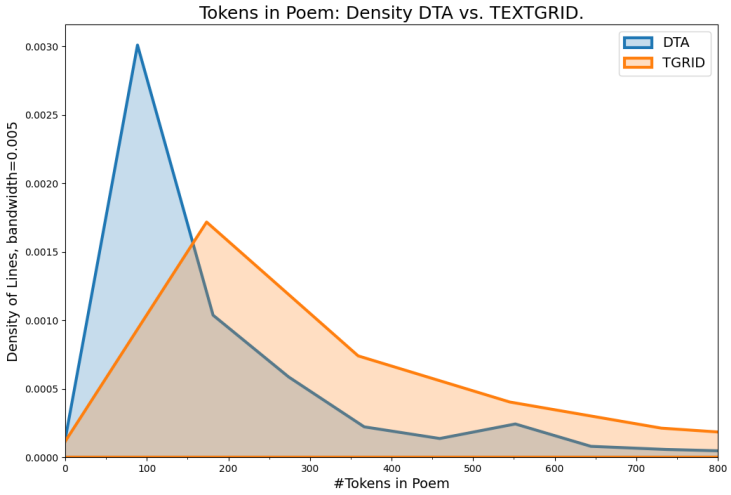

FIGURE 3.3: DTA and Textgrid Density Plot of Tokens in Poem. Bandwidth=0.005

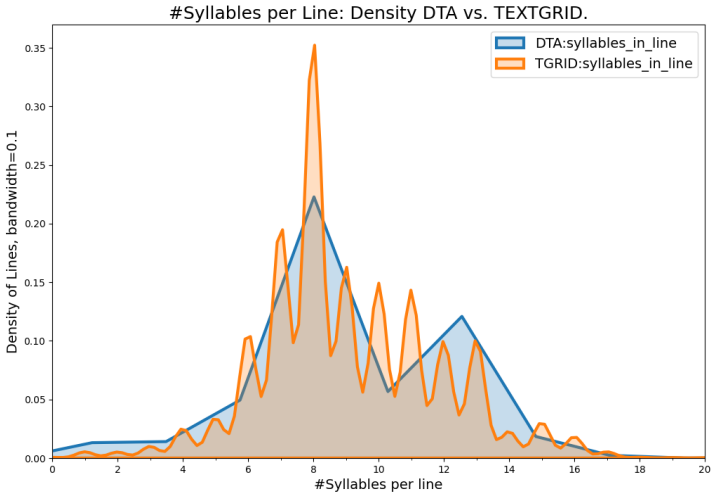

FIGURE 3.4: DTA and Textgrid Density Plot of Syllables in Line. Bandwidth=0.1

In contrast however, the stylistic differences in verse length across both corpora are not very pronounced. Figure 3.4 shows a density plot of the number of syllables in lines. Most lines in both corpora are 8 syllables long (iambic and trochaic tetrameter). DTA has another peak around 12–13 syllables, which hints at a large number of alexandrines.

This concludes a quick overview of the curation and some statistics of our large German poetry corpus. We will encounter the corpus or parts of it in the following chapters. The upcoming sections sketch the remaining corpora.

3.4 English Project Gutenberg Poetry Corpus (EPG)

We published an English poetry corpus at `https://github.com/tnhaider/metrical-tagging-in-the-wild` which stems from the English Project Gutenberg (EPG) collection, but was rigorously cleaned from duplicates and foreign language material and additionally annotated on syllable boundaries and verse measures. Some basic statistics on its size are documented in Table 3.4.

	EPG Large
#syllables	11,542,525
#tokens	9,426,889
#lines	1,109,275
#stanzas	155,615
#poems	35,022
#authors	537

Table 3.4: Sub-Corpora of the German Poetry Corpus by Size

To compile the corpus, we firstly collected all files with the metadatum 'poetry' in (temporal) batches with the GutenTag tool (Brooke et al., 2015b). This unadulterated corpus was then used in the experiments in Chapter 4 section 3.3, containing around 22 million words in 85k poems.

To clean up the corpus, we standardized the inconsistent XML annotation of GutenTag, and we removed duplicates, since the collection contained numerous different editions and issues containing the same material. To that end, we used the method that was introduced in the previous section for DLK (Jaccard Coefficient on Unigrams). However, here we check for duplicates by comparing documents based on their title first (whether they are just different editions of the same book). We also filter out any lines (or tokens) that indicate illustrations, stage directions and the like. We used langdetect 1.0.8 to filter any non-English material.[7] The result of this rigorous cleaning is a corpus that unfortunately lost about half of its size. Furthermore, the corpus is rather sparse before 1800 CE, as seen in Figure 3.5. However, as seen in section 4.4.2, this was also the case before cleaning the corpus.

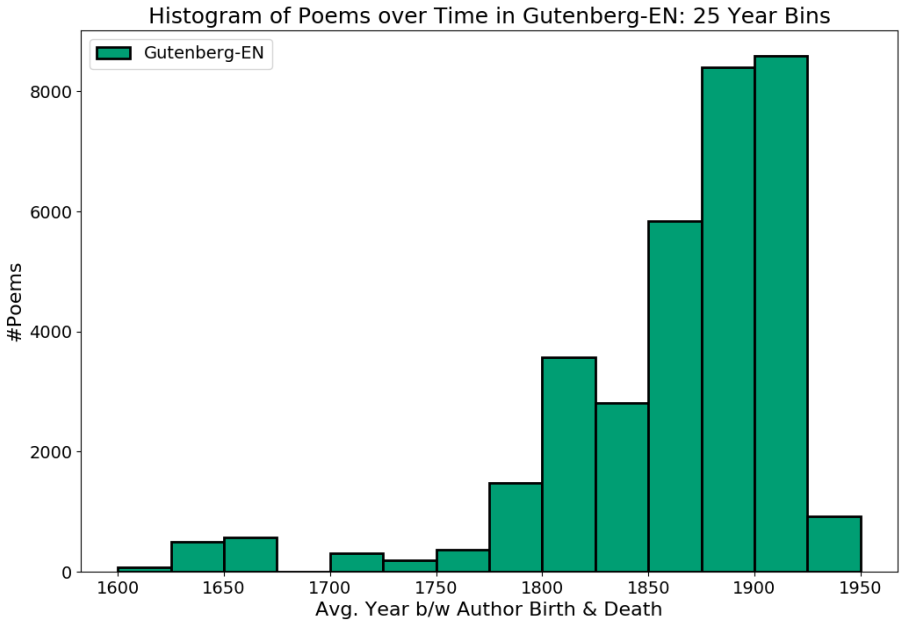

FIGURE 3.5: Histogram of #Poems over Time in Gutenberg-EN

[7] `https://pypi.org/project/langdetect/`

3.5 Small Corpora for Manual Annotation (ANTI-K & EPG64)

For our annotation and modeling studies, we build on top of two poetry corpora (in English and German). This collection represents important contributions to the literary canon over the last 400 years. We make this resource available in TEI P5 XML[8] and an easy-to-use tab separated format. Table 3.5 shows a size overview of these data sets. Figure 3.6 shows the distribution of our data over time via density plots. Note that both corpora show a relative underrepresentation before the onset of the romantic period (around 1750 CE).

	German	English
# tokens	20403	8082
# lines	3650	1240
# stanzas	731	174
# poems	158	64
# authors	51	22

Table 3.5: Statistics on our poetry corpora *PO-EMO*. Tokens without punctuation.

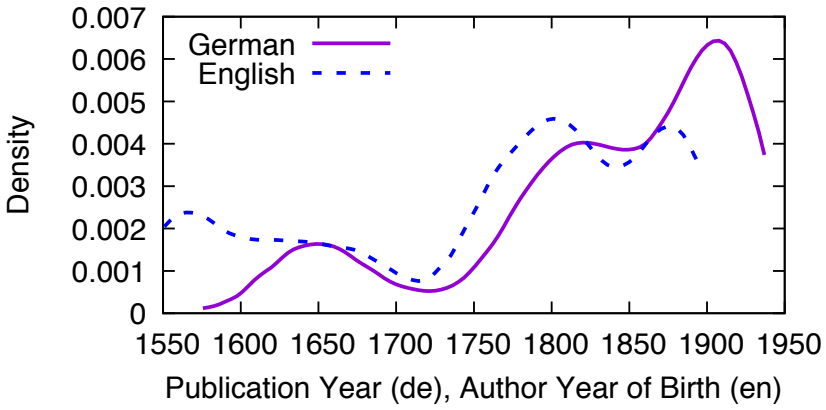

Figure 3.6: Temporal distribution of small poetry corpora (Kernel Density Plots with bandwidth = 0.2).

[8] https://tei-c.org/guidelines/p5/

In these corpora, we manually annotated the sequence of syllables for metrical (meter, met) prominence (`+/-`), including a grouping of recurring metrical patterns, i.e., foot boundaries (`|`). We also annotated a more natural speech rhythm (rhy) by annotating pauses in speech, caesuras (`:`), that segment the verse into rhythmic groups, and in these groups we assigned main accents (2), side accents (1) and null accents (0). In addition, we developed a set of regular expressions that derive the verse measure (msr) of a line from its raw metrical annotation (see Appendix). This prosodic annotation is evaluated in Chapter 6. Furthermore, we annotated aesthetic emotions, as detailed in Chapter 5. In Section 6.2.1 we describe the annotation of rhyme.

3.5.1 Small German Corpus: Antikoerperchen (ANTI-K)

The German corpus contains poems that we crawled from the website `lyrik.antikoerperchen.de`, hence the name ANTI-K. `Antikoerperchen.de` provides a platform for students to upload essays about poems. The data was available in the Hypertext Markup Language, with clean line and stanza segmentation, which we transformed into TEI P5.

The small German corpus is fairly diverse, considering its size, and covers not only a wide range of different poem lengths and verse measures but also a number of influential German poets of both genders. The corpus can be considered to be 'school canon', containing poems that are discussed and interpreted in the German school system. Besides the annotation of poetic features, every poem also has information on the author name, a title, the year of publication, and literature periods. See Figure 3.7 for a rough overview of the annotation of literary periods in that corpus. Note that this annotation was done by crowd workers without clear guidelines (contributors to the antikoerperchen website), and thus does not have standardized period labels.

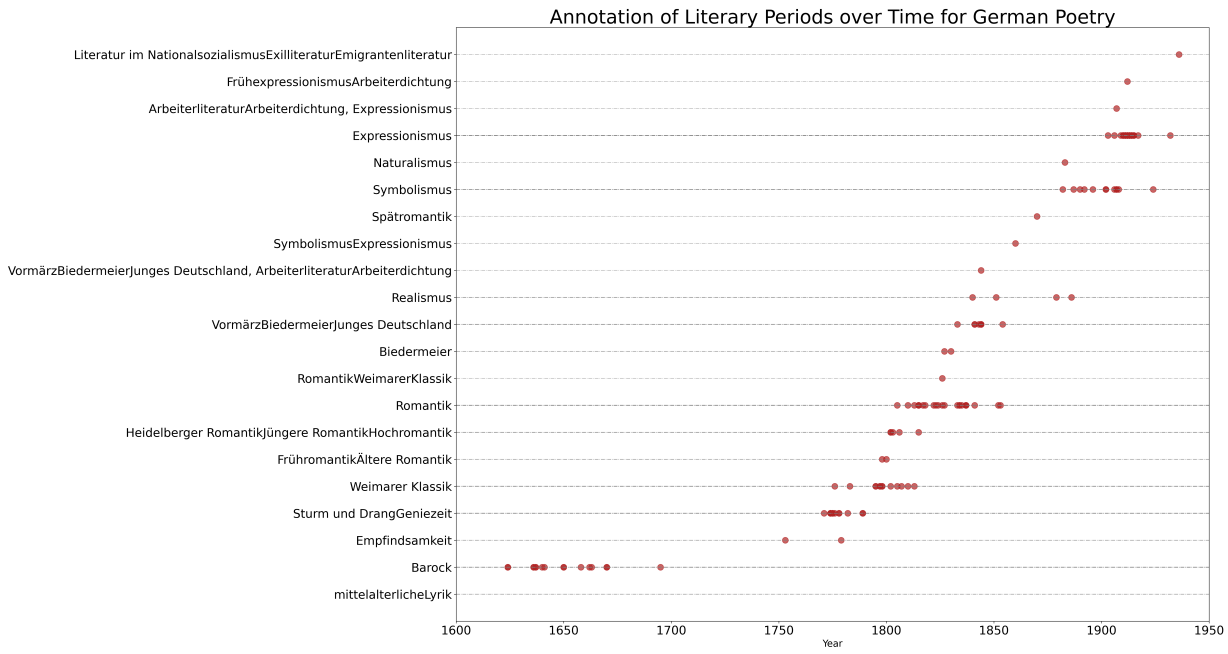

Figure 3.7: Literary Periods Annotation of German ANTI-K Poems over Years.

The included 158 poems are dispersed over 51 authors and the New High German timeline (1575–1936 CE). For the annotation of prosody, we exclude two Middle High German poems by Walther von der Vogelweide and three poems in free rhythm (by Goethe) that do not allow for a metrical analysis, effectively amounting to 3.489 lines in 153 poems, spanning a time period from 1636 to 1936 CE. This yields a corpus that is somewhat representative for classical New High German poetry, while remaining manageable for manual annotation. Also see Table 8.2 in the Appendix (Chapter 8) for an overview of all poems with author names and publication date.

3.5.2 EPG64 (Small English)

The small English corpus contains 64 poems of popular English writers. It was partly collected from Project Gutenberg with the GutenTag tool,[9] and, in addition, contains a number of hand selected poems from the modern period and represents a cross section of popular English poets. We took care to include a number of female authors, who would have been underrepresented in a uniform sample. Time stamps in the corpus (as seen in Figure 3.6) are organized by the birth year of the author, as assigned in Project Gutenberg. This corpus is also annotated with the same prosodic features and aesthetic emotions like ANTI-K.

AlexanderPope	1688
AlfredLordTennyson	1809
AnneKilligrew	1660
EdmundSpenser	1552
EdwardEstlinCummings	1894
ElizabethBarrettBrowning	1806
EmilyDickinson	1830
JohnDonne	1572
JohnDryden	1631
JohnKeats	1795
JohnMilton	1608
PercyByssheShelley	1792
RobertBurns	1759
RobertFrost	1874
RudyardKipling	1865
ThomasStearnsEliot	1888
WaltWhitman	1819
WilliamBlake	1757
WilliamButlerYeats	1865
WilliamCarlosWilliams	1883
WilliamWordsworth	1770

Table 3.6: Authors in EPG64 (Small English Corpus) with Year of Birth

[9] `https://gutentag.sdsu.edu/`

3.6 Rhyming Corpora (DTA-RHYME & Hip-Hop)

We created a rhyming goldstandard for German that is used in Chapter 6, by drawing a diachronically balanced sample from DTA. We divide our timeline by 20-year wide slots (1630 - 1650, ..., 1790 - 1810, etc.), aiming at 500 stanzas per slot (allowing ± 10%). We left the original poems intact, sampling until the desired number of stanzas was fulfilled with complete poems. An additional sampling constraint defines that an author needs to contribute enough poems within a std. deviation from the mean, so that sampling does not favor overly represented poets. Also no poems with stanzas longer than one standard deviation over 12 lines (24) were allowed. The sampled corpus eventually contains 1,948 poems over 8,147 stanzas. Students then annotated rhyme schema on stanza level (e.g. 'abba'). We extracted the rhyming word pairs and we further clean this set which results in a total number of 13,785 rhyming word pairs.

We also collected 116 German Hip-Hop song texts and annotated them on rhyme and assonance (repetition of vowels). We retrieved the documents in plain text from `hiphoplyrics.de`, mainly covering the 90's and 2000's, with one to four texts per author. Since assonance and rhymes often form a complex schema in Hip-Hop lyrics, we decided to mark assonance with capital letters in the stanza level rhyme schema to extract them separately. We retrieve 2,489 rhyme pairs and 1,032 assonance pairs.

Corpus	Poems	Stanzas	Rhyme Pairs
ANTI-K	156	731	1,440
HIPHOP	116	789	2,489
DTA-RHYME	1,948	8,147	13,784

Table 3.7: Size of German Rhyming Corpora

Other datasets that we use in this research originate from Reddy and Knight (2011) and Sonderegger (2011), who provide English and French poems

annotated for rhyme schema. These corpora are called the English and French Chicago Rhyming Corpus. We found thast the English corpus does not include any stanzas that do in fact *not* rhyme, as seen in the German schemas 'ab', 'abc', and 'abcd'. This may potentially impact any models that are trained on this data. Overall, the German DTA-RHYME corpus contains about a third of stanzas that do not rhyme, thus only two thirds of stanzas actually rhyme. For these reasons we believe that this corpus is of a quality that will allow further research on the distribution of rhyming patterns. Experiments with these corpora will be discussed in Section 6.2.1, including an overview of the frequency of schemas.

3.7 Corpus Formats: TEI P5, .json, and .tsv

On the following pages you can see examples for the format that we use to store our corpora. First, in Figure 3.8 a poem set in .json, , in Figure 3.9 a poem set in .json with additional meter/measure annotation, in Figure 3.10 you can see a TEI P5 XML `<header>`, and its `<body>` in Figure 3.11. Second, in Section 3.7.1 the tabular format that is used for the experiments on prosody in Chapter 6. For more guidelines on TEI P5 for poetry please see the Appendix.

The .json Format is organized as python dictionary. Every poem has a unique index as key, some metadata, and in the 'standard' version, every line encompasses its text, tokenization, syllabification (inside the tokenization), information on the type of tokens, and part-of-speech (pos) annotation. In the 'meterized' version, each line also provides the sequence of metrical stresses (the raw 'meter'), and a verse measure label that was derived from the raw sequence with regular expressions. The XML format includes more fine-grained (and standardized) tags, and is more geared towards archiving. Details of the actual poetry annotation will be discussed in Section 8.1.

Our resources are designed in a standardized format to sustainably and interoperably archive poetry in both .json (for the larger corpora) and TEI P5

XML (for the smaller corpora). The .json format is intended for ease of use and speed of processing, while being expressive enough to deliver the logical document structure of poems from full texts down to syllable level, including the most important metadata. Our XML format is geared towards providing an annotation standard that may be understood as suggestion of how poetic text can be augmented with extra annotation. We define a header structure that offers the most used encoding strategies for metadata. Also, we propose annotation schemes for a multitude of linguistic and psychological concepts. Where possible, we utilize in-line annotation (rather than using stand-off).

Our framework is grounded in the so-called DTA-Basisformat[10] (Haaf et al., 2014), that provides a "Base Format", which not only constrains the data to TEI P5 guidelines, but also regarding a stricter so-called relaxNG schema, which we modified to fit our annotation layers. This relaxNG schema defines a strict layout of poetic annotation, where tags, attributes and values can only assume a certain format, such as constraining the annotation to only characters or numbers of specified length. This additional grammar allows us to validate XML files regarding their correctness. It is thus useful for manual annotation with the OxygenXML editor, avoiding parsing errors later on. We implemented XML parsers in python to parse existing formats in order to extract poems with their metadata and fix stanza and line boundaries. We performed cleaning procedures that remove extant XML information, obvious OCR mistakes, and normalize umlauts and special characters in various encodings, particularly in DTA.[11] We use langdetect[12] 1.0.8 to tag every poem with its language to filter out any poems that are not German or English (such as Latin or French).

[10] `http://www.deutschestextarchiv.de/doku/basisformat/`

[11] We normalized a mixture of HTML fragments, latin-1 and utf-8 text encodings, and cases where bytecode was saved as string. We fix the orthography both on string and bytecode level. We replace the rotunda (U+A75B) and the long s (U+017F), the latter of which is pervasive in DTA. Also, we fix the awkward handling of umlauts and other special characters in DTA.

[12] `https://pypi.org/project/langdetect/`

FIGURE 3.8: A .json Format for Poetry.

```
"dta.poem.21698": {
    "metadata": {
        "author": {
            "name": "Various",  # In Margin: 'Carl Bleibtreu' (see faksimile via urn)
            "birth": "N.A.",
            "death": "N.A."
        },
        "title": "8.",
        "genre": "Lyrik",
        "period": "N.A.",
        "pub_year": "1885",
        "urn": "urn:nbn:de:kobv:b4-200905196929",
        "language": ["de:0.99"],
        "booktitle": "Arent, Wilhelm (Hrsg.):
                     Moderne Dichter-Charaktere. Leipzig, [1885]."
        },
    "poem": {
        "stanza.1": {
            "line.1": {
                "text": "Den Auserkorenen hat eine Feder",
                "tokens": [
                "Den", "Aus·er·ko·re·nen", "hat", "ei·ne", "Fe·der"
                ],
                "token_info": [
                "word", "word", "word", "word","word"
                ],
                "pos": [
                 "ART", "NN", "VAFIN", "ART", "NN"
                ]
            },
            "line.2": {
                "text": "Aus seiner Schwinge der Simurg geweiht:",
                "tokens": [
                "Aus", "sei·ner", "Schwin·ge", "der", "Si·murg", "ge·weiht", ":"
                ],
                "token_info": [
                "word", "word", "word", "word", "word", "word", "punct"
                ],
                "pos": [
                "APPR", "PPOSAT", "NN", "ART", "NE", "VVPP", "$."
                ]
            },
        },
        "stanza.2: {
            [....] [....] [....] [....]
        }
    }
}
```

Figure 3.9: A Poem with Meter Annotation from DLK in .json


```
"dta.poem.878": {
    "metadata": {
        "author": {
            "name": "Trakl, Georg",
            "birth": "N.A.",
            "death": "N.A."
        },
        "title": "DIE RABEN",
        "genre": "Lyrik",
        "period": "N.A.",
        "pub\_year": "1913",
        "urn": "urn:nbn:de:kobv:b4-30357-9",
        "language": ["de:0.99"],
        "booktitle": "Trakl, Georg: Gedichte. Leipzig, 1913."
        },
    "poem": {
        "stanza.1": {
            "line.1": {
                "text": "Über den schwarzen Winkel hasten",
                "tokens": ["Ü·ber", "den", "schwar·zen", "Win·kel", "has·ten"],
                "token_info": ["word", "word", "word", "word", "word"],
                "pos": ["APPR", "ART", "ADJA", "NN", "VVFIN"],
                "meter": "+--+-+-+-",
                "measure": "iambic.tetra.invert"
                },

            "line.2": {
                "text": "Am Mittag die Raben mit hartem Schrei.",
                "tokens": ["Am", "Mit·tag", "die", "Ra·ben",
                "mit", "har·tem", "Schrei", "."],
                "token_info": ["word", "word", "word", "word",
                "word", "word", "word", "punct"],
                "pos": ["APPRART", "NN", "ART", "NN",
                "APPR", "ADJA", "NN", "\$."],
                "meter": "-+--+--+-+",
                "measure": "amphibrach.tri.plus"
                },

    [...]
```

FIGURE 3.10: TEI P5 XML Header Format for Poetry.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title type="main">Weltende</title>
        <author>
          <persName>
            <surname>Lasker-Schüler</surname>
            <forename>Else</forename>
          </persName>
        </author>
        <editor>
          <persName>
            <surname>Haider</surname>
            <forename>Thomas Nikolaus</forename>
          </persName>
          <orgName>Max Planck Institute for Empirical Aesthetics,
                    Frankfurt am Main</orgName>
          <email>thomas.haider@ae.mpg.de</email>
        </editor>
      </titleStmt>
      <editionStmt>
        <edition>
          <name>Deutsches Lyrik Korpus Edition (DLK)</name>
          <date>1-11-2017</date>
        </edition>
      </editionStmt>
      <extent>
        <measure type="stanzas">3</measure>
        <measure type="verses">10</measure>
        <measure type="verses_per_stanza">1-4, 2-3, 3-3</measure>
        <measure type="tokens">56</measure>
        <measure type="sentences">4</measure>
        <measure type="characters">259</measure>
      </extent>
      <publicationStmt>
        <publisher>
          <name/>
        </publisher>
        <pubPlace/>
        <date type="publication">1903</date>
      </publicationStmt>
      <sourceDesc>
        <p corresp="http://lyrik.antikoerperchen.de/else-lasker-schueler-weltende
        ,textbearbeitung,337.html">
        Weltende - Else Lasker-Schüler (Interpretation #337)</p>
      </sourceDesc>
    </fileDesc>
    <profileDesc>
      <textClass>
        <classCode scheme="literary_period">Expressionismus</classCode>
        <classCode scheme="text_genre">Gedicht</classCode>
      </textClass>
    </profileDesc>
  </teiHeader>
```

Figure 3.11: TEI P5 XML Body Format for Poetry.

```
The basic structure of our TEI Body is as follows:
  <text>
    <body>
      <div n="1">
        <lg type="poem">
          <head>Weltende</head>
          <lg type="stanza">
            <l>Es ist ein Weinen in der Welt,</l>
            <l>Als ob der liebe Gott gestorben wär,</l>
            <l>Und der bleierne Schatten, der niederfällt,</l>
            <l>Lastet grabesschwer.</l>
          </lg>
          <lg type="stanza">
            <l>Komm, wir wollen uns näher verbergen...</l>
            <l>Das Leben liegt in aller Herzen</l>
            <l>Wie in Särgen.</l>
          </lg>
          <lg type="stanza">
            <l>Du! wir wollen uns tief küssen -</l>
            <l>Es pocht eine Sehnsucht an die Welt,</l>
            <l>An der wir sterben müssen.</l>
          </lg>
        </lg>
      </div>
    </body>
  </text>
</TEI>
```

3.7.1 Data Format for Experiments on Prosody

Figure 3.12 shows an example line in the data layout that is used for the experiments on prosody, including the 'measure' that was derived with regular expressions from the raw meter line. 'Syll' is the position of the syllable in a word, 0 for monosyllaba, otherwise index starting at 1. We removed punctuation to properly render line measures, even through punctuation is a good signal for caesuras (see e.g., Figure 6.4 in Chapter 6 for reference on caesura annotation). The format here is a tab-separated format (.tsv), including our prosodic annotation layers (but i.a., excluding emotions).

```
# tok   met ft pos  syll csr main smsr    measure    met_line

1  Look   +  .  VB   0   .  1   iambic  i.penta.inv +--+-+-+-+
2  on     -  .  IN   0   .  0   iambic  i.penta.inv +--+-+-+-+
3  my     -  .  PRP$ 0   .  0   iambic  i.penta.inv +--+-+-+-+
4  works  +  :  NNS  0   :  2   iambic  i.penta.inv +--+-+-+-+
5  ye     -  .  PRP$ 0   .  0   iambic  i.penta.inv +--+-+-+-+
6  Might  +  :  NNP  1   .  1   iambic  i.penta.inv +--+-+-+-+
7  y      -  .  NNP  2   :  0   iambic  i.penta.inv +--+-+-+-+
8  and    +  :  CC   0   .  0   iambic  i.penta.inv +--+-+-+-+
9  de     -  .  VB   1   .  0   iambic  i.penta.inv +--+-+-+-+
10 spair'+  :  VB   2   :  1   iambic  i.penta.inv +--+-+-+-+
```

FIGURE 3.12: Tabular data format for experiments.
Author of this line: Percy Blythe Shelley.

3.8 First Automatic Annotation: POS-Tagging and Syllabification

Properly working preprocessing tools are paramount to any pipeline approach in computational modeling. Mistakes that are made at the beginning of the pipeline may propagate through the whole analysis and in a best-case scenario lead to data that is unusable (low recall) and/or cannot be evaluated, or in a worst-case scenario taint the final results.

Detecting syllable boundaries and parts-of-speech are at the basic level of poetry analysis. Such tools can give us first insights regarding the prosodic and syntactic structure of verse. It has been frequently hypothesized that poetry makes use of 'non-canonical' syntactic structures, and that these 'non-canonical' forms emerge in favor of a prescribed metrical structure. This is naturally problematic: Models trained on out-of-domain data surely degrade in performance when applied to a very different domain, such is the case for Named Entity Recognition (a task closely related to part-of-speech tagging), as Augenstein et al. (2017) and Bamman et al. (2019) have shown. Furthermore, without in-domain test data, it is difficult to directly estimate the severity of this degradation.

Fortunately, there is great precedent on the proper design and evaluation of syllabification systems, and also training part-of-speech taggers for non-standard language varieties. For syllabification, Bartlett et al. (2009) found that rule-based expert systems modeled after linguistic theories (cf. legality and sonority principles) are outmatched by data-driven machine learning models (notably SVM-HMMs), even though the rule-based systems can be carefully improved to matching performance in some cases. A more recent paper (Gyanendro Singh et al., 2016) finds that Conditional Random Fields (CRF) and 'Hybrid Approaches' are effective for syllabification in Manpuri. With 'Hybrid' they mean an ensemble-like approach that however does not vote wrt.

competing systems, but does rule-based stacking of different systems to work around their individual weaknesses. In our German syllabification system we adopted a similar approach.

The task of part-of-speech tagging for non-standard varieties has also seen some interest. Westpfahl (2014); Westpfahl and Schmidt (2016) proposed to expand the Stuttgart-Tübingen Tagset (STTS) categories for spoken language. Van der Goot et al. (2017) investigated the impact of normalization on part-of-speech tagging of tweets (it helps a little bit). Lameris and Stymne (2021) found that having a little bit of annotated Scottish data is better than trying to do zero-shot domain-adaptation from English. Most relevant to our work are Bollmann (2013) and Schulz and Kuhn (2016). Bollmann (2013) is working in an extreme low-resource setting, where basically no training data in the target domain is available. However, utilizing a tiny set of normalized tokens allows him to significantly boost the performance of a tagger for early Modern German text (15th/16th century). Schulz and Kuhn (2016) find that already a few annotated text samples from the target domain benefit most models, most importantly also a LSTM model.

Unfortunately, some relevant research regarding the syntax and syllable structure of poetry has ignored the proper evaluation of syllabification and part-of-speech (pos) tagging. In consequence, some results may not be as robust as claimed. We will just enumerate a few examples. In section 6.4.2 we show that the popular tool `prosodic` (which determines metrical annotation from text), using a lexicon driven syllabifier (CMU dictionary), determined the wrong length (in number of syllables) for about 1/3 of all lines in our gold testset (EPG64). In work on the syntax of poetry, Gopidi and Alam (2019) used an off-the-shelf pos-tagger and dependency parser to show differences in syntactic constructions over two different time periods of poetic writing. These claims are made without evaluating their tagger and parser on in-domain data. Certainly, since they use rule-derived patterns on top of the parses, a problem might mainly arise from low recall, such that syntactic anomalies are just

not found to a certain degree. As we show in section 3.8.1, our off-the-shelf Stanford pos-tagger for English only achieved .72 accuracy on our gold poetry data. This made the annotation near impractical for most experiments in our case. Lau et al. (2018) crawled a dataset of sonnets from the English Project Gutenberg Corpus with a heuristic and word and character statistics derived from Shakespeare's 154 sonnets. The quality check of that dataset was done by manual inspection, and reported informally in the paper. Note that the performance of their system is then evaluated on this corpus. And only on a sidenote: Assylbekov et al. (2017), comparing syllable-based language models with character-based ones, used Liang (1983)'s hyphenation algorithm, which is the rule- and lexicon-based hyphenator that is widely used for TeX. The use of that algorithm is mentioned in a caption and is not evaluated. Granted, the mistakes this algorithm makes are likely systematic and therefore might not degrade the performance of the resulting language model. It is still remarkable that claims about syllable-based language models are made without actually scrutinizing the underlying structure.

Tokenization for both languages is performed with SoMaJo (Proisl and Uhrig, 2016), with a more conservative handling of apostrophes (to leave words with elided vowels intact). This tokenizer is more robust than e.g., a standard tokenizer from the NLTK toolkit, particularly in regards to special characters and punctuation marks (such as infixed apostrophes, as they are frequently used in poetry).

3.8.1 POS tagging

Since we are dealing with historical data, POS taggers trained on current data might degrade in quality. Additionally, it has been frequently noted that poetry makes use of non-canonical syntactic structures (Gopidi and Alam, 2019). For German, we evaluate the robustness of POS taggers across different text

genres. We use the gold annotation of the TIGER corpus (modern newspaper), and pre-tagged sentences from DTA, including annotated poetry (Lyrik), fiction (Belletristik) and news (Zeitung).[13] The STTS tagset is used. We train and test Conditional Random Fields (CRF)[14] to determine a robust POS model.[15] See Table 3.8 for an overview of the cross-genre evaluation. We find that training on TIGER is not robust to tag across domains, falling to around .80 F1-score when tested against poetry and news from DTA. It is however sufficient to use a tagger trained on 'Belletristik' (belles lettres) to tag poetry. These results suggest that the out-domain degradation of model quality is mainly due to (historically) deviant orthography, and to a lesser extent due to local syntactic inversions. For our experiments, we use the model trained on DTA.

Test	Train				
	TIGER	DTA	DTA+TIG.	Belletr.	Lyrik
Lyrik	.795	.949	.948	.947	**.953**
Belletristik	.837	**.956**	.954	.955	.955
DTA Zeitung	.793	**.934**	.933	.911	.900
TIGER	**.971**	.928	.958	.929	.913

Table 3.8: Evaluation of German POS taggers across genres. F1-scores.

For English, we test the Stanford core-nlp tagger.[16] Its tagset follows the convention of the Penn TreeBank. This tagger is not geared towards historical poetry and consequently fails in a number of cases. We manually correct 50 randomly selected lines and determine an accuracy of 72%, where particularly the 'NN' tag is overused. This renders the English POS annotation unreliable for further experiments.

[13] DTA was tagged with TreeTagger and manually corrected afterwards. See `http://www.deutschestextarchiv.de/doku/pos`

[14] From sklearn crf-suite: `https://sklearn-crfsuite.readthedocs.io/`

[15] As features, we use the word form, the preceding and following two words and POS tags, orth. information (capitalization), character prefixes and suffixes of length 1, 2, 3 and 4.

[16] `https://nlp.stanford.edu/software/tagger.shtml`

3.8.2 Hyphenation / Syllabification

Syllabification is the process of dividing a word into its constituent syllables. Although some work has been done on syllabifying orthographic forms (Müller et al., 2000; Bouma, 2002; Marchand and Damper, 2007; Bartlett et al., 2008), syllables are, technically speaking, phonological entities that can only be composed of strings of phonemes. Most linguists view syllables as an important unit of prosody because many phonological rules and constraints apply within syllables or at syllable boundaries (Blevins, 1995). Apart from their purely linguistic significance, syllables play an important role in speech synthesis and recognition (Kiraz and Möbius, 1998; Pearson et al., 2000)

We test the following systems: *Sonoripy*,[17] *Pyphen*,[18] *hypheNN*,[19] and a BiLSTM-CRF (Reimers and Gurevych, 2019)[20] with pretrained word2vec character embeddings. These character embeddings were trained on the corpora in section 3.3 (DLK) and 3.4 (EPG), segmented at character level, with word2vec from gensim.[21]

Syllabipy/Sonoripy determines boundaries based on the sonority principle, *Pyphen* uses the Hunspell dictionaries, and *HypheNN* is a simple feed forward network that is trained on character windows (whether the syllable boundary is in the middle of eight characters).

To train and test our models, we use CELEX2 for English and extract hyphenation annotation from wiktionary for German. For German, wiktionary contains 398.482 hyphenated words, and 130.000 word forms in CELEX. Unfortunately, German CELEX does not have proper umlauts, and models trained on these were not suitable for poetry. For English, wiktionary only contains 5,142 hyphenated words, but 160,000 word forms in CELEX.

[17] `https://github.com/alexestes/SonoriPy`
`https://github.com/henchc/syllabipy`
[18] `pyphen.org`
[19] `github.com/msiemens/HypheNN-de`
[20] `https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf`
[21] `https://radimrehurek.com/gensim/models/word2vec.html`

We evaluate our models on 20,000 randomly held-out words for each language against *word accuracy* and *syllable count* metrics. Word accuracy rejects any word with imperfect character boundaries, while syllable count is the more important figure to determine the proper length of a line. As seen in Table 3.9, the BiLSTM-CRF performs best for English and does not need any post-processing. For German, the LSTM model is less useful as it tends to overfit, where over 10% of annotated lines were still rejected even though in-domain evaluation suggests good performance. We therefore use an ensemble with HypheNN, Pyphen and heuristic corrections for German, with only 2% error on the gold data, as seen in Table 3.10. The datasets are discussed in section 6.4.1 and in the following section 3.8.2.1.

	German		English	
	w. acc.	sy. cnt	w. acc.	sy. cnt
SonoriPy	.476	.872	.270	.642
Pyphen	.839	.875	.475	.591
HypheNN	.909	.910	.822	.871
BiLSTM-CRF	**.939**	**.978**	**.936**	**.984**

Table 3.9: Evaluation of Syllabification Systems on Wiktionary (German) and CELEX (English).

	German	EPG64	FORB	PROS
# correct lines	3431	1098	1084	1564
# faulty lines	58	114	49	173
% faulty lines	1.70	9.41	4.32	10.0

Table 3.10: Size of manually annotated corpora with meter. Faulty lines denotes the number of lines where our automatic syllabification failed. Correct lines are used for experiments, since only there the gold annotation aligns.

3.8.2.1 English Prosody Datasets w/ Syllabification

The annotated corpora in English with prosodic annotation include: (1) The for-better-for-verse (FORB) collection[22] with around 1200 lines which was used by Agirrezabal et al. (2016a, 2019), and (2) the 1700 lines of poetry against which `prosodic`[23] (Algee-Hewitt et al., 2014; Anttila and Heuser, 2016) was evaluated (PROS). We merge these with our own (3) 1200 lines in 64 English poems (EPG64) that are discussed in Section 3.5.2. The first two corpora were already annotated for metrical syllable stress. These resources can be found at `https://github.com/tnhaider/metrical-tagging-in-the-wild/tree/main/data/English/SmallGold`.

Unfortunately, FORB does not contain readily available foot boundaries, and in PROS foot boundaries are occasionally set after each syllable. Additionally, FORB makes use of a `<seg>` tag to indicate syllable boundaries, so we do not derive the position of a syllable in a word. It also contains two competing annotations, `<met>` and `<real>`. The former is the supposedly proper metrical annotation, while the latter corresponds to a more natural rhythm (with a tendency to accept inversions and stress clashes). We only chose `<real>` when `<met>` doesn't match the syllable count (ca. 200 cases), likely deviating from the setup in (Agirrezabal et al., 2016a, 2019).

Table 3.10 shows the number of lines in each of these datasets and the number of lines that were incorrectly segmented by our best syllabification systems. Note that for the English data, between 5 and 10% of lines are incorrectly segmented (wrong number of syllables) by the best BiLSTM-CRF model, while the German Hybrid Ensemble method achieves an error rate under 2%. An error rate of 10% is certainly not perfect, but still substantially better than competing methods like `prosodic`, as discussed in Section 6.4.2.

[22] `https://github.com/manexagirrezabal/for_better_for_verse/tree/master/poems`

[23] `https://github.com/quadrismegistus/prosodic`

Diachronic Variation 4

4.1 Introduction

Literature is frequently tied to political, social, and intellectual-historical factors, poetry maybe even more so than other genres like long forms such as the novel (Moretti, 2005, cf. Figure 5). The aim of literary history is to get a view of such factors, of the movements in literary writing, and the traditions that emerged, and to explain them in their historical context.

In this chapter we investigate how, over the years, poetic traditions in multiple languages have fostered different literary movements and periods, and how we can find distinctive diction (word choice) that characterizes these movements and periods. In particular, we look for patterns how poetic language changed over time. We hypothesize that different literary periods and traditions used different means to compose poetry, and that we can extract this change in diction with unsupervised methods from distributional semantics and by tracking the frequency of formal linguistic features. We are especially interested in illustrating the popularity of poetic language features and word meaning over time, giving us a glimpse into how the aesthetics of poetry, but also how political and societal topics changed.

Still, characterizations of literary periods and their temporal boundaries are to a wide extent conventional. The boundaries of literary movements (when does one period end and another start) can be rather obscure and thus labels for literary periods should be understood as an auxiliary structure (Gigl, 2008, cf. chapter 4.4), in that these labels can help to gain insight, but should not be taken final and fixed. In any case, the analysis of diachronic patterns in poetry can provide insight on the aesthetic and socio-historical evolution of our literary heritage. Categorizing works of literature with regards to a certain period, or failing to do so, already gives us insight on aesthetic preferences or word choice in contrast to other temporal categories.

At the same time, natural languages change over time as they evolve to meet the needs of new generations of language users and their environment. Therefore, the language of a poet is not only a response to a particular tradition or fashion of the time period in which a poem is written. It also depends on the meaning of certain concepts and how their meaning changed over time. But when exactly does a word stop meaning one thing and start meaning something else? On the following pages, we try to approach such changes in word meaning in tandem with questions about literary history, since literary history is intricately linked to historical linguistics and language change.

Our focus in on distributional semantic methods, assuming that the meaning of words is determined by their context. Most words occur with each other only in certain contexts, and given a sufficient amount of data, we can infer paradigmatic and syntagmatic similarities between words.

In the first experiments we utilize 'Latent Dirichlet Allocation' models that were originally proposed by Blei et al. (2003). This family of methods allows us to find words that stand in each others company across documents. The intuition behind such models is that we can distill "what people talk about". See Figure 4.1, taken from Blei (2012), for an illustration behind the intuitions of Latent Dirichlet Allocation (it is an illustration and does not necessarily correspond to real data). It is assumed that some number

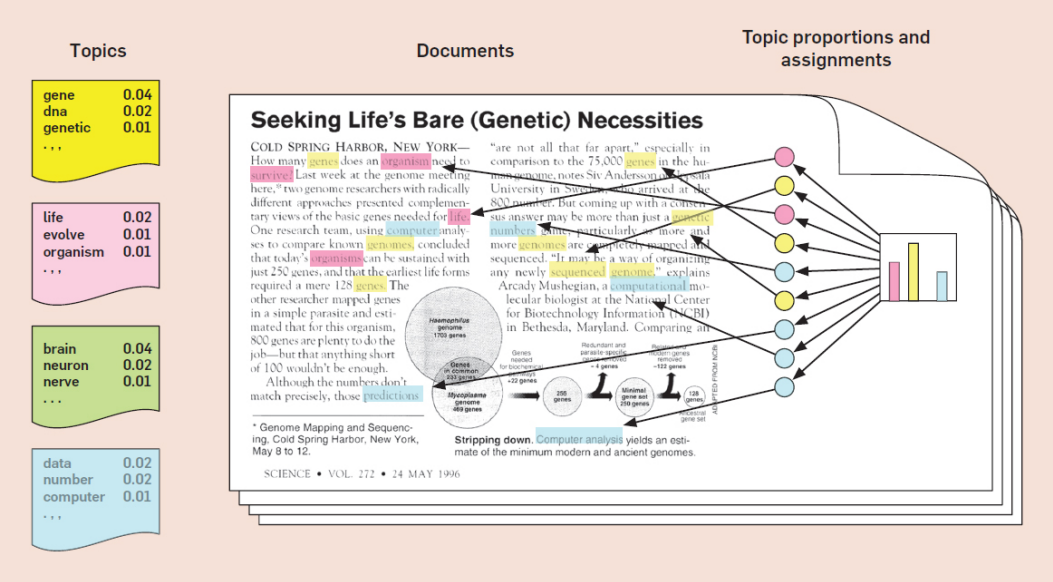

FIGURE 4.1: The intuitions behind latent Dirichlet allocation. Here illustrated on the topic composition in a text on genetics. Illustration taken from Blei (2012).

of 'topics', which are distributions over words, exist for the whole collection. Within a 'generative' modeling paradigm, each document is generated by first choosing a distribution over the topics (the histogram on the right), and then each word is weighted towards a topic assignment (the colored coins). Thus, every word belongs to every topic, albeit with varying degrees, following a normal distribution of weights.

A topic model creates a set of probability distributions over the vocabulary of the collection, which, when combined together in different proportions, best match the content of the collection. We can sort the words in each of these distributions by probability, take some most-probable words, and get a sense of what (if anything) the topic is 'about'. Each of the texts also has its own distribution over the topics, and we can examine these texts regarding a given topic to get a sense of how that topic is used.

The second family of methods that is used falls under the broad term of 'word embeddings'. With these models we want to understand the contextual meaning of words and how this meaning changes over time and in relation to

other words. The basic workings of learning word embeddings is illustrated in Figure 4.2. In our experiments, we use a variant of word2vec with skip-gram, where the task consists of predicting the (the vector of the) context words for every input word for which we desire an embedding. The competing method, CBoW (Continuous Bag of Words), would try to predict the word from its context. Notably, these methods capture not only paradigmatic similarity (substituting near-synonyms), but also syntagmatic similarity (which words occur with each other).

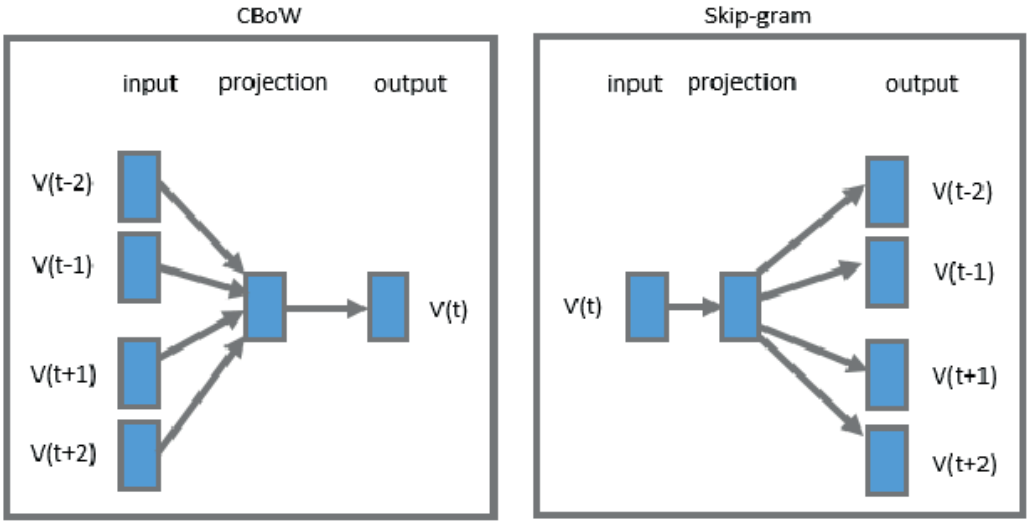

Figure 4.2: Word embeddings via cbow vs. skip-gram (with negative sampling).

We develop a method to explore poetic tropes, i.e., word pairs such as 'love (is) magic' that gain association strength (between the constituent words) over time, finding that most are gaining traction in the Romantic period. Further, we track the self-similarity of words, both with a change point analysis and by evaluating 'total self-similarity' of words over time (the similarity of the meaning of a word to itself over time). The former helps us to reconstruct literary periods, while the latter provides us with further evidence for the law of linearity of semantic change (Eger and Mehler, 2016) using our new method. To that end, we use a model that learns diachronic word2vec embeddings

jointly over all our time slots (Bamman et al., 2014a). With this model, we circumvent problems of aligning vector representations of words over time, which was previously done e.g., by training embeddings for every time slot and then aligning word vectors on second order.

In comparison to the following chapters, these approaches operate on the level of word forms (or lemmas), rather than on more sophisticated annotation. This provides us with a first overview of the corpora, without having to do much processing or learning of poetic features beforehand. However, once we have established other variables, like emotions (see Chapter 5), or prosody (see Chapter 6), we will illustrate their diachronic variation, too.

This chapter on diachronic change is structured as follows: After (i) we discuss some related work on language change and topic models, we give (ii) a quick introduction to German literary periods on the basis of crowd-based annotation. Then (iii), we track topics through a Latent Dirichlet Allocation (LDA) over time that allow us to outline literary periods, first with a focus on German, and then in comparison to English, Czech, and Russian. Next, (iv) we experiment with using LDA to generate features for texts to predict the publication year of a poem (i.e., dating). Finally, (iv) we build a diachronic word embedding model to track the semantic distance of words to each other over time to i.a., determine the emergence of metaphors/tropes.

4.2 Related Work

Underwood and Sellers (2012) tracked the evolution of diction in different literary genres. By tracking the ratio of words that entered the English vocabulary before 1150 CE vs. the vocabulary that entered between 1150 and 1699 CE, they find that a specialized literary language formed over the course of the late 18th and then 19th century: "while all genres of writing tended to adopt a more learned diction in the eighteenth century, poetry, drama, and fiction decisively reversed course in the nineteenth." By the end of the nineteenth century, poets had developed a specialized diction, inherited largely from the period before Middle English was a written language, indicating that poetry reverted to a more colloquial style.

More commonly used in natural language processing research that is interested in diachronic variation, and also in this chapter, are methods that were developed in the field of lexical semantic change. A variety of methods have been applied to the task of measuring lexical semantic change, ranging from the use of statistical tests in order to detect significant changes in the distribution of terms from two time periods (Popescu and Strapparava, 2013; Cook and Stevenson, 2010), to training distributional similarity models on time slices (Gulordava and Baroni, 2011; Sagi et al., 2009), and neural language models (Kim et al., 2014; Kulkarni et al., 2015). Other work (Mihalcea and Nastase, 2012) takes a supervised learning approach and predicts the time period to which a word belongs given its surrounding context. Besides models for lexical semantic change, core linguistic research has investigated diachronic change, like Hartmann (2014) on word-formation patterns like the -ung suffix in German.

Previous work has investigated the evolution of topics over time. There has been research with topic models on poetry with Latent Dirichlet Allocation, e.g., Navarro-Colorado (2018b) explores the overarching topical

motifs in a corpus of Spanish sonnets, or Jockers and Mimno (2013) looks at significant themes in 19th century literature (with a focus on fiction).

Bayesian models (to which topic models belong) have been developed for various tasks in lexical semantics (Frermann and Lapata, 2016), but their concern is more to disambiguate polysemous words over time (topic=word sense), rather than tracking the change of topics themselves. Dynamic topic models (Blei and Lafferty, 2006; Wang et al., 2012) were proposed. These models track the change of topics by modeling changing topics, such that the topics themselves change (where each time period is assigned n topics).

We also track the evolution of certain topics that emerge cross-lingually. Cross-lingual topic models also have gained popularity (Zhang et al., 2010; Gutiérrez et al., 2016; Bianchi et al., 2020), but these typically need parallel data. Instead, we follow a simpler approach that relies on manual interpretation and translation of the topics, but already offers insight on literary history.

Lexical semantic change has been explored in various works in recent years with a focus on studying laws of semantic change, even though the search for universal laws can be problematic with frequentist distributional semantic models from a methodological standpoint (Dubossarsky et al., 2017). Xu and Kemp (2015) explore two earlier proposed laws quantitatively: the law of differentiation (near-synonyms tend to differentiate over time) and the law of parallel change (related words have analogous meaning changes), finding that the latter applies more broadly. Hamilton et al. (2016) find that frequent words have a lower chance of undergoing semantic change and more polysemous words are more likely to change their meaning. Hamilton et al. (2016) e.g., point out that the word 'gay' changed it's meaning from exclusively meaning 'jolly' and 'happy' in the late 19th, early 20th century, to 'homosexual' in the late 20th century. Eger and Mehler (2016) find that semantic change is linear in two senses: semantic self-similarity of words tends to decrease linearly in time and word vectors at time t can be written as linear

combinations of words vectors at time $t-1$, which allows to forecast meaning change. Regarding methods, Xu and Kemp (2015) work with simple distributional count vectors, while Hamilton et al. (2016) and Eger and Mehler (2016) use low-dimensional dense vector representations. Both works use different approaches to map independently induced word vectors (across time) in a common space: Hamilton et al. (2016) learn to align word vectors using a projection matrix while Eger and Mehler (2016) induce second-order embeddings by computing the similarity of words, in each time slot, to a reference vocabulary. Kutuzov et al. (2018) survey and compare models of semantic change based on diachronic word embeddings. Dubossarsky et al. (2017) caution against confounds in semantic change models. Schlechtweg et al. (2018) propose a method how to annotate distinct senses of meaning of words as these senses change over time: words may gain or lose certain senses over time. The data from Schlechtweg et al. (2018) was used in a shared task in which we (Rother et al., 2020) participated to cluster distinct diachronic word senses.

4.3 Characterization of Literary Periods

For first insight on literary periods, in Figure 4.3 we present a plot from a crowd-annotation of literary periods in the Antikoerperchen Corpus (for the corpus see section 3.5.1). The labels are not entirely standardized, as they were collected in a crowd effort (and only crawled by us). Also, the corpus is only 150 poems large, but we can clearly see many literary movements and periods from 1600 to 1950 CE. By going through the most important periods, we will outline certain hypotheses that shall be approached with the methods in the following sections. We follow roughly the suggestions of Gigl (2008) in characterizations of literary periods.

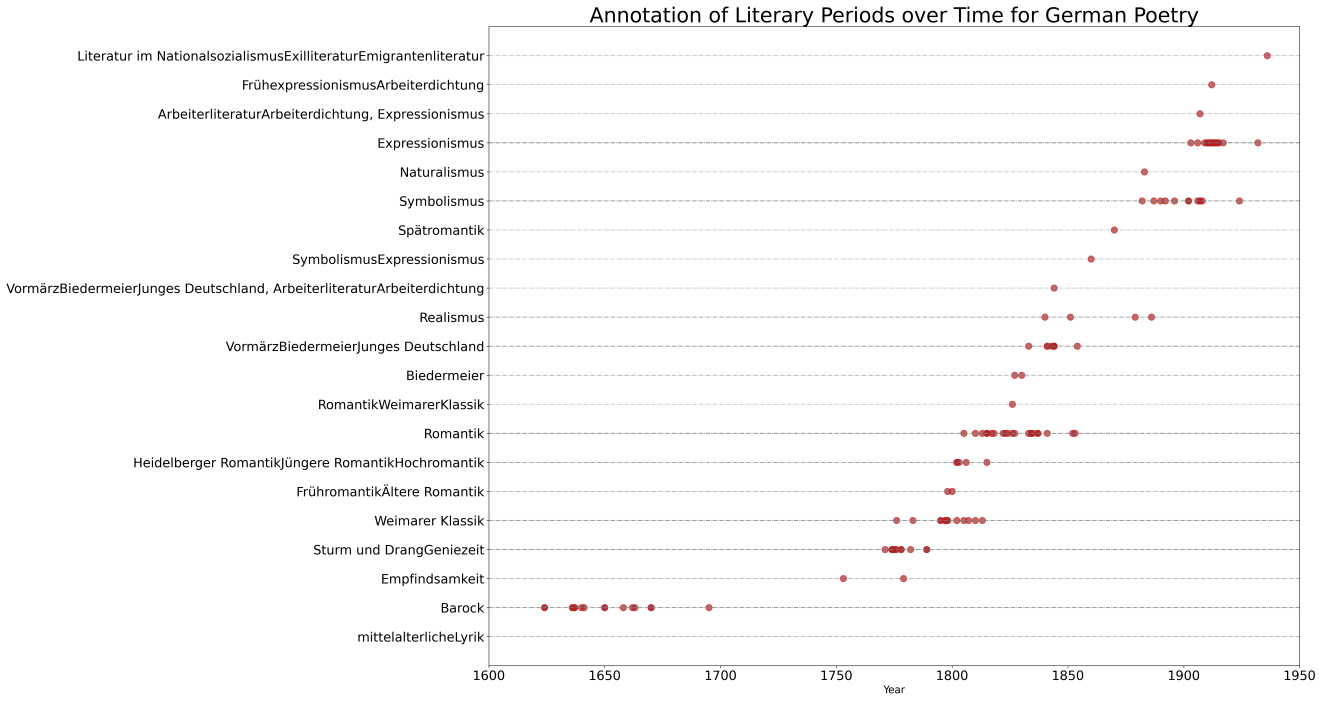

FIGURE 4.3: Annotation Literary Periods in ANTI-K

Our annotation for the **'Barock' (Baroque)** period ranges roughly from 1625 to 1700. In a broader sense, especially with regard to the importance of Baroque literature and philosophy, the Baroque can be understood as an

epoch of European intellectual history. We hypothesize that Barock literature has a focus on topics of death, power, and hedonism. The concept of death was understood differently than how it was understood later on.

The subsequent period, **'Aufklärung' (Enlightenment)** is not present in this annotation. See the gap between 1700 and 1750. Gigl (2008) mentions that the focus of (German) enlightenment literature is on problems of morality and virtue. We therefore should see those later in the large corpora. More generally, important characteristics of the enlightenment are the invocation of reason as the universal authority of judgment, which included the fight against prejudice and the turn to the natural sciences. In sociopolitical terms, the Enlightenment aimed at more personal freedom of action (emancipation), education, civil rights, general human rights, and the common good as a duty of the state.

The period of **'Empfindsamkeit' (Sensibility)** is only present with two poems (by Klopstock) in our annotated sample, from 1755 and 1780 CE respectively. 'Sensibility' (Empfindsamkeit) refers to a tendency in the European Enlightenment that lasted from about 1700 until the French Revolution (1789-1799). The origins of Empfindsamkeit are largely religious; examples can be found in the emotionally colored texts for the oratorios and cantatas of Johann Sebastian Bach, or in the poems of Friedrich Gottlieb Klopstock. According to Lamping (2016), until the 18th century, all production of lyric poetry was still firmly bound to the traditions of ancient Greek and Roman lyric poetry. We can see this in the use of the hexameter and odic stanza forms.

We can see the periods **'Sturm & Drang'** and **'Weimarer Klassik' (Weimar Classicism)** at the end of the 18th and beginning 19th century, Goethe and Schiller being the most popular contributors to both. 'Sturm und Drang' was a proto-Romantic movement in German literature. Within the movement, individual subjectivity and, in particular, extremes of emotion were given free expression in reaction to the perceived constraints of rationalism imposed by the Enlightenment and associated aesthetic movements.

'Weimar Classicism' was a German literary and cultural movement, whose practitioners established a new humanism with a synthesis of ideas from Romanticism, Classicism, and the Age of Enlightenment. Temporally, according to our annotation, this latter period of 'Weimarer Klassik' heavily bleeds into 'Romantik' (Romanticism), while 'Sturm & Drang' essentially ends with the onset of Romanticism. We therefore assume that these periods will show up in our analysis within 'early romanticism', but being distinguishable from romanticism through specific word choice and diction.

The period of **'Romanticism'** begins around 1770/1800 CE, and ends around 1870. Romanticism encompasses by far the longest period, which lead to multiple sub-periods in our annotation. It ranges from 'Frühromantik' (Early Romanticism) over 'Hochromantik' (High Romanticism) and 'Heidelberger Romantik' (Heidelberg Romanticism), into 'Spätromantik' (Late Romanticism). We hypothesize that romantic poetry is concerned with a glorified representation of beauty, but also with the supernatural, with emotionalism, and a focus on nature.

Writing of the Modern period (Modernity) is represented in our annotation with the two sub-periods of **'Symbolismus' (Symbolism)** (ca. 1875–1925) and **'Expressionismus' (Expressionism)** (ca. 1900–1930). **'Realismus' (Realism)** is the only period that streches from Romanticism into Modernity. We hypothesize that the language of symbolism will be hard to distinguish from core romantic diction, but expressionism should present language that is strongly influenced by concepts of war, industrialism, and societal problems (see for example the poetry of Engelke, Heym, or Trakl).

4.4 Topic Evolution in Poetry

We approach diachronic variation of topics in poetry from two perspectives. First, as distant reading task to visualize the development of interpretable topics over time in German poetry and in the next section in comparison to English, Czech and Russian. Subsequently, we use these topics for a downstream task, i.e., supervised machine learning task to determine the year (the time-slot) of publication for a given poem.

Statistical topic models are increasingly and popularly used by Digital Humanities scholars to perform distant reading tasks on literary data, including poetry (Navarro-Colorado, 2018b; Hettinger et al., 2016). Topic models are usually unsupervised and therefore less biased toward human-defined categories. They are especially suited for insight-driven analysis, because they are constrained in ways that make their output interpretable. Although there is no guarantee that a 'topic' will correspond to a recognizable theme or event or discourse, they often do so in ways that other methods do not. Their easy applicability without supervision and ready interpretability make topic models good for exploration. Especially Latent Dirichlet Allocation (LDA), originally proposed by Blei et al. (2003), has shown its usefulness, as it is unsupervised, robust, easy to use, scalable, and it offers interpretable results. However, being an unsupervised and exploratory method, LDA models are hard to evaluate, as changes on parameters (like the number of topics, or changes in the corpus), may substantially change the resulting topic compositions. Still, we replicated our results on multiple corpora and different languages. In the first experiment, we apply LDA to the German Textgrid corpus (see section 3.3). Later experiments on the full German corpus (DLK) showed similar topics, as can be also seen in the follow up experiments in section 4.4.2, which were carried out on the larger DLK corpus.

We use Latent Dirichlet Allocation for a visualization of topic trends in a mono-lingual and a cross-lingual setting, illustrating the similarities and dis-

parities between different poetic traditions and literary periods. Our method is largely based on reading and translating topic distributions and finally interpreting the trajectories of relative topic importance against the backdrop of literary history. Our method to visualize topic trends is comparatively simple to what is usually proposed in the computational literature. And certainly, this brings with it some problems, such as the manual labor involved in interpreting and translating topics, which can lead to the inclusion or the omission of certain interpretative material. There is no ready solution to eliminate human error. Thus, the results presented in this chapter have exploratory character, but were documented and scrutinized to the best of our knowledge.

To investigate diachronic topic evolution, we train the model over the whole corpus laterally/horizontally, assuming that topics that are important for certain literary movements are also prominent in the whole corpus. We then split the corpus at defined time stamps and project the topics in the documents in each time slot vertically. This generously assumes that, if a topic is important for a certain time slot, there is a sufficient amount of documents for which this topic is at least moderately central, and that the significant topics for that period were found in the first place. But this method also makes it easy to track a certain topic over time, since the specific topic composition does not change over time, as would be the case when training models for individual time slots (or via a rolling window).

4.4.1 German Diachronic Topics

To discover trends of German poetic topics over time, we bin the poems of the Textgrid corpus into time slots of 25 years width each.[1] See Figure 4.4 for a plot of the number of documents per bin (Textgrid in green, plus the yellow duplicates), as it was introduced in Chapter 3. The chosen binning slots offer a fair amount of documents per slot for our experiments. Wider

[1]This first experiment is done on Textgrid, since at the time of experimentation, DLK was not finished yet.

slots (e.g., 50 year bins) would obfuscate the granularity of the data, while slimmer slots (e.g., 10 year bins) would lead to sparse data in certain time slots. The poems of Textgrid are not as representative as the full DLK corpus, but models trained on the latter corpus resulted in similar topics, but some topics like 'virtue, arts' in the Enlightenment period were not visible anymore on the larger corpus (likely because DTA contains poems with different topics in that time period).

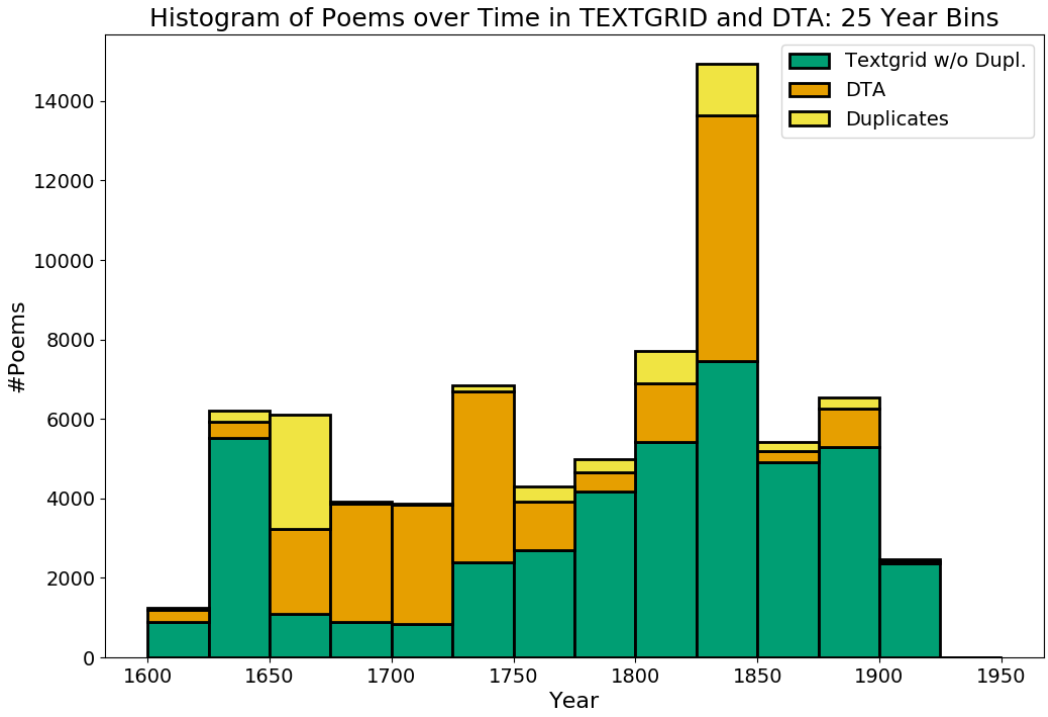

Figure 4.4: DTA and Textgrid Poems in 25 Year Bins. Identified duplicates are subtracted from Textgrid.

We use the implementation of Latent Dirichlet Allocation (LDA) as it is provided in genism (Rehurek and Sojka, 2011). We transform our documents (of wordforms) to a bag of words representation, filter stopwords (function words), and set the desired number of topics=100 and train for 50 epochs with a decay=0.5 to attain a reasonable distinctness of topics. We choose 100 topics (rather than a lower number that might be more straightforward to interpret) as we want to later use these topics as features for downstream

tasks, and we aim for more fine grained topics, rather than risking to mix certain topics together. We find that wordforms (instead of lemma) are useful for topic models of poetry (at least in German), as they capture style features like rhyme ('Mund' (mouth), 'Grund' (ground, cause), 'rund' (round)). Rhyme words often cluster together (as they stand in proximity).We also find clusters with orthographic variations ('Hertz' instead of 'Herz') that provide a strong diachronic signal.

We retrieve the most important (likely) words for all 100 topics and interpret these (sorted) word lists as aggregated topics, e.g. Topic 27 (Figure 4.5) contains: Tugend (virtue), Kunst (art), Ruhm (fame), Geist (spirit), Verstand (mind) and Lob (praise). This topic as a whole describes the concept of 'artistic virtue'.

To visualize trends of singular topics over time, we aggregate all documents d in slot s and add the probabilities of topic t given d and divide by the number of all d in s.

$$p(t|s) = \frac{\sum p(t|d)}{n(d,s)} \tag{4.1}$$

This gives us the average probability of a topic per timeslot. We then plot the trajectories for each single topic (importance of topic for time slot).

In the following, we will show selected plots that caught our attention because of their distinctive shapes. Some of these topics are also important for a temporal classification (by information gain), as is shown in Section 4.4.1.1. See Figures 4.5 –4.15 for a selection of interpretable topic trends. Please note that the scaling on the y-axis differ for each topic, as some topics are more pronounced in the whole dataset overall.

Topic 27, 'artistic virtue' (virtue, art, glory, spirit, wit, praise, kind), see Figure 4.5, shows a sharp peak in importance around 1700–1750, outlining the importance of that topic for the period of Enlightenment. This shows that poetry from the Enlightenment period is present in the Textgrid corpus,

even though the small corpus (ANTI-K) is lacking annotation for those. This topic highlights the prevalence of artistic virtue and associated concepts for poetic writing in the Enlightenment period, and confirms the hypothesis based on Gigl (2008), such that the concepts of 'morality' and 'virtue' are quite prominent and central in this period.

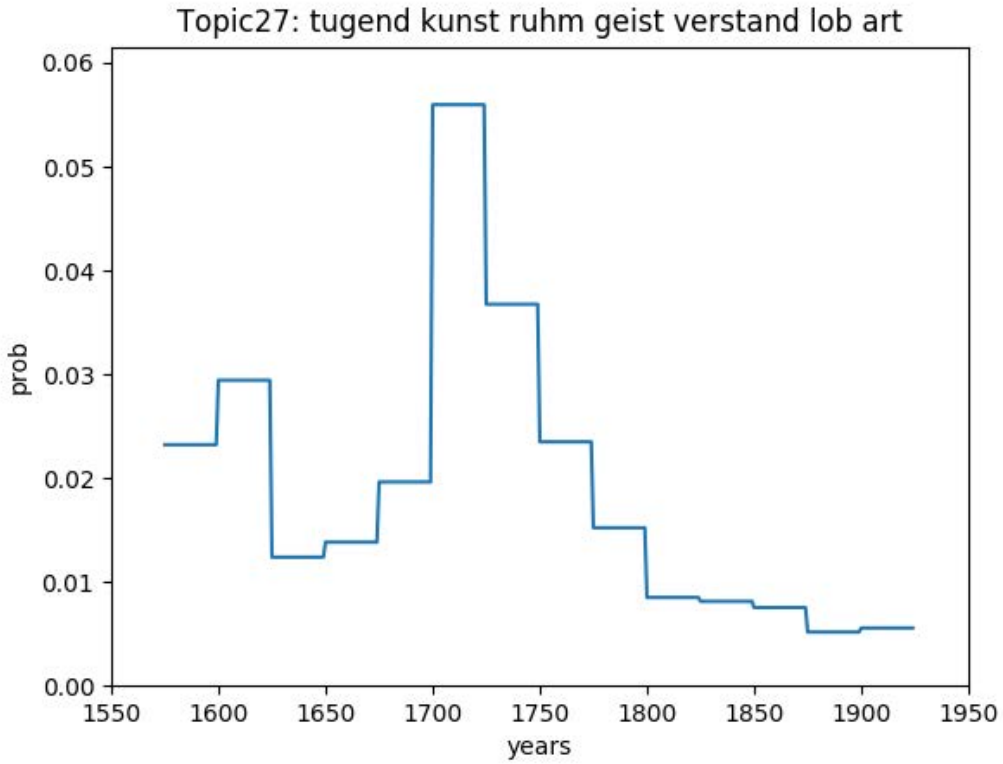

Figure 4.5: Topic 27 'Virtue, Arts'
(Period: Enlightenment)

Several topics indicate Romanticism, such as 'flowers, garden, roses' (Figure 4.6), 'singing, song' (Figure 4.7) or 'dust, ghosts, temple, altar, depths' (Figure 4.8). The topics themselves seem to be stereotypical for the romantic period, with positive and idyllic topics around flowers and singing, but also a more gloomy and mysterious topic centered around the imagery of 'crypts'. What unifies these particular topics is that their onset is around 1750 CE,

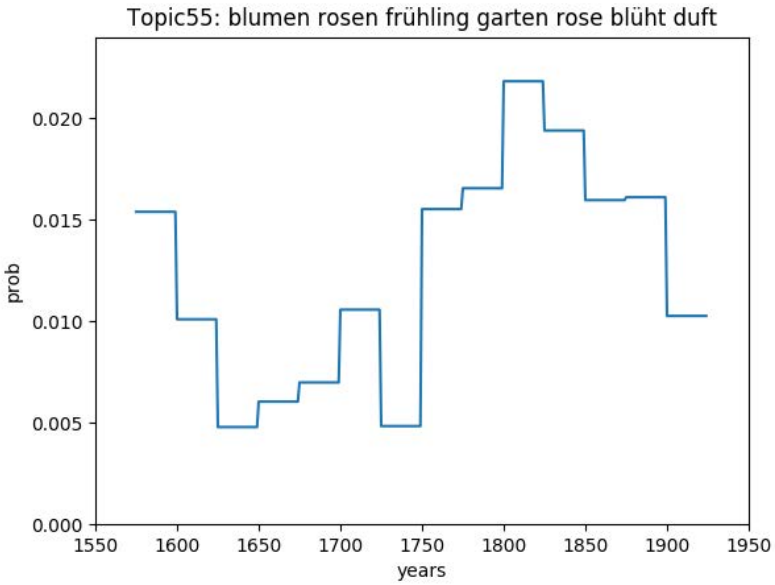

FIGURE 4.6: Topic 55 'Flowers, Spring, Garden'
(Period: Early Romanticism)

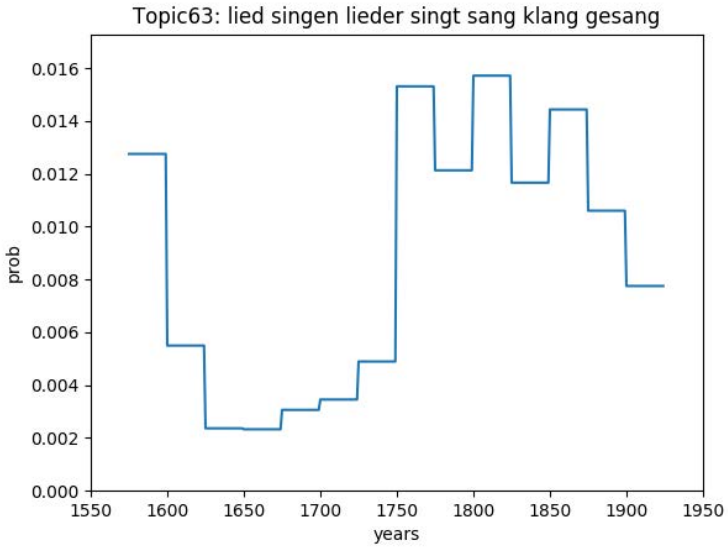

FIGURE 4.7: Topic 63 'Song' (Period: Romanticism)

and that they lose importance before the 20th century, clearly delineating the respective literary period.

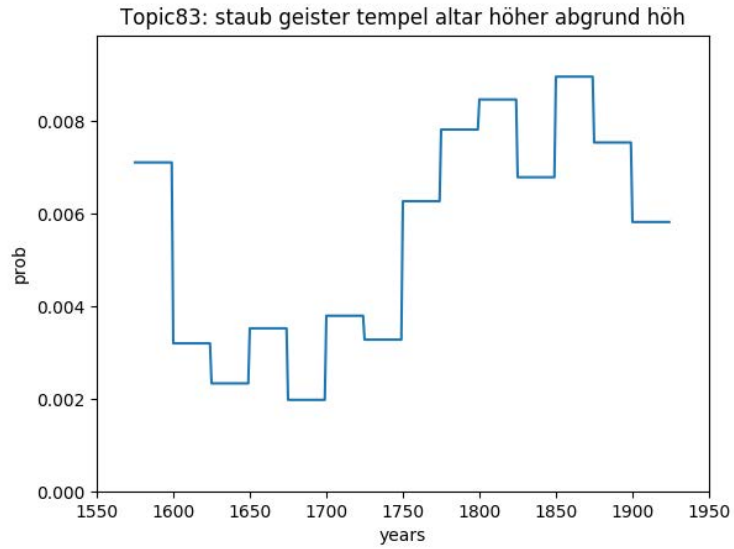

Figure 4.8: Topic 83 'Staub Geister Tempel' (Period: Romanticism)

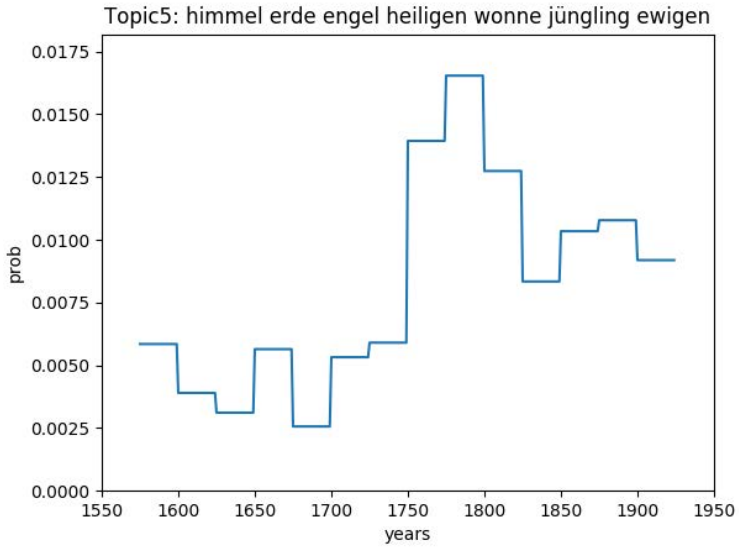

Figure 4.9: Topic 19 'Heaven, Depth, Silence' (Period: Sturm und Drang, Weimarer Klassik). Among the most informative topics for temporal classification.

Topic 19 'Heaven, Depth, Angel, Silence' (Figure 4.9) falls in the pre-romantic period of 'Sturm & Drang' and 'Weimarer Klassik', as seen with the

rather sharp peak between 1750 and 1800 CE. But it is not a core romantic topic, since it already loses importance before 1825. This topic indicates that religion and divinity are not core romantic topics, but that this is rather found in earlier periods.

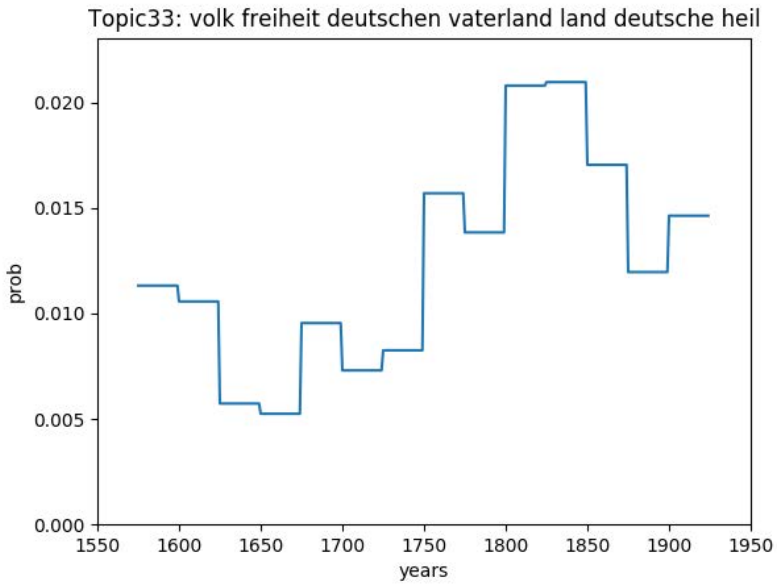

FIGURE 4.10: Topic 33 'German Nation'
(Period: Vormärz, Young Germany))

The period of 'Vormärz' or 'Young Germany' is seen rather clearly in the topic 'German Nation' (Figure 4.10), showing the rise of the German national movement (towards democracy and the republic) in the early 19th century. The most prominent words include 'Volk' (peoples), 'Freiheit' (freedom), 'deutsche(n)' (German), 'Vaterland' (fatherland/homeland), However, its trajectory is hardly distinguishable from romantic topics, only that is ends quicker. Note that the topic again grains traction in the 20th century, attributable to the rise of nationalism. Closer study of the data revealed that there are no poets that wrote exclusively for this 'movement'. Instead, many poets raised their voice to speak about German politics.

We find that the topics 'Beautiful Girls' (Figure 4.11) and 'Life & Death' (Figure 4.12) were always quite present over time, while 'Girls' is more prounounced in Romanticism, and 'Death' in the Baroque period. This indicates that the fair sex was always central to young male poets, especially that love poetry is more prominent in romantic poetry, notably in an idealized fashion, where a sleeping girl and here eyes allow more surface for the projection of ideas, rather than being in an actual dialog with her. Compare the characterization of Ottilie in Goethe's Wahlverwandschaften (Benjamin, 2016, orig. 1922). On the other hand, 'Life & Death' are pervasive to any artform, and especially poetry. The poets from the Baroque period had a particular taste for dying things and burials. It should be noted, perhaps, that against the backdrop of the Thirty Years' War (1618–1648), people's everyday lives were dominated by violence and destruction. Motifs of the period deal with the resulting widespread fear of death and its effects in different ways (Beutin et al., 1994).

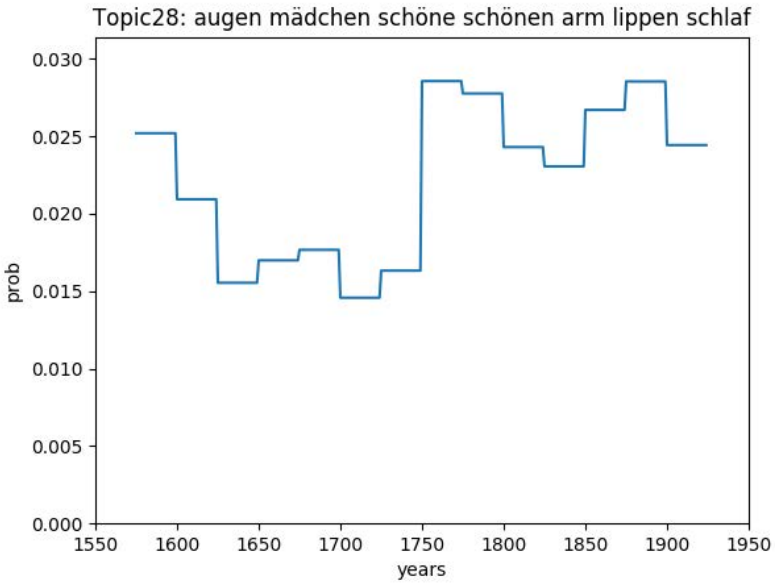

Figure 4.11: Topic 28 'Beautiful Girls, Sleep, Bodyparts' (Period: Omnipresent, Romanticism)

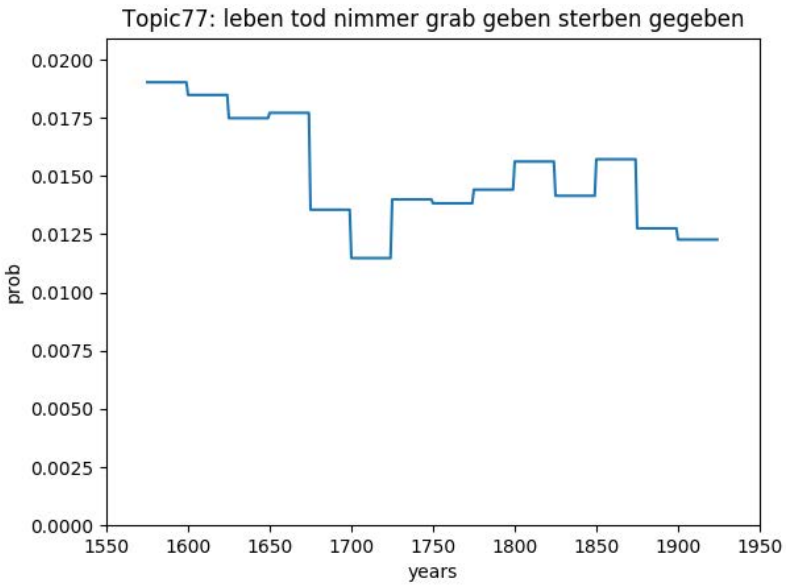

FIGURE 4.12: Topic 77 'Life & Death' (Period: Omnipresent, Barock)

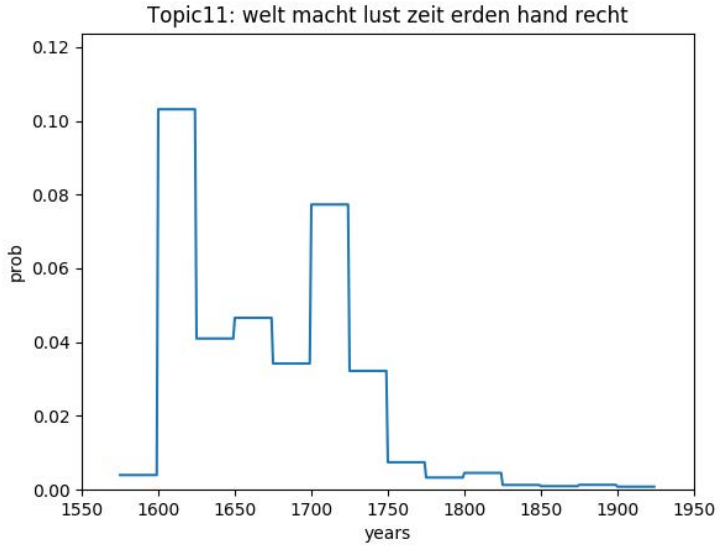

FIGURE 4.13: Topic 11 'World, Power, Lust, Time' (Period: Barock). One of the most informative topics for classification.

Some topics, that are most informative for the temporal classification task (as elaborated below), demarcate the period of 'Barock' (baroque period). Among these is Topic 11: 'World, Power, Lust, Time' (Figure 4.13), which is clearly a Baroque topic, as it ends at 1750. It seems to be a very prototypical

topic for the period, with a focus on 'wordly' matters such as hedonic pleasure (lust) and power, rather than heavenly ones.

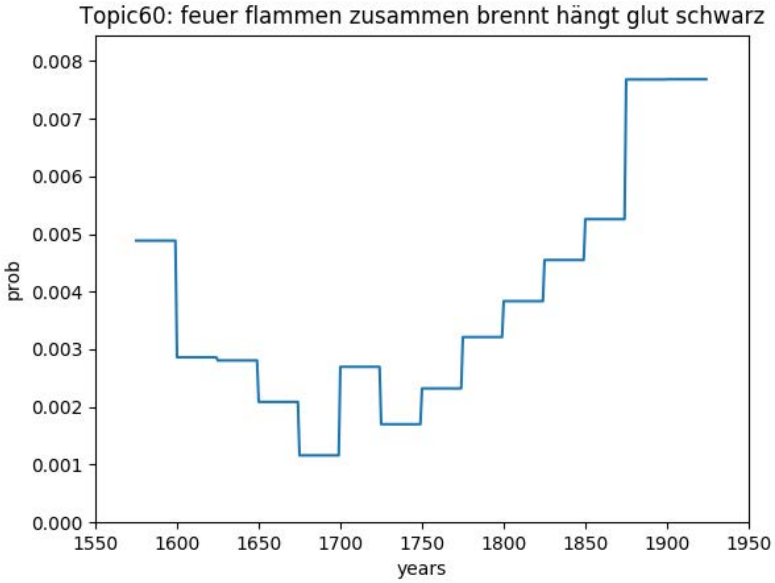

Figure 4.14: Topic 60 'Fire, Flames' (Period: Modernity)

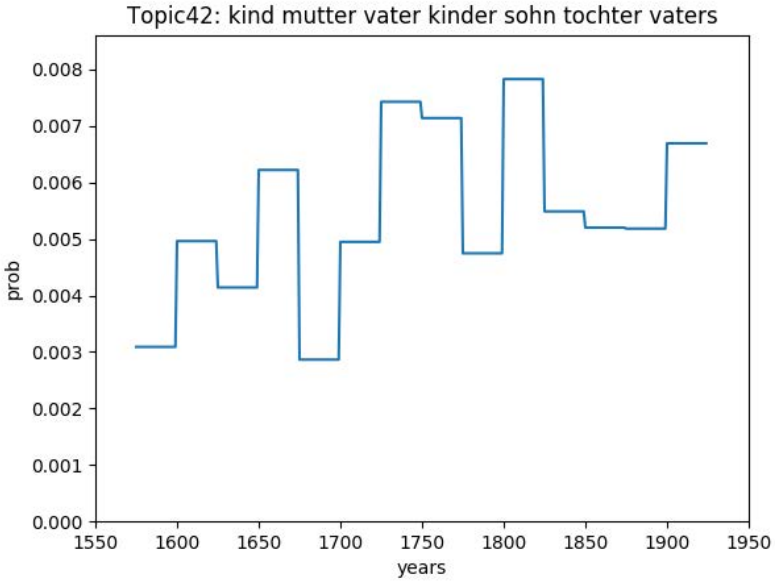

Figure 4.15: Topic 42 'Family' (no period, fluctuating over time)

We find that the topic 'Fire & Flames' (Figure 4.14) steadily rises into Modernity, though it is not very prominent overall. Finally, the topic 'Family'

(Figure 4.15) shows wild fluctuation over time, being more or less present over the whole time frame, but not in high concentration ($p \sim 0.006$). Topics such as these are not indicative of any literary periods, but show that there are also topics that poets talked about things not necessarily of importance for their time.

4.4.1.1 Predicting Time Periods and Authorship for German

A reliable system that can accurately assign a time stamp (e.g., a year) to a given poem could have huge potential. While dating poetry might not rank with the prestige of the famous radiocarbon dating (C-14) method, it could help to categorize newly found textual sources. To test whether topic models can be used for dating poetry or attributing authorship, we perform supervised classification experiments with Random Forest Ensemble classifiers. We find that we obtain better results by training and testing on stanzas instead of full poems, as we have more data available. Also, we use 50 year slots (instead of 25) to ease the task (in total seven time slots).

As a baseline, we extract rather straightforward style features, such as line length, poem length (in token, syllables, lines), cadence (number of syllables of last word in line), soundscape (ratio of closed to open syllables, see (Hench, 2017)), and as proxy for meter, the number of syllables of the first word in the line (monosyllabic first words indicate a iambic line, while bisyllabic words indicate a trochaic line, since bisyllabic German words are largely trochaic, and unstressed function words are typically monosyllabic).

We split the data randomly into 70:30 training:testing, which has a better performance (5 points) over a 50:50 split. We then train Random Forest Ensemble classifiers and perform a grid search over their parameters to determine the best classifier. Please note that our class sizes are quite imbalanced (where most poems are around the romantic period, and fewer in the pre-

romantic period). However, a Random Forest Classifier is good at handling such imbalance.

	Style	LDA	Style+LDA
7 Time Slots Stanzas	.83	.89	**.90**

Table 4.1: Dating Poetry: Diachronic Classification (Random Forest)

The Style baseline achieves an Accuracy of 83%, LDA features 89% and a combination of the two achieves 90%, when training and testing on stanzas. However, training on full poems reduces this to 42—52%. This is most likely due to the increased number of training instances. For authorship attribution, we also use a 70:30 random train:test split and use the author name as class label. We only choose the most frequent 180 authors. We find that training on stanzas gives us 71% Accuracy, but when trained on full poems, we only get 13% Accuracy. It should be further investigated if this is only because of a surplus of data. As seen in Table 4.1, for temporal classification we find that a simple style baseline on eight features gives us almost as good performance as with topics.

The most informative features for temporal classification (by information gain) are the following:

1. Topic 11 ('Welt, Macht, Lust, Zeit') (.067)
   ('World, Power, Lust, Time')
2. Topic 37 ('Hertz, Gantz, Hertzen, Augen, Himmel, Geist') (.055)
   ('Heart (old spelling), Whole/Entire (old spelling), Hearts (old spelling), Eyes, Heaven, Spirit')
3. Number of Syllables Per Line (.046)
4. Length of poem in syllables (.031)
5. Topic 19 ('Himmel, Tief, Empor, Stille') (.029)
   ('Sky, Deep, Up, Silence')

6. Topic 98 (‘Sah, Ward, Kam, Stand’) (.025)
   (‘Saw, Became, Came, Stood’)
7. Topic 27 (‘Tugend, Kunst, Ruhm, Geist’) (.023)
   (‘Virtue, Art, Fame, Spirit’)
8. Soundscape (Ratio open vs. closed syllables) (.023)

Table 4.2 shows results of an experiment for predicting the year of publication via regression. We use Lasso regression on the basis of topic probabilities in poems and compare it to a MLP regressor on top of document level (CLS token) BERT embeddings (Devlin et al., 2019). The BERT model that was pretrained on historical data[2] explains far more variance in a regression over years. Furthermore, a Lasso regression on 100 topics (the previously discussed model) misses the target on average by around 45 years (mean absolute error), while BERT is only off by 34 years on average. Also, it should be noted that a BERT base model that was pretrained only on contemporary data such as wikipedia and legal text[3] is not able to explain any variance at all.

Algorithm	Features	R^2	Mean Absolute Error
Lasso	Topics	.34	45 Years
MLPRegressor	BERT	.52	34 Years

Table 4.2: Dating Poetry: Diachronic Regression

See Figure 4.16 for a scatter plot, showing a comparison of the real year annotation on poems versus what the Lasso regression model on topics predicted. We can see quite a difference in what the model should predict and what it actually predicts. Especially the early modern texts (before 1600) are off by around 150 years (into the future). We can see from this plot, that this method is far from reliable, but that it is able to more or less distinguish between pre-romantic and romantic texts, where romantic texts are rarely predicted to be before 1750, but the pre-romantic texts show more variance.

[2] `https://huggingface.co/redewiedergabe/bert-base-historical-german-rw-cased`
[3] `https://huggingface.co/bert-base-german-cased`

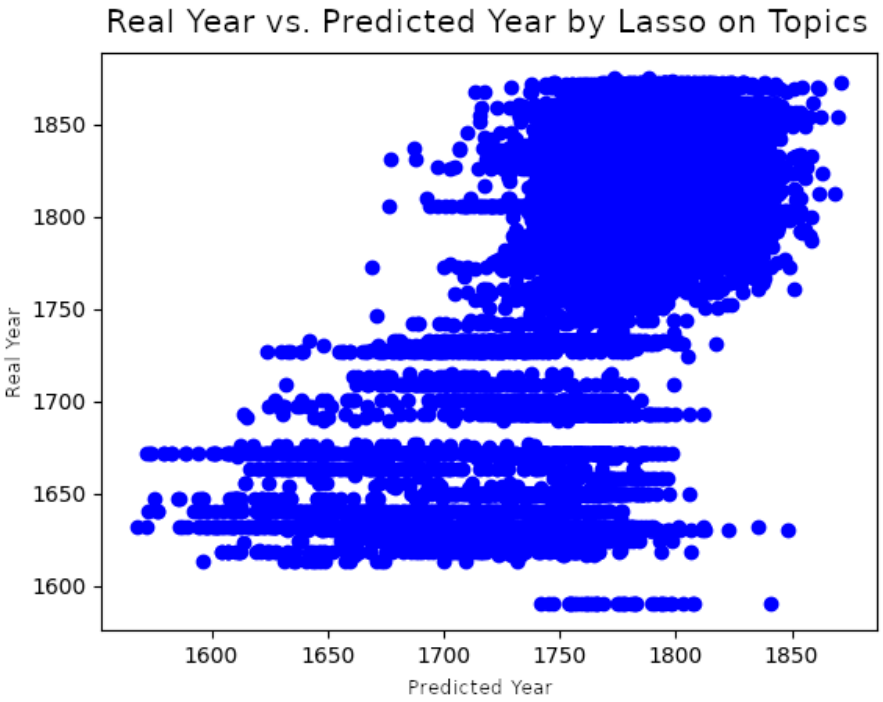

Figure 4.16: Lasso Regression Model on Topics. Real Year vs. Predicted Year. Each dot represents one poem.

4.4.2 Cross-lingual Topic Evolution

In the previous section, we have seen experiments on diachronic topic variation for German poetry. In this section, we re-implement this methodology in a multilinguial setting, in particular for German, English, Russian and Czech poetry. We hope to see the same (or similar) topics for German, even though the corpus is a bit larger (now also including DTA), and how such topics translate to topics in poetry of other languages, and how the importance of these topics compares over languages: Are the same topics present, and if so, did they gain in popularity at an earlier or later time?

To determine the evolution of topics across poetic traditions, we collect four poetry corpora in Czech, Russian, German and English. See Table 4.3 for an overview, and where the corpora were mined from. The German and English corpora were introduced in Chapter 3. The Czech and Russian corpora were supplied by Petr Plechac, and since I speak neither of these languages, the task of translation was carried out by him. The used corpora are contaminated with foreign language poems and we filter these with `langdetect`.[4]

Language	Poems	Tokens	Comment
Czech	$\sim 80k$	15M	Corpus of Czech Verse (`http://versologie.cz`)
Russian	$\sim 18k$	2.7M	Poetic subcorpus of Russian National Corpus (`http://ruscorpora.ru`)
German	$\sim 74k$	12M	German Poetry Corpus v3, see Chapter 3.
English	$\sim 85k$	22M	Project Gutenberg, see Chapter 3.

Table 4.3: Diachronic Poetry Corpora for Multilingual Topic Analysis

In Figure 4.17 and 4.18 you can see the distribution of poems over time for all four corpora. It is apparent that the German corpus offers the best coverage of pre-romantic poetry, while English does offer a little bit of data before 1750 CE (albeit not too much). The Czech and Russian corpora begin only around 1770 CE.

[4] `https://pypi.org/project/langdetect/`

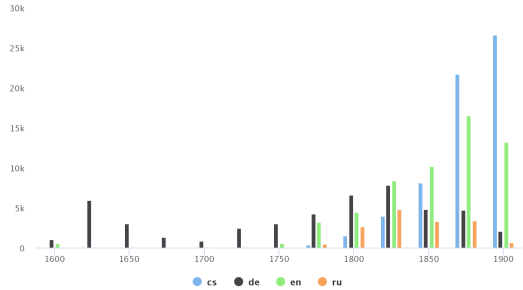

Figure 4.17: Size of Corpora over Time

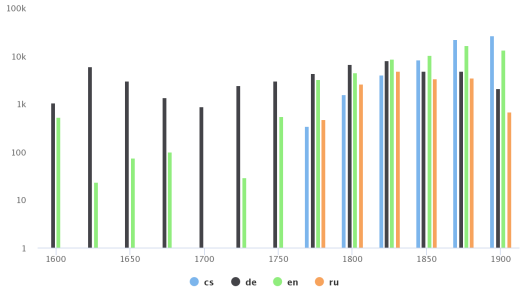

Figure 4.18: Size of Corpora over Time; log(Size) at y-axis

We transform our documents to a bag of words representation. As we deal also with highly inflected languages (Czech, Russian), lemmas were used instead of word forms. For lemmatization and POS-tagging of English and German texts we use the TreeTagger (Schmid, 1994), for lemmatization and POS-tagging of Czech texts we use the MorphoDita (Straková et al., 2014), for lemmatization of Russian texts we use the MyStem (Segalovich, 2003). In Czech, German, and English all the parts-of-speech except for nouns, adjectives, and verbs were filtered out, to minimize the influence of meaningless function words. In Russian, the list of stopwords is provided by the NLTK library and manually extended by us.

We set the number of topics=100 and train for 100 epochs (passes) to attain a reasonable distinctness of topics. We choose 100 topics as previous research on poetic topics (Haider, 2019; Navarro-Colorado, 2018b) determined this parameter to be be optimal for distant reading of poetry.

4.4.2.1 Experiment Setup

We approach diachronic variation in poetry as distant reading task to visualize the development of interpretable topics over time and across languages. As in the previous mono-lingual approach, we retrieve the most important (likely) words for all topics and interpret these (sorted) word lists as aggregated topics. We then manually translate several topics that align over all four corpora. This

manual translation process is by no means optimal, as we will discuss later, but it already gives a first overview on prominent topics in these corpora.

To discover trends over time, we bin our documents into time slots of 25 years width each, except for early English where two large slots (1600–1674 and 1675–1749) were used due to sparse data. See Figures 4.17 and 4.18 for a plot of the number of documents per bin. To visualize trends of singular topics over time, we follow the strategy of Haider (2019), as outlined in the previous section: We aggregate all documents `d` in slot `s` and sum the probabilities of topic `t` given `d` and divide by the number of all `d` in `s`. This gives us the average probability of a topic per time slot. We then plot the trajectories for each single topic.

4.4.2.2 Alignment and Interpretation of Topic Trajectories

Based on a few selected topics, we can trace similarities and disparities over poetic traditions. See Figures 4.10–4.24 for a selection of interpretable topic trends, where the four languages align and diverge. Please note that the scaling on the y-axis differ for each topic, as some topics are more pronounced in the whole dataset overall.

Figure 4.19 shows the topic "Nation", which has a similar trend in German, Czech, and Russian, but is not present in the English corpus (cf. completely different geopolitical situation of the British empire). In the German corpus it emerges in the second half of the 18th century and peaks around 1825 to 1850 (outlining the period of 'Vormärz') as we already saw in Figure 4.10, outlining the political revolution and uprisings in the decades before 1850 (revolution in Germany in March of 1848). The same peak before 1850 can be found in the Czech corpus (late National Revival), and slightly delayed in Russian. In all the three corpora, the topic loses importance after 1850/60, but is gaining traction once again at the beginning of the 20th century, with the move into nationalism at the dawn of the 20th century.

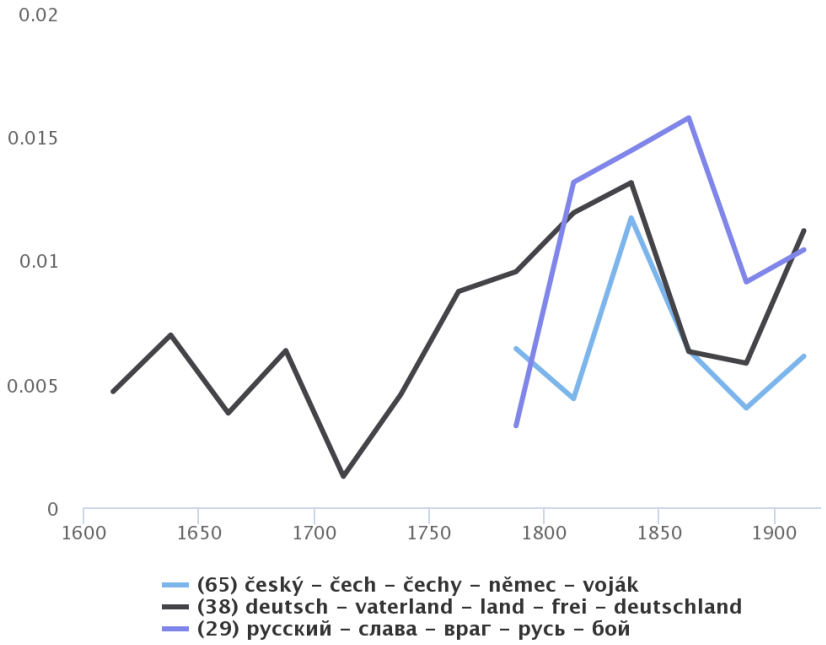

Figure 4.19: Multilingual Poetic Topics: Topic **Nation**

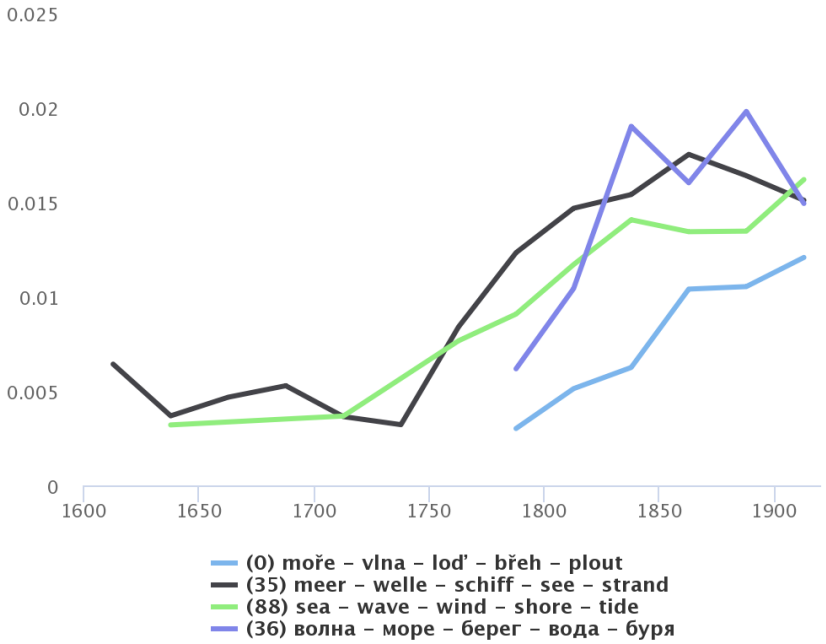

Figure 4.20: Multilingual Poetic Topics: Topic **Sea**

Figure 4.20 shows the topic ”Sea”, which has a similar rising tendency towards the second half of the 19th century and stays stable into Modernity. This topic is most pronounced for the Russian and German period of Romanticism, after which it seems to taper off, while it still shows an upward trajectory for English and Czech.

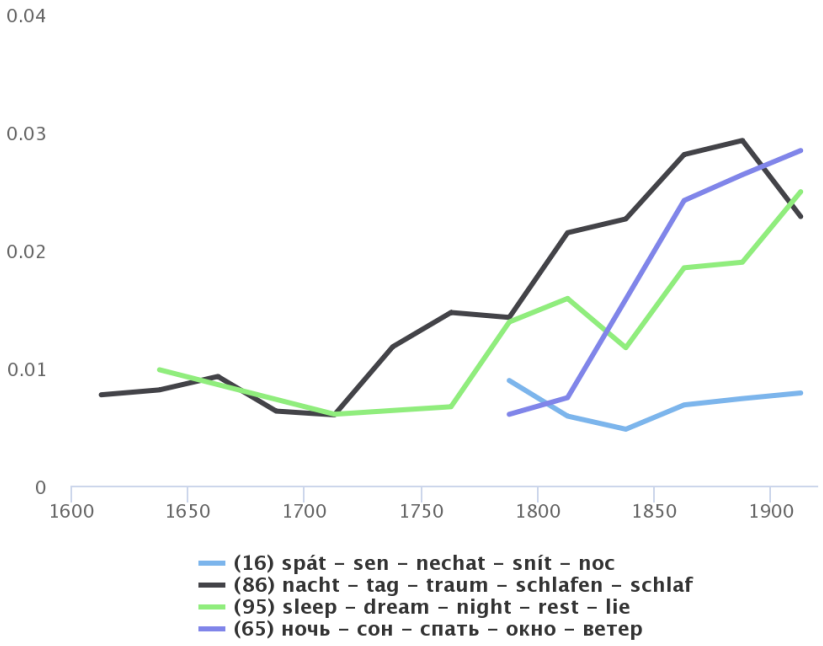

FIGURE 4.21: Multilingual Poetic Topics: Topic **Sleep**

Judging by the trajectories, the topic ”Sleep” (Figure 4.21) appears fairly correlated with the topic ”Sea” in English, German, and Russian (though there seems to be no obvious connection), with a focus on late Romanticism and Modernity, but it is rather marginal in the Czech corpus. It seems that, with the onset of Romanticism around 1750 ‘Sleep’ became a mainstay topic in romanticism. This topic is also related to the German romantic topic ‘Girls, Sleep, [Bodyparts]’ seen in Figure 4.11.

Figure 4.22 shows the topic “Sorrow” that has clearly separable trends, with English and German on one hand and Czech and Russian on the other. In the first case (the Germanic languages) it is associated with the period of Romanticism (although becoming prominent earlier in English), and in

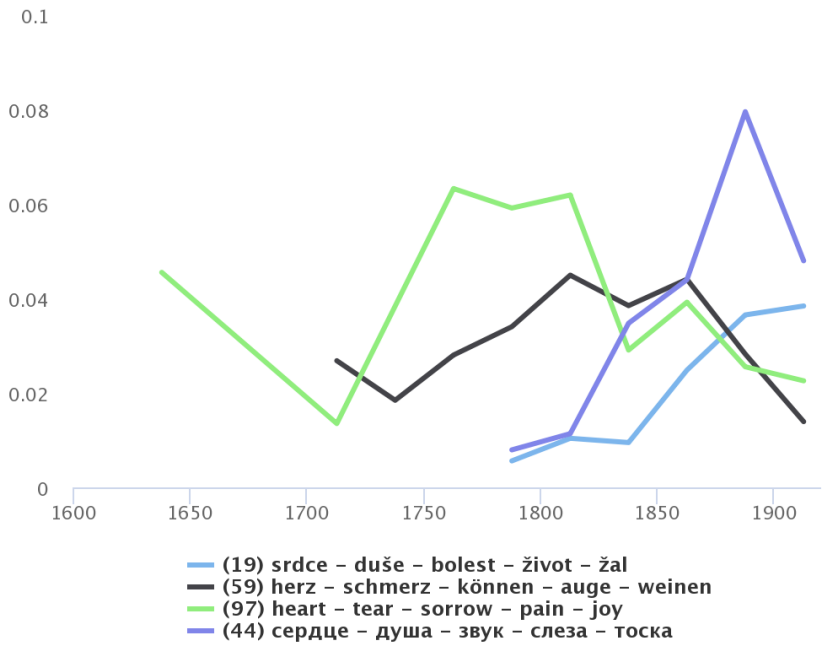

Figure 4.22: Multilingual Poetic Topics: Topic **Sorrow**

the latter with late 19th century Modernism (although in Russian it emerges already 70 year earlier in the period of Romanticism, 1825 to 1850). While poetry gives the space to celebrate feelings of joy or being in love, it also helps process pain. It is thus not surprising that the themes of grief and sorrow occur in our corpora. Most commonly, grief poems are elegies that reflect and lament someone's death. This topic should allow us to extract all elegies from the corpora.

Figure 4.23 shows the topic "Stars & Sky", which is pronounced in English and German High Romanticism (1800 to 1825) and in Russian Late Romanticism (1825 to 1850). In Czech the peak occurs delayed in the generation of "Máj" (period 1850 to 1875). Note that these authors claimed themselves as the followers of Karel Hynek Mácha (1810–1836), who in turn is well-known for bringing English Romanticism themes into Czech poetry.

Lastly, Figure 4.24 shows the topic "Wine" which is associated with the Anacreontics. It is accented in early 18th century English poetry, second half 18th century German poetry, and late 18th century Czech poetry (almanacs

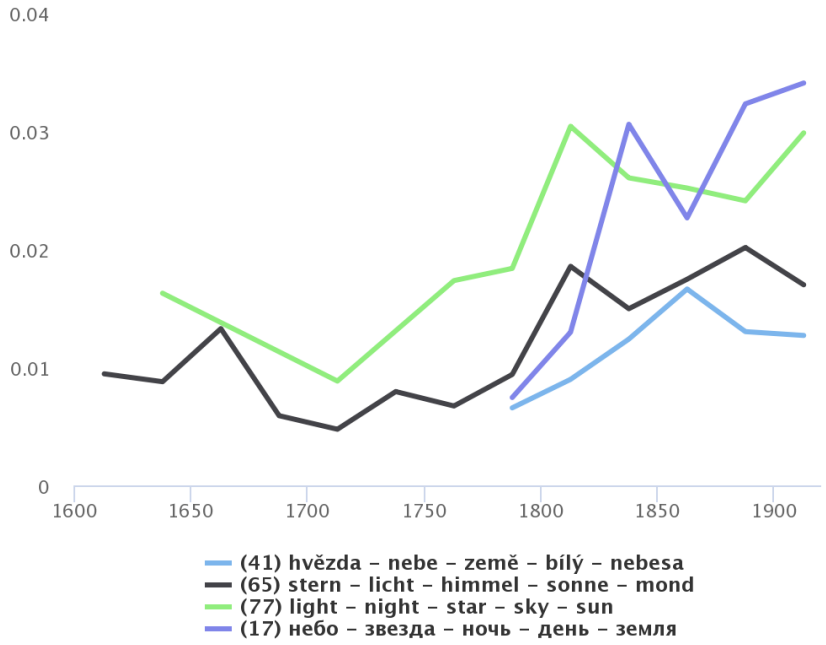

FIGURE 4.23: Multilingual Poetic Topics: Topic **Stars & Sky**

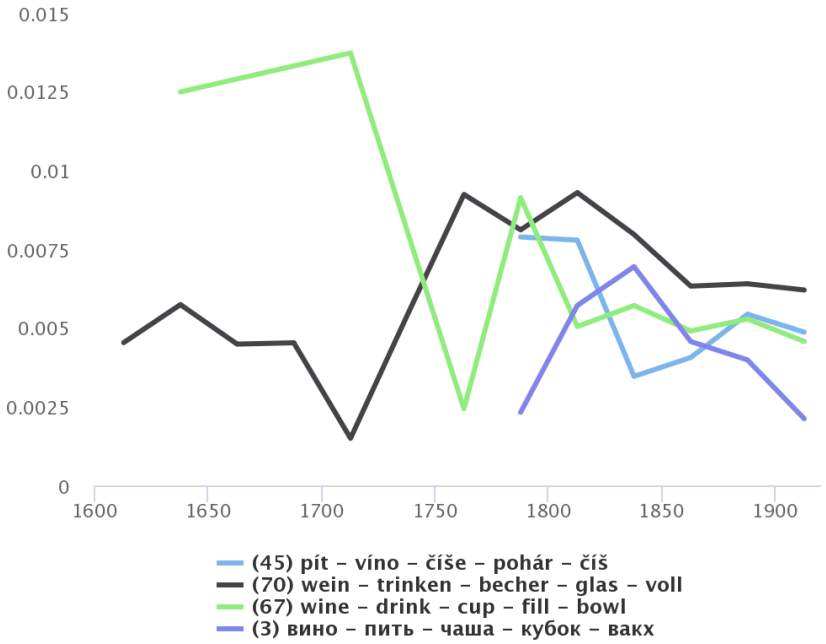

FIGURE 4.24: Multilingual Poetic Topics: Topic **Wine**

edited by A. J. Puchmajer). In Russian poetry it surprisingly peaks in the period of romanticism (1825 to 1850). One poet who is associated with the Anacreontics is Johann Gleim, for example with his 'Versuch in Scherzhaften Liedern'. According to Trop (2010), the Anacreontic form represents a lost poetic exercise in pleasure, one that would soon be overcome by competing poetic paradigms.[5] However, while this topic seems to decline in importance over time (and was never very prominent overall as seen on the average probability scale at the y-axis), it is still present later poems in our corpora, especially in German.

[5]'The attempt to link poetry to religious experience (Klopstock), poetry as education into ethical autonomy (Schiller), the cult of genius and personalized expression (Herder and Goethe), or the fantastic as a source of novelty (Romanticism). Pleasure itself, however, 'generates its own forms of novelty, its own way of loosening the hold of social and cultural norms and revealing the contingency of these norms.' (Trop, 2010).

4.5 Lexical Semantic Change and Emerging Tropes

Due to its succinctness and novelty of expression (Roberts, 2000; Underwood and Sellers, 2012), poetry is a great test bed for semantic change analysis. The experiments here originated from the question of how certain semantic fields (like topics) correlate, so that they would form pervasive poetic metaphors.

```
My love is like to ice, and I to fire:
How comes it then that this her cold so great
Is not dissolved through my so hot desire,
But harder grows the more I her entreat?
Or how comes it that my exceeding heat
Is not allayed by her heart-frozen cold,
But that I burn much more in boiling sweat,
And feel my flames augmented manifold?
What more miraculous thing may be told,
That fire, which all things melts, should harden ice,
And ice, which is congeal's with senseless cold,
Should kindle fire by wonderful device?
Such is the power of love in gentle mind,
That it can alter all the course of kind.

Edmund Spenser
```

FIGURE 4.25: Edmund Spenser's Sonnet 'My love is like to ice, and I to fire'.

Consider the sonnet in Figure 4.25, 'My love is like to ice, and I to fire', by Edmund Spenser. Here, we see a metaphor of love being likened to burning fire or love being cold as ice, where an interpersonal relationship can be hot or cold, whether there is attraction or rejection. The poem alludes to both the fire and ice imagery through related words like 'hot desire', 'heart-frozen cold', 'flames', 'melt', 'kindle', etc. We would thus assume that a topic model renders this poem as a mixture of a 'hot & cold' and a 'love & desire' topic. However, first pilot experiments on the correlation of LDA topics did not lead to conclusive results. Instead, here we develop a methodology that is based on

the syntagmatic and paradigmatic similarity of words over time via diachronic word embeddings.

We offer a method to explore poetic tropes, i.e., word pairs such as 'love (is) fire' or 'love (is) magic' by tracking their gain in association strength (cosine similarity) over time, based on the distributional semantic meaning of words, finding that most of these poetic tropes are gaining traction in the Romantic period. Further, we track the self-similarity of words, both with a change point analysis and by evaluating 'total self-similarity' of words over time. The former helps us to reconstruct literary periods, while the latter provides us with further evidence for the law of linearity of semantic change (Eger and Mehler, 2016) using our new method.

4.5.1 Method: Semantic Change

Our model learns diachronic word2vec embeddings jointly over time slots based on an architecture of Bamman et al. (2014a), eliminating the need to align embeddings over different time periods. With the method of Bamman et al. (2014a), we *jointly* compute embeddings across different linguistic variables (here time slots): each word w has an embedding

$$\mathbf{w} = \mathbf{e}_w \mathbf{W}_{\text{main}} + \mathbf{e}_w \mathbf{W}_C,$$

where $\mathbf{W}_{\text{main}} \in \mathbb{R}^{|V| \times d}$ is a main embedding matrix and $\mathbf{W}_C \in \mathbb{R}^{|V| \times d}$ is an embedding matrix for linguistic variable C, and $\mathbf{e}_w$ is a 1-hot vector of word w. In their original work, C ranges over geographic locations (US states), but we use time slots instead. A joint model has several advantages: it better addresses data sparsity and it directly learns to map words in a joint vector space without necessity of ex-post projection. In our work, we use this latter model for temporal embeddings in that each linguistic variable C corresponds to a time epoch t:

$$\mathbf{w}(t) = \mathbf{e}_w \mathbf{W}_{\text{main}} + \mathbf{e}_w \mathbf{W}_t$$

Thus, we do not need to align independently trained embeddings from every time slot. Instead, a joint (MAIN) model is learned that is then re-weighted for every time epoch. However, this does not necessarily mean that embeddings of a certain low-frequency word in a given time slot are stable. If there is not enough context for a given word in a certain time period t, the model just learns the MAIN embedding with little to no re-weighting, i.e., the matrix $\mathbf{W}_t$ may not be well estimated (at certain rows).

For training our model, we organize the corpus by stanzas, where every stanza represents a document. The reasoning behind this is that for poetic tropes (or metaphors more generally), words are likely to stand in local context, rather than at opposite end of the poem. For a corpus, we use an earlier version of DLK (see section 3.3). We merged the DTA and Textgrid collections and removed duplicate stanzas that match on their first line. This removed 9600 duplicate stanzas. Filtering Dutch and French material further eliminated 3200 stanzas. Since the earliest time slot 1575–1625 is too small, we merge it with the adjacent slot, resulting in six time slots total.

See figure 4.26 for the distribution of stanzas in 50 year time slots. The slots are labeled with approximate literature period information based on the clustered annotation in ANTI-K. We can see that the Romantic period (approx. 1750–1875) is overly heavy, while the Barock period is to some extent underrepresented. We use time slots of 50 year width to avoid learning sparse representations of words, since some words of interest only occur a few hundred times in the corpus (e.g., 'Begeisterung' (excitement) only occurs around 300 times in the whole corpus overall). To get reliable embeddings for a word, it should be present in each time slot at least a few dozen times.

We lemmatize the corpus based on a gold token-lemma mapping that we extracted from DTA (the tcf format version). When this does not cover a token, we pos-tag the line to feed the word with its pos-tag into germalemma.[6]

[6] `https://github.com/WZBSocialScienceCenter/germalemma`

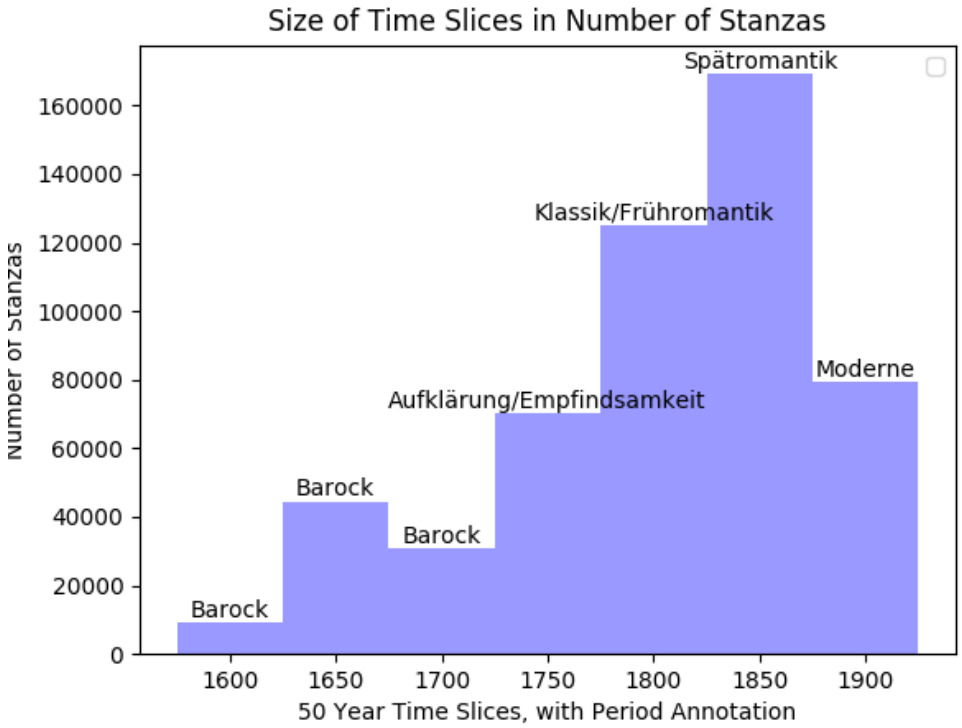

Figure 4.26: Distribution of stanzas in 50 year slots, 1575–1925 AD, with literary period approximation.

4.5.2 Self-Similarity of Words in Semantic Change

We investigate semantic self-similarity of words over time in two ways: (1) How does poetic diction change over successive time steps (change point detection), and (2) how does contextual word meaning change in total over the whole time frame with respect to the word's frequency (laws of conformity and linearity)? Self-similarity refers to computing the similarity of a word to itself across different time slots. We use a model with a 25+50 sliding time window, where time steps increase by 25 years, with a window size of 50 years. This effectively doubles the amount of data and allows a more fine grained analysis.

4.5.2.1 Pairwise Self-Similarity (Change Point Detection)

We compute how the contextual use of words changes over successive time steps. We do this by determining the self-similarity of a word w over time by

calculating the cosine similarity of the embedding vectors $\mathbf{w}(t)$ for w at time periods $t = t_i$ and $t = t_{i+1}$ as in equation (4.2):

$$\text{cossim}(\mathbf{w}(t_i), \mathbf{w}(t_{i+1})) \tag{4.2}$$

where $\text{cossim}(\mathbf{a}, \mathbf{b})$ is defined as $\mathbf{a}^\intercal\mathbf{b}$ for two normalized vectors $\mathbf{a}$ and $\mathbf{b}$.

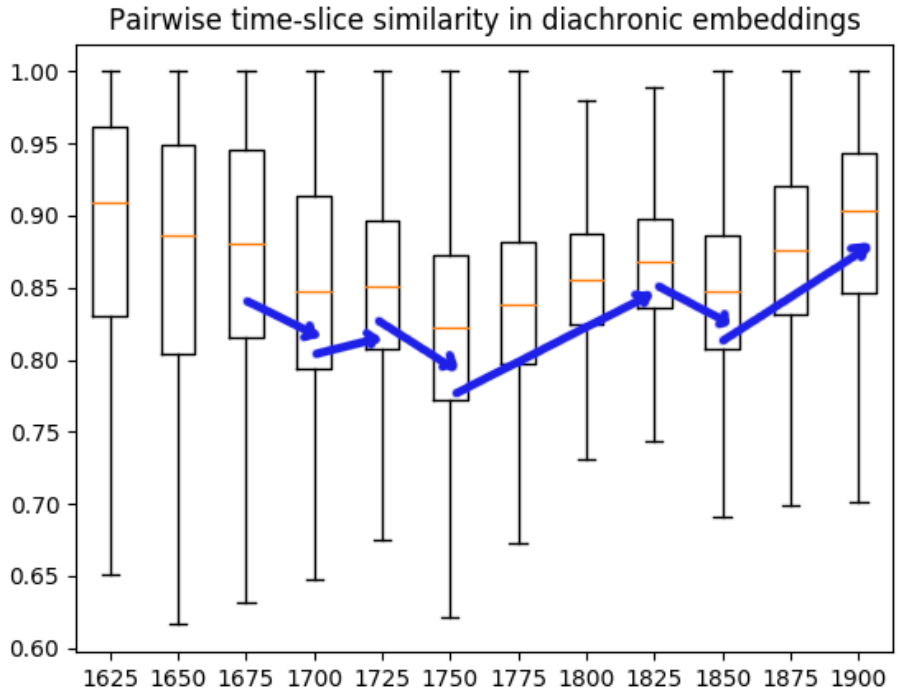

FIGURE 4.27: Pairwise Self-Similarity. Top-3000 most frequent words. Cossine similarities of word w with itself in adjacent time slots $cossim(w(t_i), w(t_{i+1}))$
: Context densification in literary periods.

Thus, we can aggregate the self-similarity of words at every time step (the similarity of a word to itself over two subsequent time slots) and aggregate the change for all these words combined. See figure 4.27 for a boxplot of this pairwise self-similarity for the 3000 most frequent words in aggregate.

Our interpretation is that rising similarity (higher median) signifies a homogenization of overall word use (diction), while a falling similarity signifies

semantic diversification. 'Rising' then means that the median of the box at t_{+1} is higher than t_0, while a 'falling similarity' means that the median at t_{+1} is lower than at t_0. A rising median means that the vocabulary is densifying. Thus, words become more similar to themselves over subsequent time steps, indicating that subsequent generations of poets are re-using word meanings and contexts, while a 'dip' signifies a sudden break with tradition, fostering a new movement, a new use of words and contexts.

We see a steady falling trajectory in the period between 1600 and 1675, with a dip at 1700. This period is generally regarded as the 'Barock' period. Over this period, the embedding of words seems to be fairly heterogenous, as there is no clear trend. The dip at 1700 nevertheless shows that there is a sudden change over the respective 50 year windown.

After that, word use slowly homogenizes (resulting in a rising trajectory), until we see a sharp dip around 1750, the onset of the Romantic period. Then it homogenizes during the Romantic period, until a dip at 1850, the end of the Romantic period, and then a homogenization into the onset of Modernity.

4.5.2.2 Total Self-Similarity (Linearity of Semantic Change)

We determine change of word meaning across any possible time distances as a probing for the linearity of semantic change in our corpus.

For this, we calculate the semantic self-similarity of a word across all time periods t_i and t_j with $t_i < t_j$. We then aggregate all pairwise distances in years

$$\text{dist}(t_i, t_j) = |t_i - t_j|$$

for all words w that occur at least 50 times in every time slot.[7] To obtain robust estimates of embeddings, we only allow words that occur at least 50 times in every time slot and remove stopwords, leaving us with only 472 lemmas.

[7]For all 25, 50, ..., 300 year distances, cossims per word in these distances are averaged, so we are left with one value per distance and word.

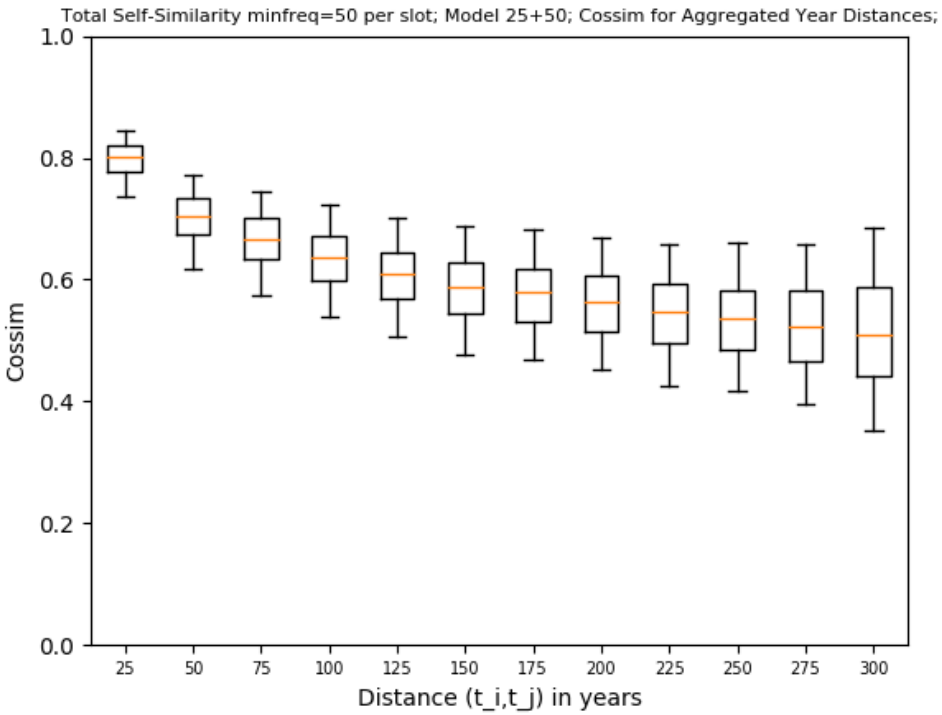

FIGURE 4.28: Total Self-Similarity of words that occur at least 50 times in every time slot. Cossine similarities aggregated by the distance of compared time slots (t_i, t_j) averaged for every time slot given a word. Removed stopwords. Whiskers: [5,95] percentiles.

The x-axis in Figure 4.28 gives the distances $\text{dist}(t_i, t_j)$ while the y-axis shows the distribution of cossims over all words w within each distance.

We find that there is approximately a linear relation between the distance of time slots for an average word, where close slots are more similar, and far apart slots are increasingly dissimilar. However, the variance also increases with distance. This increase in variance should be investigated in future research.

Additionally, to test a frequency effect on semantic change (which Dubossarsky et al. (2017) noted to be problematic) we divided our 472 words equally into a low-frequency and a high-frequency band (the top frequent 236 and the 236 least frequent words). We find that the low-frequency band shows a generally higher self-similarity than the high-frequency band over all

distances. This would mean that, overall, high frequency words tend to be more semantically diverse over time, i.e. stand in more diverse contexts (the word 'the' stands in more contexts than the word 'love', thus the chance for a high frequency word to undergo change is higher). In contrast, low-frequency words stand in fewer contexts, therefore undergo less change. However, this could also come from the tendency of the model to revert to MAIN for low frequency words (due to the concatenation).

4.5.3 Emerging Tropes: Near-Synonyms & Metaphors

To detect emerging tropes, we calculate the cosine similarity of word pairs over time. For the sake of visualization we use a 50+50 model with 6 time slots. We calculate the distance of the embeddings of a particular word against every other word w in the vocabulary, where w has to occur at least 30 times in the corpus, and it needs to be represented in every time slot at least twice. We allow one slot to be empty, assuming that there the MAIN embedding will be sufficient. These parameters were determined so that enough words remain in the vocabulary, accounting for the factor that not all words are present in all time slots at sufficient numbers. If we would require that a word occurs at least 50 times per time slot reduces the available vocabulary to only a few hundred words, removing many words of interest. On the other hand, allowing vocabulary that is missing in multiple time slots leads to rather sparse embeddings where the temporal embeddings of these words are hardly distinguishable from the MAIN embedding.

For a proof of concept, we calculate the distance of the embeddings of the word 'love' against every other word w in the vocabulary at each time step and then perform Principal Component Analysis (PCA) over the resulting trajectories (that have six data points each, since there are six time slots).

#	**rising traj.**	**falling traj.**	**stable high traj.**	**stable low traj.**
1	frische	aufrechen	liebe	brummen
2	veilchen	alsbald	freundschaft	krähen
3	niedersinken	billigkeit	lust	rasseln
4	duftig	erzeigen	treue	rum
5	jenseits	unterstehen	trieb	bock
6	zauber	betragen	seligkeit	dum[m]
7	entgleiten	stracks	hoffnung	prasseln
8	künden	zuerkennen	glaube	trommel
9	hoffend	hierin	keusch	säbel
10	efeu	schmeissen	treu	traben
11	enthüllen	anlaß	erkalten	bellen
12	erfüllung	jederzeit	wahr	block
13	heimat	muhen (mühen?)	immerdar	bügel
14	trübe	schimpfen	regung	gaul
15	gloria	stecken	gegenliebe	grasen
16	rieseln	anderst	herz	übern
17	verbluten	hierauf	freude	binse

Table 4.4: Top 17 words per dimension for 'love' tropes from PCA extremes, as plotted in the following figures.

#	**rising traj.**	**falling traj.**	**stable high traj.**	**stable low traj.**
1	fresh(ness)	raking (leaves)	love	hum
2	violets	soon	friendship	crowing
3	sinking down	cheapness/indulgence	lust	rattling
4	airy/scenty	show	faithfulness	around
5	beyond	subordinate	drive/urge	buck
6	magic	manners/account	bliss	dumb
7	slipping	straightaway	hope	pattering
8	to announce	to acknowledge	believe	to drum
9	hoping	in this	chaste	saber
10	ivy	throwing	faithful	trotting
11	reveal	occasion	cooling	bark
12	fulfillment	anytime	true	block
13	home(land)	moo (efforts?)	evermore	strap
14	blear/murky	scold	stir	horse/nag
15	glory	to stick	mutual love	to graze
16	trickle	other/unlike	heart	over
17	bleed out	hereupon	joy	rush

Table 4.5: Translations for top 17 words per dimension for 'love' tropes from PCA extremes, as plotted in the following figures.

We then perform Principal Component Analysis (PCA) over the resulting trajectories (the similarity of one word against all other words over time, thus the number of trajectories is equivalent to the size of the vocabulary). The resulting principal components show that similar trajectories are co-variant. Component 1 aggregates stable high/low trajectories, while component 2 aggregates rising/falling trajectories. We illustrate our finding with the tropes for the concept 'love' ('Liebe' in German) and determine the most salient word pairs over the whole dataset. 'Love' is a very frequent word in poetry. Nevertheless, this approach works equally well for any word, except for very low frequency words that exhibit idiosyncratic behavior as they are not well distributed and consequently don't have a good embedding.

The first 4 components of PCA explain over 95% variance, where component 1 explains 73%, component 2, 13%, and component 3, 5%. We retrieve the top-25 word pairs at every component extreme. We find that component 1 orders trajectories based on high/low semantic similarity, while component 2 orders based on rising/falling trajectories. See Figures 4.29 (stable high trajectory), 4.30 (rising trajectory), 4.33 (stable low trajectory) and 4.32 (falling trajectory). See Table 4.4 for the respective word pairs (collocations) with 'love' as they are plotted in the following, and Table 4.5 for a translation of these words. Note that 'love' (liebe) is always at a similarity of 1.0, since it is always identical to itself in respective time slots.

Stable High Trajectories (near synonyms) Figure 4.29 shows a plot of (also see Tables 4.4 and 4.5 column 3) have a consistently high cosine, meaning that these collocations have remained unchanged since the Baroque period: 'love is fidelity',[8] 'love is friendship',[9] or 'love is lust'. These are conventional near-synonyms. A k-nearest neighbor (KNN) analysis would also retrieve such near-synonyms. Performing our analysis for multiple words, we find that the

[8]('Treue', 'Liebe')
[9]('Freundschaft', 'Liebe')

idiom ('apples', 'pears') is a special case, as it strongly loads into both rising and stable high PCA dimensions (both top 20), as seen in Figure 4.31. Note that only the plural forms of apples and pears is an idiom, but not with the singular forms ('Apfel und Birne') (but we are using lemmas here)[10]

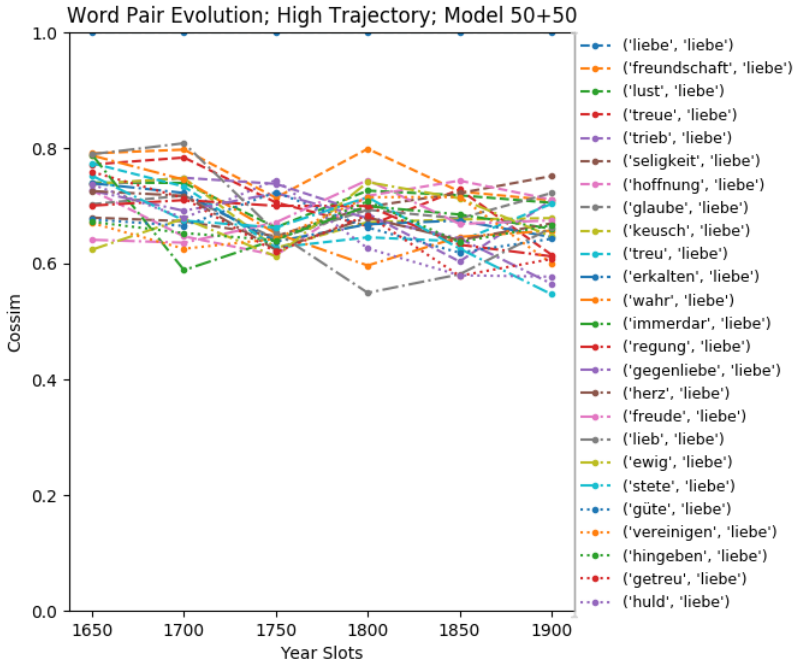

FIGURE 4.29: Stable High Trajectories, Word Similarities to 'Love'

Rising Trajectories (emerging tropes) Figure 4.30 (also see Table 4.4 column 1) shows rising collocations that emerge during the Romantic period, i.e. 'fresh love' ('frische', 'Liebe'), 'love is magic/enchantment' ('Zauber', 'Liebe') and 'love is violets' ('Veilchen', 'Liebe'), or 'love is slipping' ('entgleiten', 'Liebe'), or 'scented love' ('duftend', 'Liebe'). A metaphorical (trope) interpretation is most likely here. These word similarities reveal poetic metaphors that emerged over time as the constituent words were increasingly used in similar contexts and then stayed associated into modernity.

[10]Also see `https://www.redensarten-index.de/suche.php?suchbegriff=Aepfel+mit+Birnen+vergleichen&suchspalte%5B%5D=rart_ou` Note also that this plot was made with different parameters, where the influence of the MAIN model is seen in the first time slot.

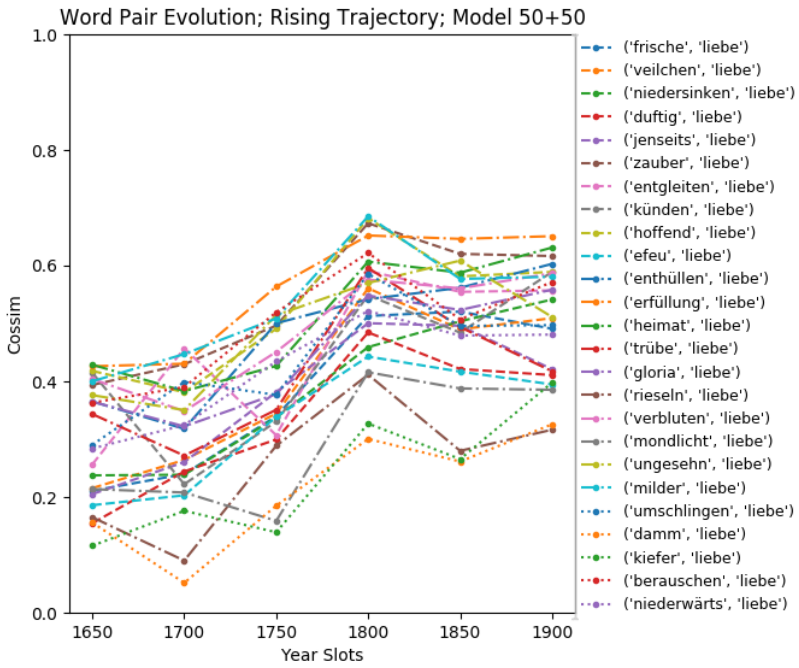

Figure 4.30: Rising Trajectories, Word Similarities to 'Love'

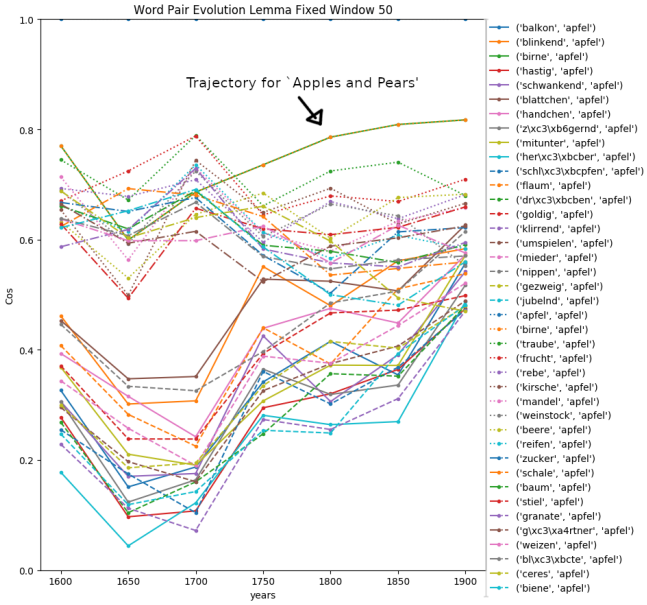

Figure 4.31: High and Rising Trajectories for embedding similarities to the word 'apple'. Illustration of the already high and then rising trajectory of the idiom 'Äpfel und 'Apples and Pears' (equivalent to 'apples and oranges').

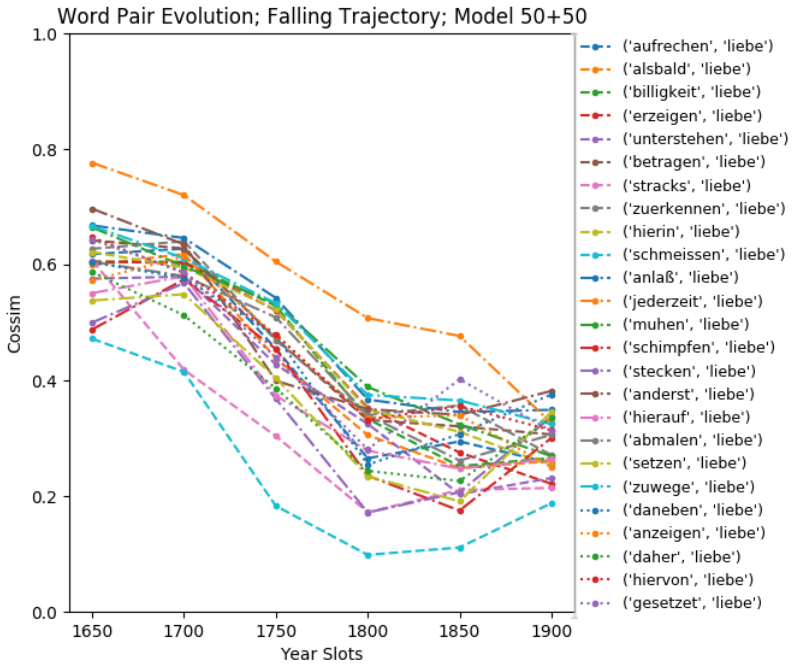

FIGURE 4.32: Falling Trajectories, Word Similarities to 'Love'

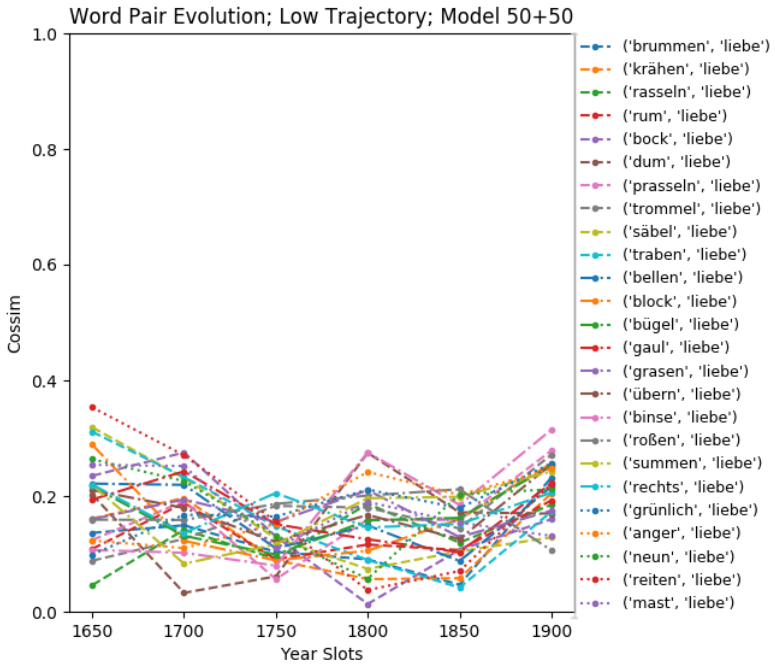

FIGURE 4.33: Low Trajectories, Word Similarities to 'Love'

Falling Trajectories (losing word senses) As illustrated in figure 4.32 (column 2 in Table 4.4), these collocations fall into obscurity. We find 'cheap love' (billigkeit, which changed in meaning from 'equity/indulgence' to 'cheap') or things like 'raking' (aufrechen) or 'manners/accounting' (betragen). On the one hand, the constituent words may have changed their meaning (or at least lost or gained some senses) as in the case of 'billigkeit', such that the new meaning is further removed from 'love', either because the concept of love is not associated with it anymore (in the Western world, love is largely not dependent on the indulgence of the parents anymore), or because the word also occurs in different contexts now.

Stable Low Trajectories (no association) The lines in figure 4.33 (column 4 in Table 4.4) signify word pairs that are always far apart. We find things that make noise, like 'love (is not) drums' ('trommel'), and a topic that circles around horses. These results do not suggest any sensible interpretation, and such associations might change strongly across different corpora.

4.6 Concluding Remarks

4.6.1 Diachronic Variation in Poetry

We have shown that distributional semantic methods are powerful tools to show changes in language use over a period of around 300 years of poetic writing. Using poetry corpora reveals particularities of that specific genre, such as the topics that were popular at certain points in time, outlining literary periods. We have shown this also across different languages. Despite the methodological simplicity in the way we used these methods, we could already show some clear trends. Also, diachronic word embeddings allowed us a window into how the meaning of words in poetry changed, including the discovery of poetic tropes. And while these methods are promising, more work is necessary to test whether these approaches are robust across other datasets. Furthermore, the used data here is relatively small in comparision to the amount of data discussed in other papers. We certainly had issues with data sparsity, considering the large time frame, where some words simply did not yet occur at certain times. 12–17 million tokens over 350 years is not a lot for these methods.

4.6.2 Diachronic Topic Models

We used Latent Dirichlet Allocation for a visualization of topic trends in a mono-lingual and a cross-lingual setting, illustrating the similarities and disparities between different poetic traditions and literary periods. Our method is largely based on reading and translating topic distributions and finally interpreting the trajectories of relative topic importance against the backdrop of literary history. While most topics are easily interpretable and show a clear trend, others are quite noisy. We were able to identify salient topics for literary periods, such as ‘virtue & arts’ for Enlightenment, ‘World, Power, Lust, Time’ for Baroque, or ‘Garden, Flowers’ and ‘Singing’ for Romanticism.

We find that some topics, especially the examples chosen, do align across languages, sometimes with temporal delay (as they were picked up later in another language), while other topics were not as heavily discussed in other poetic discourses (such as the lack of 'Nation' in English). Future work should look into cross-lingual alignment methods, e.g., through multi-lingual embeddings or poly-lingual topic models without parallel data. Finally, one flaw of our method is that over- or underrepresentation of certain authors or near-duplicates of poems (different editions) can lead to corpus imbalance. Consequently, this can impact our measure to calculate the relative importance of a topic given a certain time stamp and should be addressed in future work. For an exploratory experiment, dating poetry is very promising, however far from perfect. It should be investigated whether using stanzas instead of whole poems only improves results because of more available data. Also, it needs to be determined if better topic models can deliver a better baseline for diachronic change in poetry, and if better style features will outperform semantics. We suspect that only selecting clear trending and peaking topics (e.g., through co-variance) might further improve the results for dating.

4.6.3 Semantic Change and Emerging Tropes

We investigated distributional semantic change through word embeddings. With self-similarity, we can reconstruct literature period transitions and find that the law of linear semantic change also applies to poetry. We extracted emerging and vanishing poetic tropes based on the co-variance of time trajectories of the semantic distance in word pairs. This method is applicable more broadly to cluster similar trajectories for any given word pairs. We found trajectories of word similarities that are beyond simple nearest-neighbor analysis, and illustrated findings for reasonable tropes with 'love'. While large, our dataset is still somewhat sparse in the distribution of words over all time slots, partially because many word forms simply emerge or vanish at a certain point,

e.g., 'excitement' (Begeisterung) does not occur in the Baroque period. Thus, for confident analysis of laws of semantic change or to get a broader view of poetic metaphors, more data and a more robust model is called for. Also, future research should investigate how such tropes and metaphors formed in other text genres.

Aesthetic Emotions 5

5.1 Introduction

Emotions are central to human experience, creativity and behavior. Models of affect and emotion, both in psychology and natural language processing, commonly operate on predefined categories, designated either by *continuous scales* of, e.g., *Valence*, *Arousal* and *Dominance* (Mohammad, 2016) or *discrete emotion labels* (which can also vary in intensity). Discrete sets of emotions often have been motivated by theories of basic emotions, encompassing concepts as proposed by Ekman (1992)—*Anger*, *Fear*, *Joy*, *Disgust*, *Surprise*, *Sadness*—and Plutchik (1991), who added *Trust* and *Anticipation*. These categories are likely to have evolved as they motivate behavior that is directly relevant for survival. However, *art reception* typically presupposes a situation of safety and therefore offers special opportunities to engage in a broader range of more complex and subtle emotions. These differences between real-life and art contexts have not been considered in natural language processing work so far.

To emotionally move readers is considered a prime goal of literature since Latin antiquity (Johnson-Laird and Oatley, 2016; Menninghaus et al., 2019, 2015). Deeply moved readers shed tears or get chills and goosebumps even

in lab settings (Wassiliwizky et al., 2017). In cases like these, the emotional response actually implies an aesthetic evaluation: narratives that have the capacity to move readers are evaluated as good and powerful texts for this very reason. Similarly, feelings of suspense experienced in narratives not only respond to the trajectory of the plot's content, but are also directly predictive of aesthetic liking (or disliking). Emotions that exhibit this dual capacity (representing both a feeling and an aesthetic evaluation) have been defined as "aesthetic emotions" (Menninghaus et al., 2019). Contrary to the negativity bias of classical emotion catalogues, emotion terms used for aesthetic evaluation purposes include far more positive than negative emotions. At the same time, many overall positive aesthetic emotions encompass negative or mixed emotional ingredients (Menninghaus et al., 2019), e.g., feelings of suspense include both hopeful and fearful anticipations. Feelings of Nostalgia can include the bitter feeling that a moment has forever passed, but also the delight to be able to experience the moment again in your memories.

For these reasons, we argue that the analysis of literature (with a focus on poetry) should rely on specifically selected emotion items rather than on the narrow range of basic emotions only. Our selection is based on previous research on this issue in psychological studies on art reception and, specifically, on poetry. For instance, (Knoop et al., 2016) found that *Beauty* is a major factor in poetry reception.

With the goal of modeling the aesthetic experience of reading poetry, we consider emotions as they are *elicited in the reader*, rather than what is *expressed in the text* or *intended by the author*. In this work, we look at the emotions readers experience in themselves when reading, following work on the reception of literature, as opposed to emotions that are immanent to the text (such as emotion between protagonists) or emotions that are hypothetically expressed or intended by authors (e.g., what they felt when writing or want to make readers feel).[1] Thus, we conceptualize a set of *aesthetic emo-*

[1]Which would be a more fitting scenario for the analysis of social media.

tions that are predictive of aesthetic appreciation in the reader. By default, we capitalize emotion terms to highlight their use as categories.

We also allow the annotation of multiple labels per line of poetry to capture mixed emotions (within their context) and evaluate this novel setting in an annotation experiment both with carefully trained experts and via crowdsourcing. Our annotation with experts leads to an acceptable agreement of $\kappa = .70$, resulting in a consistent dataset for future large scale analysis. Finally, we conduct first emotion classification experiments based on BERT, showing that identifying aesthetic emotions is challenging in our data, with up to .52 F1-micro on the German subset. Furthermore, we present first experiments in transfer learning to examine the viability of other annotations for emotion prediction, and we show preliminary empirical results on the mixing of aesthetic emotions and their prevalence w.r.t. literary periods.

We primarily adopt and adapt emotion terms that (Schindler et al., 2017a) have identified as aesthetic emotions in their study on how to measure and categorize such particular affective states. Further, we consider the aspect that, when selecting specific emotion labels, the perspective of annotators plays a major role. Whether emotions are *elicited in the reader*, *expressed in the text*, or *intended by the author* largely changes the permissible labels. For example, feelings of *Disgust* or *Love* might be intended or expressed in the text, but the text might still fail to elicit corresponding feelings as these concepts presume a strong reaction in the reader. Our focus here was on the actual emotional experience of the readers rather than on hypothetical intentions of authors. We opted for this reader perspective based on previous research in NLP (Buechel and Hahn, 2017a,b) and work in empirical aesthetics (Menninghaus et al., 2017), that specifically measured the reception of poetry. Our final set of emotion labels consists of *Beauty/Joy, Sadness, Uneasiness, Vitality/Energy, Suspense, Awe/Sublime, Humor, Annoyance*, and *Nostalgia*.[2]

[2]The concepts *Beauty* and *Awe/Sublime* primarily define object-based aesthetic virtues. Kant (2001, orig. 1790) emphasized that such virtues are typically intuitively felt rather

In addition to selecting an adapted set of emotions, the annotation of poetry brings further challenges, one of which is the choice of the appropriate unit of annotation. Previous work considers words[3] (Mohammad and Turney, 2013b; Strapparava and Valitutti, 2004), sentences (Alm et al., 2005b; Aman and Szpakowicz, 2007), utterances (Cevher et al., 2019), sentence triples (Kim and Klinger, 2018), or paragraphs (Liu et al., 2019) as the units of annotation. For poetry, reasonable units follow the logical document structure of poems, i.e., verse (line), stanza, and, owing to its relative shortness, the complete text. The more coarse-grained the unit, the more difficult the annotation is likely to be, but the more it may also enable the annotation of emotions in context. We find that annotating fine-grained units (lines) that are hierarchically ordered within a larger context (stanza, poem) caters to the specific structure of poems, where emotions are regularly mixed and are more interpretable within the whole poem. Consequently, we allow the mixing of emotions already at line level through multi-label annotation.

The remainder of this chapter includes (1) a report of the annotation process that takes these challenges into consideration, (2) an implementation of baseline models for the novel task of aesthetic emotion annotation in poetry, including experiments in transfer learning, i.a., to measure the relevance of meter for elicited emotions, (3) first insight into mixed emotions and which emotions are prevalent for literary periods, and (4) a comparison of crowd sourcing emotions with expert annotations. In a first study, the annotators work on the annotations in a closely supervised fashion, carefully reading each verse, stanza, and poem. In a second study, the annotations are performed via crowdsourcing within relatively short time periods with annotators not seeing the entire poem while reading the stanza. Using these two settings, we aim at

than rationally computed. Such *feelings* of *Beauty* and *Sublime* have therefore come to be subsumed under the rubrique of *aesthetic emotions* in recent psychological research (Menninghaus et al., 2019). For this reason, we refer to the whole set of category labels as *emotions* throughout this thesis.

[3]to create emotion association dictionaries

obtaining a better understanding of the advantages and disadvantages of an expert vs. crowdsourcing setting in this novel annotation task. Particularly, we are interested in estimating the potential of a crowdsourcing environment for the task of self-perceived emotion annotation in poetry, given time and cost overhead associated with in-house annotation process (that usually involve training and close supervision of the annotators).

We provide the final datasets of German and English language poems annotated with reader emotions on verse level at `https://github.com/tnhaider/poetry-emotion`.

5.2 Related Work

Emotion and Poetry in Natural Language Processing: Corpus-based analysis of emotions in poetry has been considered, but there is no work on German, and little on English. Kao and Jurafsky (2015) analyze English poems with word associations from the Harvard Inquirer and LIWC, within the categories *positive/negative outlook*, *positive/negative emotion* and *phys./psych. well-being*. Hou and Frank (2015) examine the binary sentiment polarity of Chinese poems with a weighted personalized PageRank algorithm. Barros et al. (2013) followed a tagging approach with a thesaurus to annotate words that are similar to the words 'Joy', 'Anger', 'Fear' and 'Sadness' (moreover translating these from English to Spanish). With these word lists, they distinguish the categories 'Love', 'Songs to Lisi', 'Satire' and 'Philosophical-Moral-Religious' in Quevedo's poetry. Similarly, Alsharif et al. (2013) classify unique Arabic 'emotional text forms' based on word unigrams.

Mohanty et al. (2018) create a corpus of 788 poems in the Indian Odia language, annotate it on text (poem) level with binary negative and positive sentiment, and are able to distinguish these with moderate success. Sreeja and Mahalakshmi (2019) construct a corpus of 736 Indian language poems

and annotate the texts on Ekman's six categories + Love + Courage. They achieve a Fleiss Kappa of .48.

In contrast to our work, these studies focus on basic emotions and binary sentiment polarity only, rather than addressing aesthetic emotions. Moreover, they annotate on the level of complete poems (instead of fine-grained verse and stanza-level).

Emotion Annotation: Emotion corpora have been created for different tasks and with different annotation strategies, with different units of analysis and different foci of emotion perspective (reader, writer, text). Examples include the ISEAR dataset (Scherer and Wallbott, 1994) (document-level); emotion annotation in children stories (Alm et al., 2005a) and news headlines (Strapparava and Mihalcea, 2007) (sentence-level); and fine-grained emotion annotation in literature by Kim and Klinger (2018) (phrase- and word-level). We refer the interested reader to an overview paper on existing corpora (Bostan and Klinger, 2018).

We are only aware of a limited number of publications which look in more depth into the emotion perspective. Buechel and Hahn (2017a) report on an annotation study that focuses both on writer's and reader's emotions associated with English sentences. The results show that the reader perspective yields better inter-annotator agreement. Yang et al. (2009) also study the difference between writer and reader emotions, but not with a modeling perspective. The authors find that positive reader emotions tend to be linked to positive writer emotions in online blogs.

Emotion Classification: The task of emotion classification has been tackled before using rule-based and machine learning approaches. Rule-based emotion classification typically relies on lexical resources of emotionally charged words (Strapparava and Valitutti, 2004; Esuli and Sebastiani, 2006; Moham-

mad and Turney, 2013b) and offers a straightforward and transparent way to detect emotions in text.

In contrast to rule-based approaches, current models for emotion classification are often based on neural networks and commonly use word embeddings as features. Schuff et al. (2017) applied models from the classes of CNN, BiLSTM, and LSTM and compare them to linear classifiers (SVM and MaxEnt), where the BiLSTM shows best results with the most balanced precision and recall. Abdul-Mageed and Ungar (2017) claim the highest F_1 with gated recurrent unit networks (Chung et al., 2015) for Plutchik's emotion model. More recently, shared tasks on emotion analysis (Mohammad et al., 2018; Klinger et al., 2018) triggered a set of more advanced deep learning approaches, including BERT (Devlin et al., 2019) and other transfer learning methods (Dankers et al., 2019).

5.3 Data Collection

For our annotation and modeling studies, we build on top of two poetry corpora (in English and German), namely ANTI-K for German and EPG64 for English, as they are described in Chapter 3. This collection represents important contributions to the literary canon over the last 400 years. We make this resource available in TEI P5 XML[4] and an easy-to-use tab separated format: `https://github.com/tnhaider/poetry-emotion`

5.4 Expert Annotation

In the following, we will explain how we compiled and annotated three data subsets, namely, (1) 48 German poems with gold annotation. These were originally annotated by three annotators. The labels were then aggregated with majority voting and based on discussions among the annotators. Finally,

[4] `https://tei-c.org/guidelines/p5/`

they were curated to only include one gold annotation. (2) The remaining 110 German poems that are used to compute the agreement in table 5.3 and (3) 64 English poems contain the raw annotation from two annotators.

We report the genesis of our annotation guidelines including the emotion classes. With the intention to provide a language resource for the computational analysis of emotion in poetry, we aimed at maximizing the consistency of our annotation, while doing justice to the diversity of poetry. We iteratively improved the guidelines and the annotation workflow by annotating in batches, cleaning the class set, and the compilation of a gold standard. The final overall cost of producing this expert annotated dataset amounts to approximately €3,500.

5.4.1 Workflow

The annotation process was initially conducted by three female university students majoring in linguistics and/or literary studies, which we refer to as our "expert annotators". We used the INCePTION platform (Klie et al., 2018) for annotation.[5]

Starting with the German poems, we annotated in batches of about 16 (and later in some cases 32) poems. After each batch, we computed agreement statistics including heatmaps, and provided this feedback to the annotators. For the first three batches, the three annotators produced a gold standard using a majority vote for each line. Where this was inconclusive, they developed an adjudicated annotation based on discussion. Where necessary, we encouraged the annotators to aim for more consistency, as most of the frequent switching of emotions within a stanza could not be reconstructed or justified.

When experiencing poems (or any piece of art for that matter), emotions are regularly mixed and are more interpretable within the whole sequence of

[5] `https://inception-project.github.io/use-cases/po-emo/`

the poem. We therefore annotate lines hierarchically within the larger context of stanzas and the whole poem. Hence, we instruct the annotators to read a complete stanza or full poem, and then annotate each line in the context of its stanza. To reflect on the emotional complexity of poetry, we allow a maximum of two labels per line while avoiding heavy label fluctuations by encouraging annotators to reflect on their feelings to avoid 'empty' annotations. Rather, they were advised to use fewer labels and more consistent annotation. This additional constraint is necessary to avoid "wild", non-reconstructable or non-justified annotations.

All subsequent batches (all except the first three) were only annotated by two out of the three initial annotators, coincidentally those two who had the lowest initial agreement with each other. We asked these two experts to use the generated gold standard (48 poems; majority votes of 3 annotators plus manual curation) as a reference ("if in doubt, annotate according to the gold standard"). This eliminated some systematic differences between them[6] and markedly improved the agreement levels, roughly from 0.3–0.5 Cohen's κ in the first three batches to around 0.6–0.8 κ for all subsequent batches. This annotation procedure relaxes the *reader* perspective, as we encourage annotators (if in doubt) to annotate how they think the other annotators would annotate. However, we found that this formulation improves the usability of the data and leads to a more consistent annotation.

5.4.2 Emotion Labels

We opt for measuring the *reader perspective* rather than the text surface or author's intent. To closer define and support conceptualizing our labels, we use particular 'items', as they are used in psychological self-evaluations. These items consist of adjectives, verbs or short phrases. We build on top of Schindler et al. (2017a) who proposed 43 items that were then grouped by a factor

[6]One person labeled lines with more negative emotions such as *Uneasiness* and *Annoyance* and the person labeled more positive emotions such as *Vitality/Energy* and *Beauty/Joy*.

analysis based on self-evaluations of participants. The resulting factors are shown in Table 5.1. We attempt to cover all identified factors and supplement with basic emotions (Ekman, 1992; Plutchik, 1991), where possible. See Table 5.2 for an overview of the finally used set of emotion labels with their associated items.

Factor	Items
Negative emotions	anger/distasteful
Prototypical Aesthetic Emotions	beauty/sublime/being moved
Epistemic Emotions	interest/insight
Animation	motivation/inspiration
Nostalgia / Relaxation	nostalgic/calmed
Sadness	sad/melancholic
Amusement	funny/cheerful

Table 5.1: Aesthetic Emotion Factors by Schindler et al. (2017a).

#	Emotion Label	Items
1	**Beauty/Joy**	**found it beautiful/pleasing/makes me happy/joyful**
2	Sadness	makes me sad/touches me
3	**Uneasiness**	**found it ugly/unsettling/disturbing/frightening/distastef.**
4	Vitality/Energy	found it invigorating/spurs me on/inspires me
5	**Awe/Sublime**	**found it overwhelming/sense of greatness**
6	Suspense	found it gripping/sparked my interest
7	**Humor**	**found it funny/amusing**
8	Nostalgia	makes me nostalgic
9	**Annoyance**	**annoys me/angers me/felt frustrated**

Table 5.2: Final Set of Aesthetic Emotions with their Associated Items. Sorted by Label Frequency.

We started with a larger set of labels to then delete and substitute (e.g., by toning down) labels during the initial annotation process to avoid infrequent classes and inconsistencies. Further, we conflated labels if they showed considerable confusion with each other.

We used the following operations to alter the labelset:

Delete: (Boredom, Confusion, Other)

Tone Down: (Disgust → Uneasiness, Anger → Annoyance)

Merge: (Beauty, Joy → Beauty/Joy)

These iterative improvements particularly affected *Confusion*, *Boredom* and *Other* that were very infrequently annotated and had little agreement among annotators ($\kappa < .2$), and were thus removed from the final labelset. Annotators did not agree on which poems were boring or confusing. Also, once annotators were attuned to the labelset, they rarely used Other. The label Disgust, which is an unusual emotion for poetry, was toned down to Uneasiness, and the label Anger was toned down to Annoyance. Before these changes, both emotions were not used, but in their new framing they were annotated. Furthermore, the labels Beauty and Joy were originally treated as separate labels, but we found that they were frequently annotated with each other as primary and secondary emotions by the same annotators, and that they were frequently confused across annotators. We thus decided to merge them to form an inclusive label that should cover both Beauty and Joy, without necessarily claiming that these emotion terms are identical. For German, we also removed *Nostalgia* ($\kappa = .218$) after gold standard creation, but after further consideration, added it back for English, then achieving agreement. *Nostalgia* is still available in the gold standard (then with a second label *Beauty/Joy* or *Sadness* to keep consistency). However, *Confusion*, *Boredom* and *Other* are not available in any sub-corpus.

Our final set consists of nine classes, i.e., (in order of frequency) *Beauty/Joy, Sadness, Uneasiness, Vitality/Energy, Suspense, Awe/Sublime, Humor, Annoyance,* and *Nostalgia*.

5.4.3 Annotation Guidelines

In the following, we describe the annotation guidelines, the finally used labels and their items with an explanation of their meaning, and then give further details on the aggregation process.

Instructions for Annotators

1. The annotation should reflect your current feelings while reading the poem.
2. Read the entire poem and then stanza before annotating each line.
3. Label your emotions after reading each individual line (not sentence!).
4. Use as few emotions as possible!
5. Choose the emotion most dominant while reading the stanza.
6. Choose another emotion if necessary.
7. Choose at least one label per line.
8. You should not use more than two labels per line.
9. Change the dominant emotion within a stanza only if unavoidable.
10. If you change the non-dominant emotion within a stanza, remember to keep labeling the dominant emotion additionally to the new emotion.
11. Notice that nostalgia always has to be used with an additional label: Beauty/Joy or Sadness (only applies to provisional German annotation)

5.4.3.1 Label Definitions

Annoyance (annoys me/angers me/felt frustrated): Annoyance implies feeling annoyed, frustrated or even angry while reading the line/stanza. We include the class *Anger* here, as this was found to be too strong in intensity.

Awe/Sublime (found it overwhelming/sense of greatness): *Awe/Sublime* implies being overwhelmed by the line/stanza, i.e., if one gets the impression

of facing something sublime or if the line/stanza inspires one with Awe (or that the expression itself is sublime). The term *Sublime* originated with Kant (2001, orig. 1790) as one of the first aesthetic emotion terms. *Awe* is a more common English term. Awe/Sublime is used when an experience is overwhelming (item 1), or when you get a sense of greatness (item 2). The label can be used as a strong expression of Beauty/Joy, when simple Joy is not enough to express the appreciation you have for this particular piece of art. Sublime is used to describe a feeling of greatness in cases when you are totally blown away by the expression (expressive power of the poem), or because the text shows you something so magnificent and monumental that you are struck with Awe. A sublime poem might evoke a really colorful image in your mind, or your mouth won't close anymore because the experience was truly awesome. At the same time, the experience of Awe/Sublime is often associated with grand concepts like god, truth, life and death. Sublimity can be felt when reading apocalyptic poems, texts that evoke a sense of overwhelming doom or levity, possibly rich in imagery and arousal, ideally expressive and well crafted. Awe/Sublime is admittedly an acquired taste, in the sense that it is used more often when one is more familiar with the term and art reception in general.

Beauty/Joy (found it beautiful/pleasing/makes me happy/joyful): (Kant, 2001, orig. 1790) already mentions a "feeling of beauty", and it should be noted that it is not a 'merely pleasing emotion'. Therefore, in our pilot annotations, *Beauty* and *Joy* were separate labels. However, Schindler et al. (2017a) found that items for *Beauty* and *Joy* load into the same factors. Furthermore, our pilot annotations revealed, while *Beauty* is the more dominant and frequent feeling, both labels regularly accompany each other, and they often get confused across annotators. Therefore, we add *Joy* to form an inclusive label *Beauty/Joy* that increases consistency.

Humor (found it funny/amusing): Implies feeling amused by the line/stanza or if it makes one laugh.

Nostalgia (makes me nostalgic): Nostalgia is defined as a sentimental longing for things, persons or situations in the past. It often carries both positive and negative feelings. However, since this label is quite infrequent, and not available in all subsets of the data, we annotated it with an additional *Beauty/Joy* or *Sadness* label to ensure annotation consistency.

Sadness (makes me sad/touches me): If the line/stanza makes one feel sad. It also includes a more general 'being touched / moved'.

Suspense (found it gripping/sparked my interest): Choose *Suspense* if the line/stanza keeps one in suspense (if it excites one or triggers one's curiosity). Suspense is felt in situations when you are excited or anxious about what might happen next. Do you anticipate that there will be a happy ending (excitement) or do you fear that things might end badly (anxiety)? Do you feel a tension or thrill while reading? The associated items are: (1) found it gripping, and (2) sparked my interest. The first item applies when you want to continue reading because you want to know what happens next in the story, or when the story 'grips' you and doesn't let you go. The second item applies when you are intrigued by the message of the poem, or when you continue reading out of curiosity. We removed *Anticipation* from the earlier *Suspense/Anticipation* label, as *Anticipation* appeared to be a more cognitive prediction whereas Suspense is a far more straightforward emotion item.

Uneasiness (found it ugly/unsettling/disturbing / frightening/distasteful): This label covers situations when one feels discomfort, when the line/stanza feels distasteful/ugly, unsettling/disturbing or frightens one. The labels *Ugliness* and *Disgust* were conflated into *Uneasiness*, as both are seldom felt in poetry (being inadequate/too strong/high in arousal), and typically lead to *Uneasiness*. The label Uneasiness can be used when you feel discomfort, or when you find a thought disturbing or even unsettling, or when you are afraid that something bad is going to happen. Please note that Uneasiness is seldom full blown Disgust or Fear. Poetry is rarely used to evoke such strong negative feelings. However, you might encounter poems that make you uncom-

fortable and distressed. You might find texts that describe something ugly or distasteful (like the suffering in war, or xenophobic/racist messages).

Vitality/Energy (found it invigorating/spurs me on/inspires me): This label is meant for a line/stanza that has an inciting, encouraging effect (if it conveys a feeling of movement, energy and vitality which animates to action). Other suitable label terms include *Animated*, *Inspiration*, *Stimulation*, and *Activation*.[7] The word Vitality stems from the Latin language, and Energy from Greek. Vitality might seem odd to German speakers, but it was used in Schindler et al. (2017a), and it is frequently used in video games to indicate the overall and remaining 'life force' (see 'Diablo' or 'The Witcher' series).

5.4.4 Agreement

Table 5.3 shows the Cohen's κ agreement scores among our two expert annotators for each emotion category e as follows. We assign each instance (a line in a poem) a binary label indicating whether or not the annotator has annotated the emotion category e in question. From this, we obtain vectors v_i^e, for annotators $i = 0, 1$, where each entry of v_i^e holds the binary value for the corresponding line. We then apply the κ statistics to the two binary vectors v_i^e. Additionally to averaged κ, we report micro-F1 values in Table 5.4 between the multi-label annotations of both expert annotators as well as the micro-F1 score of a random baseline as well as of the majority emotion baseline (which labels each line as *Beauty/Joy*).

[7] Activation appears stable across cultures (Jackson et al., 2019)

	κ		**Ann. 1 %**		**Ann. 2 %**	
	en	de	en	de	en	de
Beauty / Joy	.77	.74	.31	.30	.26	.30
Sadness	.72	.77	.21	.20	.20	.18
Uneasiness	.84	.77	.15	.19	.15	.18
Vitality / Energy	.50	.63	.12	.11	.18	.13
Awe / Sublime	.71	.61	.07	.06	.07	.06
Suspense	.58	.65	.04	.07	.07	.08
Humor	.81	.68	.04	.05	.04	.05
Nostalgia	.81	—	.03	—	.03	—
Annoyance	.62	.65	.03	.04	.02	.02

Table 5.3: Cohen's kappa agreement levels and normalized line-level emotion frequencies for expert annotators (Nostalgia is not available in the German data).

	English	German
avg. κ	0.707	0.688
F1	0.775	0.774
F1 Majority	0.323	0.323
F1 Random	0.108	0.119

Table 5.4: Top: averaged kappa scores and micro-F1 agreement scores, taking one annotator as gold. Bottom: Baselines.

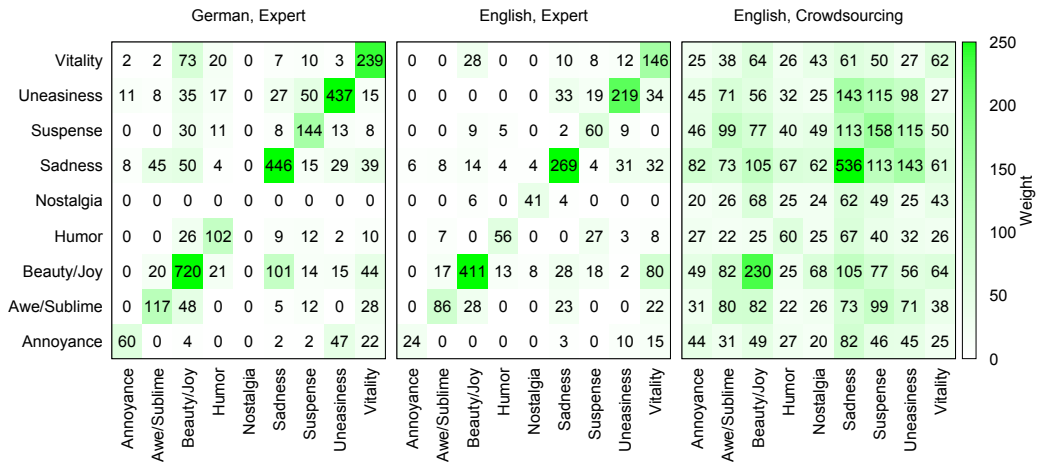

Figure 5.1: Emotion co-occurrence matrices for the German and English expert annotation experiments and the English crowdsourcing experiment.

We find that Cohen κ agreement ranges from .84 for *Uneasiness* in the English data, .81 for *Humor* and *Nostalgia*, down to German *Suspense* (.65), *Awe/Sublime* (.61) and *Vitality/Energy* for both languages (.50 English, .63 German). Both annotators have a similar emotion frequency profile, where the ranking is almost identical, especially for German. However, for English, Annotator 2 annotates more *Vitality/Energy* than *Uneasiness*. Figure 5.1 shows the confusion matrices of labels between annotators as heatmaps. Notably, *Beauty/Joy* and *Sadness* are confused across annotators more often than other labels. This is topical for poetry, and therefore not surprising: One might argue that the beauty of beings and situations is only beautiful because it is not enduring and therefore not to divorce from the sadness of the vanishing of beauty (Benjamin, 2016). We also find considerable confusion of *Sadness* with *Awe/Sublime* and *Vitality/Energy*, while the latter is also regularly confused with *Beauty/Joy*.

5.4.5 Examples of Emotion Annotation

We illustrate two examples of our German gold standard annotation, a poem each by Friedrich Hölderlin and Georg Trakl, and an English poem by Walt Whitman. Hölderlin's text stands out, because the mood changes starkly from the first stanza to the second, from *Beauty/Joy* to *Sadness*. Trakl's text is a bit more complex with bits of *Nostalgia* and, most importantly, a mixture of *Uneasiness* with *Awe/Sublime*. Whitman's poem is an example of *Vitality* and its mixing with *Sadness*. The English annotation was unified by us for space constraints. For the full corpus with annotation please see `https://github.com/tnhaider/poetry-emotion/`

Friedrich Hölderlin: Hälfte des Lebens (1804)

Original Text	Labels	English Gloss (own translation)
Mit gelben Birnen hänget	[Beauty/Joy]	With yellow pears hang
Und voll mit wilden Rosen	[Beauty/Joy]	And full of wild roses
Das Land in den See,	[Beauty/Joy]	The land into the lake,
Ihr holden Schwäne,	[Beauty/Joy]	You fair swans,
Und trunken von Küssen	[Beauty/Joy]	And drunk with kisses
Tunkt ihr das Haupt	[Beauty/Joy]	You dip your heads
Ins heilignüchterne Wasser.	[Beauty/Joy]	Into the holy sober water.
Weh mir, wo nehm' ich, wenn	[Sadness]	Woe is me, where take I, when
Es Winter ist, die Blumen, und wo	[Sadness]	It is winter, The flowers, and where
Den Sonnenschein,	[Sadness]	The sunshine,
Und Schatten der Erde?	[Sadness]	And shadow of earth?
Die Mauern stehn	[Sadness]	The walls stand
Sprachlos und kalt, im Winde	[Sadness]	Speechless and cold, in the wind
Klirren die Fahnen.	[Sadness]	The flags clatter.

Georg Trakl: In den Nachmittag geflüstert (1912)

Original Text	Labels	English Gloss (own translation)
Sonne, herbstlich dünn und zag,	[Beauty/Joy] [Nostalgia]	Sun, autumn thin and timid,
Und das Obst fällt von den Bäumen.	[Beauty/Joy] [Nostalgia]	And the fruit falls from the trees.
Stille wohnt in blauen Räumen	[Beauty/Joy]	Silence dwells in blue rooms
Einen langen Nachmittag.	[Beauty/Joy]	A long afternoon.
Sterbeklänge von Metall;	[Sadness] [Uneasiness]	Dying sounds of metal;
Und ein weißes Tier bricht nieder.	[Sadness] [Uneasiness]	And a white beast breaks down.
Brauner Mädchen rauhe Lieder	[Sadness] [Nostalgia]	Brown girls rough songs
Sind verweht im Blätterfall.	[Sadness] [Nostalgia]	Are blown away in the fall of leaves.
Stirne Gottes Farben träumt,	[Uneasiness] [Awe/Sublime]	Forehead of God dreams colors,
Spürt des Wahnsinns sanfte Flügel.	[Uneasiness] [Awe/Sublime]	Feels the soft wings of madness.
Schatten drehen sich am Hügel	[Uneasiness] [Awe/Sublime]	Shadows turn on the hill
Von Verwesung schwarz umsäumt.	[Uneasiness] [Awe/Sublime]	Fringed by black decay.
Dämmerung voll Ruh und Wein;	[Beauty/Joy]	Twilight full of rest and wine;
Traurige Guitarren rinnen.	[Beauty/Joy]	Sad guitars trickle.
Und zur milden Lampe drinnen	[Beauty/Joy]	And to the mild lamp inside
Kehrst du wie im Traume ein.	[Beauty/Joy]	You return as in a dream.

Walt Whitman: O Captain! My Captain! (1865)

O Captain! my Captain! our fearful trip is done,	[Beauty/Joy]
The ship has weather'd every rack, the prize we sought is won,	[Beauty/Joy]
The port is near, the bells I hear, the people all exulting,	[Beauty/Joy]
While follow eyes the steady keel, the vessel grim and daring;	[Beauty/Joy]
But O heart! heart! heart!	[Sadness]
O the bleeding drops of red,	[Sadness]
Where on the deck my Captain lies,	[Sadness]
Fallen cold and dead.	[Sadness]
O Captain! my Captain! rise up and hear the bells;	[Vitality]
Rise up – for you the flag is flung – for you the bugle trills,	[Vitality]
For you bouquets and ribbon'd wreaths –	
for you the shores a-crowding,	[Vitality]
For you they call, the swaying mass, their eager faces turning;	[Vitality]
Here Captain! dear father!	[Vitality]
This arm beneath your head!	[Vitality]
It is some dream that on the deck,	[Sadness]
You've fallen cold and dead.	[Sadness]
My Captain does not answer, his lips are pale and still,	[Sadness]
My father does not feel my arm, he has no pulse nor will,	[Sadness]
The ship is anchor'd safe and sound, its voyage closed and done,	[Vitality] [Sadness]
From fearful trip the victor ship comes in with object won;	[Vitality] [Sadness]
Exult O shores, and ring O bells!	[Vitality] [Sadness]
But I with mournful tread,	[Sadness]
Walk the deck my Captain lies,	[Sadness]
Fallen cold and dead.	[Sadness]

5.5 Mixed Emotions and Diachronic Emotions

As shown in Figure 5.2 below, we find that emotions are regularly mixed in our dataset. However, no single poem aggregates to more than six emotion labels, while no stanza aggregates to more than four emotion labels. Most lines and stanzas prefer one or two labels. German poems seem more emotionally diverse where more poems have three labels than two labels, while the majority of English poems have only two labels. This is however attributable to the generally shorter English texts.

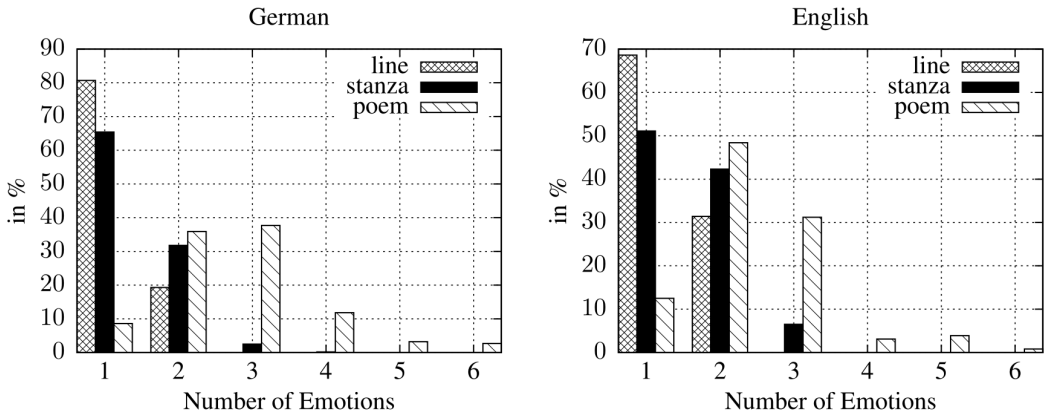

Figure 5.2: Distribution of number of distinct emotion labels per logical document level in the expert-based annotation. No whole poem has more than 6 emotions. No stanza has more than 4 emotions.

In the following, we illustrate which **primary and secondary emotions** occur with each other for a line, and also which emotions are associated with particular **literary periods**. To that end, we use the information theoretic association measure 'Normalized Pointwise Mutual Information' (NPMI). It should be noted that PMI is a discriminative measure and shows the signal that two labels have for each other. It is not a co-variance method like factor analysis. Therefore, if a label (like Beauty/Joy) occurs with many other labels, but is evenly distributed across those, the PMI score will nevertheless stay low.

Only if there is a non-even distribution, PMI will show with which other labels it is more associated and with which less.

NPMI is calculated as follows:

$$PMI(x, y) = log_2 \frac{P(x, y)}{P(x)P(y)} \tag{5.1}$$

$$NPMI(x, y) = \frac{PMI(x, y)}{h(x, y)} \tag{5.2}$$

where

$$h(x, y) = -log_2 p(x, y) \tag{5.3}$$

Pointwise Mutual Information (PMI) denotes the result of the joint probability (relative frequency) of two labels x and y, divided by the individual probabilities of x and y given all labels (here also with log_2). Normalized PMI (NPMI) is PMI divided by the Mutual Information (h) of two labels. This results in real values in the interval $[-1; +1]$, where $+1$ means that two labels occur exclusively with each other, 0 means that two labels have a chance occurrence with each other (random distribution), and -1 means that two labels never occur with each other.

Figure 6.22 shows NPMI values between primary and secondary labels per line for either one annotator (not across annotators), and Figure 6.23 shows NPMI values between any emotion of a line and the period to which the poem belongs. Red-ish colors signify a positive association, while blue/green colors show a negative association. Values at -1 are colored black, since those labels never occur with each other. It should be noted that NPMI associations only show tendencies in the data through pointwise correlations, and that the results here can not be understood as causal relationships. The outcomes are further obscured by the fairly small dataset (153 German poems), and some labels being fairly sparse.

We can see in Figure 6.22 that a non-existent **secondary emotion** (NONE) is negatively associated with every **primary emotion**, meaning that there is no primary emotion that prefers to be annotated alone. Also note that no

emotion will occur with itself in primary and secondary position and thus will always have a NPMI of −1 with itself.

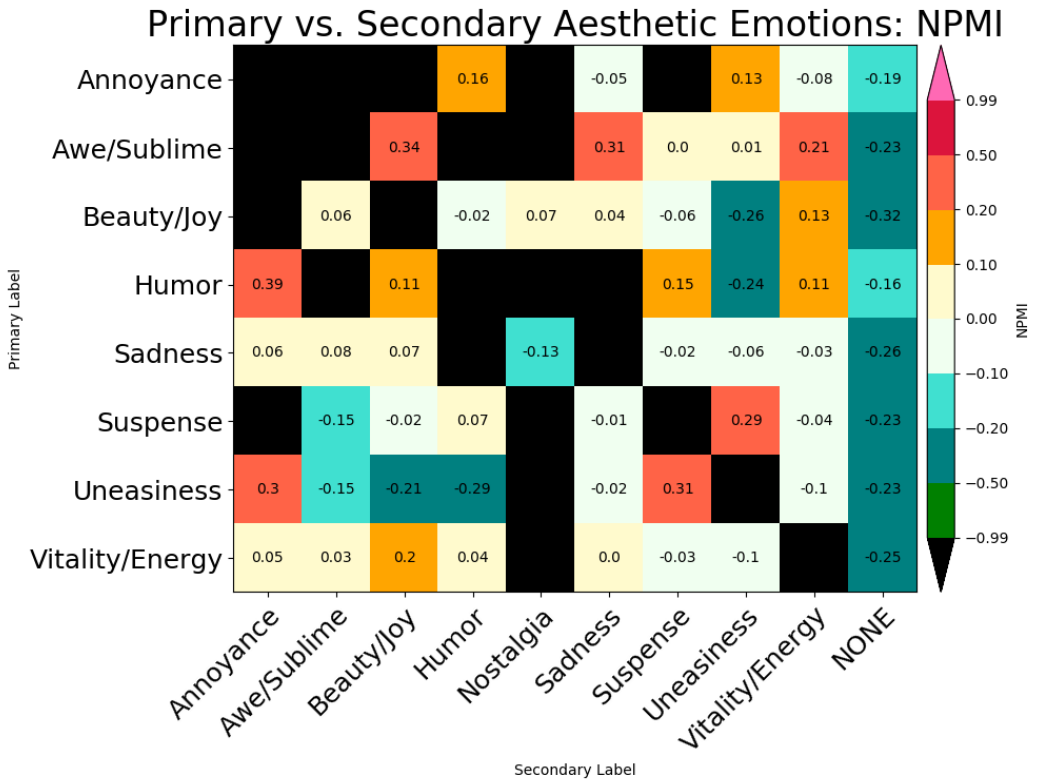

Figure 5.3: NPMI Association b/w Primary and Secondary Emotions

Furthermore, since we are looking at the German dataset here, Nostalgia is only annotated with Beauty/Joy and Sadness, but without clear preference for either (and Nostalgia was exclusively annotated as a secondary emotion). The most frequent emotions Beauty/Joy and Sadness are randomly distributed over most secondary emotions, but there is a tendency that Beauty/Joy is accompanied by Vitality/Energy, and that it rather unlikely that a primary Beauty/Joy is accompanied by a negative emotion Uneasiness or Annoyance. Opposed to the observation we made in the previous section, that Beauty/Joy and Sadness are confused across annotators, they do not occur above chance with each other in a line for a single annotator. However, this still shows that these two labels occur with each other, only that this does

not happen strikingly often. However, it is remarkable that Awe/Sublime is fairly frequently accompanied by both Beauty/Joy and Sadness, and also Vitality/Energy, while Annoyance and Humor don't seem to fit with it.

A primary emotion Humor is avoided by Uneasiness and Sadness, but also by Awe/Sublime. However, there are a substantial amount of lines where Humor is accompanied by Annoyance. This likely shows that Humor does not go well with intense negative emotions (Sadness and Uneasiness) and also that amusement is not associated with an overwhelming feeling of greatness (Awe/Sublime), but that readers can be annoyed by particular types of humor. However, Humor goes well with a feeling of Energy and Beauty/Joy, and also humorous texts can be suspenseful, but just as well that particular humor can annoy the readers. The two negative emotions Uneasiness and Annoyance accompany each other, but Uneasiness can be associated with Suspense, while Annoyance never occurs with Suspense. Suspense also avoids Awe/Sublime.

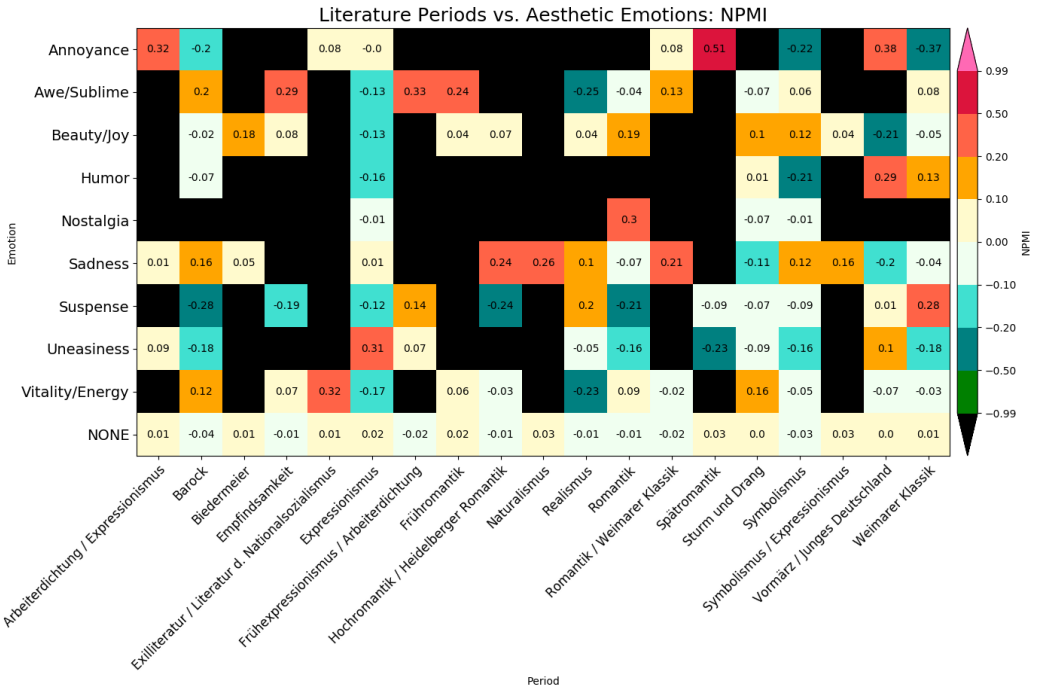

FIGURE 5.4: NPMI Association b/w Time Periods and Emotions

The association of **elicted emotions** from poems within particular **literary periods** is shown in Figure 6.23. It should be noted that for some periods we only have a few poems in the dataset, as can be seen in Figure 3.7. Here, we use annotation from both annotators, both their primary and secondary emotions.

We find that expressionistic poems are quite associated with Uneasiness, but other emotions occur below chance in that literary period. The other movement in Modernity, Symbolism, is associated with Sadness and Beauty/Joy, but avoids Humor, Annoyance and Uneasiness, which brings it closer to romantic poetry. The labels for Romanticism are dispersed over a number of labels, ranging from early Romanticism (Frühromantik), over high Romanticism (Romantik, Hochromantik, Heidelberger Romantik), up to late Romanticism (Spätromantik). Romantic poetry appears to have a tendency for Sadness, but avoids Suspense and Uneasiness. Romanticism furthermore elicits Nostalgia and Beauty/Joy above chance. Suspense is also avoided in Barock and poems in Sensibility, but it appears that poetry from the Weimar Classicism period is suspenseful and humorful above chance. Awe/Sublime is found in poetry of Barock, Sensibility, early Expessionism/Worker's poetry (Arbeiterdichtung) and early Romanticism/Weimar Classicism. Some poems in late Romanticism and Worker's poetry are apparently quite annoying. On the other hand, Barock poems, poems of Expatriats/literature in National Socialism and Sturm und Drang poetry inspire to action as they are associated with Vitality/Energy. Vormärz poetry (which is necessarily political) shows a strong preference for Uneasiness and Annoyance, but these poems can be also humorous (pointing at the divide of patriotic poems, which can be annoying to read from a contemporary perspecive, and satirical poems that make fun of such patriotism). Finally, poems of Expressionism and Realism avoid Vitality/Energy, while Realism also avoids Awe/Sublime, speaking to its avoidance of themes surrounding grandeur and overwhelming imagery.

5.6 Crowdsourcing Annotation

After concluding the expert annotation, we performed a focused crowdsourcing experiment, based on the final label set and items as they are listed in Table 5.5 and Section 5.4.2. With this experiment, we aim to understand whether it is possible to collect reliable judgements for aesthetic perception of poetry from a crowdsourcing platform. A second goal is to see whether we can replicate the expensive expert annotations with less costly crowd annotations.

We chose a simple annotation environment, where we asked participants to annotate English 4-line stanzas with self-perceived reader emotions. We chose English due to the higher availability of English language annotators on crowdsourcing platforms. Each annotator rates each stanza independently of surrounding context.

5.6.1 Data and Setup

For consistency and to simplify the task for the annotators, we opted for a trade-off between completeness and granularity of the annotation. Specifically, we selected stanzas composed of four verses from the corpus of 64 hand selected English poems. The resulting selection of 59 stanzas is uploaded to the platform 'Figure Eight'[8] for annotation. The annotators were asked to answer the following questions for each instance:

Question 1 (single-choice): Read the following stanza and decide for yourself which emotions it evokes.

Question 2 (multiple-choice): Which additional emotions does the stanza evoke?

The answers to both questions correspond to the emotion labels we defined to use in our annotation, as described in Section 5.4.2. We add an additional answer choice "None" to Question 2 to allow annotators to say that a stanza does not evoke any additional emotions.

[8] `https://www.figure-eight.com/`

Each instance is annotated by ten people. We restrict the task geographically to the United Kingdom and Ireland and set the parameters on Figure Eight to only have the highest quality annotators join the task. We pay €0.09 per instance. The final cost of the crowdsourcing experiment is €74.

5.6.2 Results

	κ					Counts				
Threshold	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5
Beauty / Joy	.21	.41	.46	.28	–	34.58	15.98	7.51	3.23	1.43
Sadness	.43	.47	.52	.02	−.04	43.34	28.99	17.77	9.52	2.82
Uneasiness	.18	.25	.08	−.01	–	36.47	16.33	5.49	1.54	1.04
Vitality	.15	.26	.19	–	–	25.62	7.34	2.02	1.05	1.00
Awe / Sublime	.31	.17	.37	.46	–	29.8	11.36	3.4	1.31	1.00
Suspense	.11	.29	.21	.26	–	39.12	17.8	6.54	1.97	1.04
Humor	.19	.46	.39	≈0	–	19.26	5.36	2.1	1.22	1.07
Nostalgia	.23	.01	−.02	–	–	30.52	10.16	1.95	1.00	1.00
Annoyance	.01	.07	.66	0	–	26.54	6.17	1.35	1.00	1.00
Average	0.20	0.27	0.32	0.14	−0.04	31.69	13.28	5.35	2.43	1.27

Table 5.5: Results obtained via boostrapping for annotation aggregation. The row *Threshold* shows how many people within a group of five annotators should agree on a particular emotion. The column labeled *Counts* shows the average number of times certain emotion was assigned to a stanza given the threshold. Cells with '–' mean that neither of two groups satisfied the threshold.

In the following, we determine the best aggregation strategy regarding the 10 annotators with bootstrap resampling. For instance, one could assign the label of a specific emotion to an instance if just one annotators picks it, or one could assign the label only if all annotators agree on this emotion. To evaluate this, we repeatedly pick two sets of 5 annotators each out of the 10 annotators for each of the 59 stanzas, 1000 times overall (i.e., 1000×59 times, bootstrap resampling). For each of these repetitions, we compare the agreement of these two groups of 5 annotators. Each group gets assigned with an adjudicated emotion which is accepted if at least one annotator picks it, at least two annotators pick it, etc. up to all five pick it.

The results can be seen in Table 5.5. The κ scores show the average agreement between the two groups of five annotators, when the adjudicated class is picked based on the particular threshold of annotators with the same label choice. We see that some emotions tend to have higher agreement scores than others, namely *Annoyance* (.66), *Sadness* (up to .52), and *Awe/Sublime*, *Beauty/Joy*, *Humor* (all .46). The maximum agreement is reached mostly with a threshold of 2 (4 times) or 3 (3 times).

We further show in the same table the average numbers of labels from each strategy. Obviously, a lower threshold leads to higher numbers (corresponding to a disjunction of annotations for each emotion). The drop in label counts is comparably drastic, with on average 18 labels per class. Overall, the best average κ agreement (.32) is less than half of what we saw for the expert annotators (roughly .70). Crowds especially disagree on many more intricate emotion labels (Uneasiness, Vitality/Energy, Nostalgia, Suspense).

We visualize how often two emotions are used to label an instance in a confusion table in Figure 5.1. Sadness is used most often to annotate a stanza, and it is often confused with Suspense, Uneasiness, and Nostalgia. Further, Beauty/Joy partially overlaps with Awe/Sublime, Nostalgia, and Sadness.

On average, each crowd annotator uses two emotion labels per stanza (56% of cases); only in 36% of the cases the annotators use one label, and in 6% and 1% of the cases three and four labels, respectively. This contrasts with the expert annotators, who use one label in about 70% of the cases and two labels in 30% of the cases for the same 59 four-liners. Concerning frequency distribution for emotion labels, both experts and crowds name Sadness and Beauty/Joy as the most frequent emotions (for the 'best' threshold of 3) and Nostalgia as one of the least frequent emotions. The Spearman rank correlation between experts and crowds is about 0.55 with respect to the label frequency distribution, indicating that crowds could replace experts to a moderate degree when it comes to extracting, e.g., emotion distributions for an author or time period. Now, we further compare crowds and experts in terms

of whether crowds could replicate expert annotations also on a finer stanza level (rather than only on a distributional level).

5.7 Comparing Experts with Crowds

To gauge the quality of the crowd annotations in comparison with our experts, we calculate agreement on the emotions between experts and an increasing group size from the crowd. For each stanza instance s, we pick N crowd workers, where $N \in \{4, 6, 8, 10\}$, then pick their majority emotion for s, and additionally pick their second ranked majority emotion if at least $\frac{N}{2} - 1$ workers have chosen it.[9] For the experts, we aggregate their emotion labels on stanza level, then perform the same strategy for selection of emotion labels. Thus, for s, both crowds and experts have 1 or 2 emotions. For each emotion, we then compute Cohen's κ as before. Note that, compared to our previous experiments in Section 5.6.2 with a threshold, each stanza now receives an emotion annotation (exactly one or two emotion labels), both by the experts and the crowd-workers.

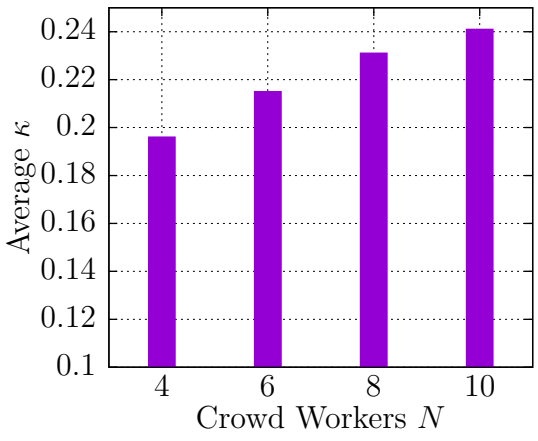

Figure 5.5: Agreement between experts and crowds as a function of the number N of crowd workers.

[9]For workers, we additionally require that an emotion has been chosen by at least 2 workers.

In Figure 5.5, we plot agreement between experts and crowds on stanza level as we vary the number N of crowd workers involved. On average, there is roughly a steady linear increase in agreement as N grows, which may indicate that $N = 20$ or $N = 30$ would still lead to better agreement. Concerning individual emotions, *Nostalgia* is the emotion with the least agreement, as opposed to *Sadness* (in our sample of 59 four-liners): the agreement for this emotion grows from .47 κ with $N = 4$ to .65 κ with $N = 10$. *Sadness* is also the most frequent emotion, both according to experts and crowds. Other emotions for which a reasonable agreement is achieved are *Annoyance, Awe/-Sublime, Beauty/Joy, Humor* ($\kappa > 0.2$). Emotions with little agreement are *Vitality/Energy, Uneasiness, Suspense, Nostalgia* ($\kappa < 0.2$).

By and large, we note from Figure 5.1 that expert annotation is more restrictive, with experts agreeing more often on particular emotion labels (seen in the darker diagonal). The results of the crowdsourcing experiment, on the other hand, are a mixed bag as evidenced by a much sparser distribution of emotion labels. However, we note that these differences can be caused by 1) the disparate training procedure for the experts and crowds, and 2) the lack of opportunities for close supervision and on-going training of the crowds, as opposed to the in-house expert annotators.

In general, however, we find that substituting experts with crowds is possible to a certain degree. Even though the crowds' labels look inconsistent at first view, there appears to be a good signal in their *aggregated* annotations, helping to approximate expert annotations to a certain degree. The average κ agreement (with the experts) we get from $N = 10$ crowd workers (0.24) is still considerably below the agreement among the experts (0.70).

5.8 Modeling Emotions & Transfer Learning

To estimate the difficulty of automatic classification of our data set, we perform multi-label[10] document classification (of stanzas) with BERT (Devlin et al., 2019). For this experiment we aggregate all labels for a stanza and sort them by frequency, both for the gold standard and the raw expert annotations. As can be seen in Figure 5.2, a stanza bears a minimum of one and a maximum of four emotions. Unfortunately, the label *Nostalgia* is only available 16 times in the German data (the gold standard) as a second label (as discussed in Section 5.4.2). None of our models was able to learn this label for German. Therefore we omit it, leaving us with eight proper labels.

We use the code and the pre-trained BERT models of Farm,[11] provided by `deepset.ai`. We test the multilingual-uncased model (Multiling), the german-base-cased model (Base),[12] the german-dbmdz-uncased model (Dbmdz),[13] and we tune the Base model on 80k stanzas of the German Poetry Corpus DLK (Haider and Eger, 2019) for 2 epochs, both on token (masked words) and sequence (next line) prediction ($\text{Base}_{\text{Tuned}}$).

We split the randomized German dataset so that each label is at least 10 times in the validation set (63 instances, 113 labels), and at least 10 times in the test set (56 instances, 108 labels) and leave the rest for training (617 instances, 946 labels).[14] We train BERT for 10 epochs (with a batch size of 8), optimize with entropy loss, and report F1-micro on the test set. See Table 5.6 for the results.

[10] We found that single-label classification had only marginally better performance, even though the task is simpler.

[11] `https://github.com/deepset-ai/FARM`

[12] There was no uncased model available.

[13] `https://github.com/dbmdz` a model by the Bavarian state library that was also trained on literature.

[14] We do the same for the English data (at least 5 labels) and add the stanzas to the respective sets.

	German		Multiling.	
Model	dev	test	dev	test
Majority	.212	.167	.176	.150
MULTILING	.409	.341	.461	.384
BASE	.500	.439	–	–
$\text{BASE}_{\text{TUNED}}$	.477	.514	–	–
DBMDZ	.520	**.520**	–	–

Table 5.6: BERT-based multi-label classification on stanza-level.

We find that the multilingual model cannot handle infrequent categories, i.e., *Awe/Sublime, Suspense* and *Humor*. However, increasing the dataset with English data improves the results, suggesting that the classification would largely benefit from more annotated data. The best model overall is DBMDZ (.520), showing a balanced response on both validation and test set. See Table 5.7 for a breakdown of all emotions as predicted by the this model. Precision is mostly higher than recall. The labels *Awe/Sublime*, *Suspense* and *Humor* are harder to predict than the other labels.

Label	Precision	Recall	F1	Support
Beauty/Joy	0.5000	0.5556	0.5263	18
Sadness	0.5833	0.4667	0.5185	15
Uneasiness	0.6923	0.5625	0.6207	16
Vitality/Energy	1.0000	0.5333	0.6957	15
Annoyance	1.0000	0.4000	0.5714	10
Awe/Sublime	0.5000	0.3000	0.3750	10
Suspense	0.6667	0.1667	0.2667	12
Humor	1.0000	0.2500	0.4000	12
micro avg	0.6667	0.4259	0.5198	108
macro avg	0.7428	0.4043	0.4968	108
weighted avg	0.7299	0.4259	0.5100	108
samples avg	0.5804	0.4464	0.4827	108

Table 5.7: Recall and precision scores of the best model (dbmdz) for each emotion on the test set. 'Support': number of instances with this label.

The BASE and $\text{BASE}_{\text{TUNED}}$ models perform slightly worse than DBMDZ. The effect of tuning of the BASE model is questionable, probably because of the restricted vocabulary (30k). We found that tuning on poetry

does not show obvious improvements. Lastly, we find that models that were trained on lines (instead of stanzas) do not achieve the same F1 (~.42 for the German models).

5.8.1 Transfer Learning

Transfer learning is the process of transferring knowledge between different but related tasks or domains to improve the performance of a model. This could potentially alleviate the problem of small data in our task of classifying stanzas by their emotion, by leveraging knowledge from training a model on a different task and transferring that knowledge to the task at hand.

The BERT language model by Devlin et al. (2019) makes use of this technique by pre-training the base model with unsupervised tasks on a large amount of unlabeled data. The pre-trained model can then be further fine-tuned using labeled data for specific downstream NLP tasks. Phang et al. (2018) further expand on this idea by suggesting an approach called Supplementary Training on Intermediate Labeled-data Tasks (STILTs), which adds a second stage of pre-training with an intermediate supervised task before the final target downstream task. STILTs can improve the overall performance and reliability of a downstream model, particularly when the target task only has a small amount of training data available.

Figure 5.7 illustrates this workflow. First, the initial model is tuned on a first task, where the tuned parameters are propagated back through the whole model, consequently impacting the weights in the underlying model. This model is then taken to be tuned on a second task, in our case to learn the aesthetic emotions of PO-EMO. The results from this section are taken from the Bachelor thesis of Thanh Tung Linh Nguyen, which I co-supervised, and supplied resources for (e.g., the datasets and regular expressions for verse measures). The wording here is also mostly mine.

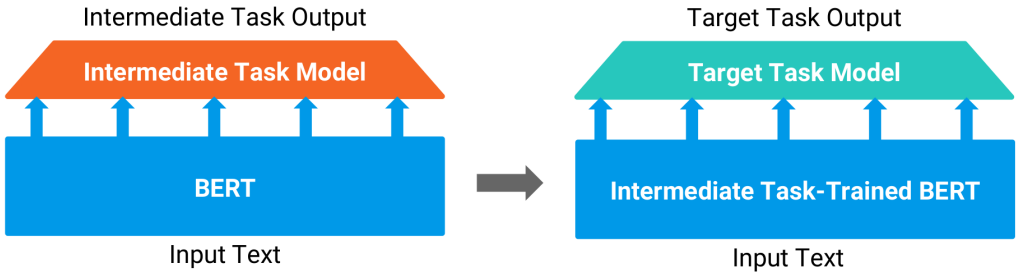

FIGURE 5.6: Illustration of Workflow for BERT on STILTS

We adapt fine-tuning BERT and STILTs in our experiments using this two-step setting. We use BERT-large as the baseline model and fine-tune it with either unsupervised (on more poetry) or supervised tasks (basic emotion recognition and verse meter classification), before fine-tuning it with labeled data from PO-EMO for our classification task. After that, we compare the results to see if there is any improvement over the baseline results. We thus perform the following three experiments: poetry fine-tuning, emotion fine-tuning and meter fine-tuning.

In this experiment, we use the English data from EPG64 (see section 3.5.2). We split the data (by poems) in half, use one half for training and the other half for testing. Then we turn the setting around by training on the second set and evaluating on the first set. In the end, we average the results. The results from these experiments can be seen in Figure 5.7.

5.8.2 Baseline Model

The baseline system is simply an English BERT-large model that was not tuned on any additional data. We just tune the model on the aesthetic emotion recognition task, aggregated to stanzas as was done in the section above for the German data. The baseline model achieves .37 F1-macro score. This is just a bit better than a majority baseline (see section 5.4.4 for human agreement and F1 baselines).

5.8.3 Unsupervised Fine Tuning on More Poetry

We fine-tune a model on a random sample of 30,000 stanzas from the English Gutenberg Poetry Corpus (see section 3.4). These stanzas are unlabeled, and thus the training (fine-tuning) is done in an unsupervised setting, by further training the masked language model of BERT. We do not carry out a next-sentence prediction. In the Experiment on German we saw a bit of improvement, but since here, with the English data, we are dealing only with a fraction of the data, transfer learning should help more. And in fact, the poetry tuned model achieves .50 F1-macro on aesthetic emotions, getting it to a performance comparable to what was achieved on the German data, despite less resources.

5.8.4 Supervised Fine Tuning on Basic Emotion

To test whether basic emotion terms help to understand aesthetic emotions, we perform a supervised tuning step on fairy tales annotated with basic emotions. The corpus comes from Alm et al. (2005a), containing 185 children stories by B. Potter, H.C. Andersen and the Grimm brothers. In total, the dataset contains 14,000 sentences, each annotated for a basic emotion term: Anger, Disgust, Fear, Happiness, Neutral, Sadness, Positive Surprise and Negative Surprise. Unfortunately, their low inter-annotator agreement ($\kappa = .24 - .51$) might hamper the results. But these agreement numbers show that the task of assigning basic emotion terms to sentences is hard. Like our PO-EMO dataset, the Alm et al. (2005a) dataset is imbalanced wrt. to the used emotion terms. There are many instances with Neutral or Happiness labels, but only a few with Disgust and Fear.

We randomly choose 10,000 sentences for training and leave the rest for testing. The result for the intermediate task can be seen in Table 5.8. Our emotion classification model for fairy tales performs well on labels that are common in the tales corpus, such as Happiness, Neutral and Sadness, but

overall the performance of the model reflects the problems of agreement in the manual annotation, and that infrequent emotions (such as Disgust) are hard to detect.

Emotion	F1-score
Anger	0.50
Disgust	0.32
Fear	0.44
Happiness	0.58
Neutral	0.94
Sadness	0.58
Positive Surprise	0.28
Negative Surprise	0.30
Avg. F1-macro Score	0.49

Table 5.8: Classification of Sentences from Fairy Tales with Basic Emotions with BERT-large.

We then use the resulting model for our downstream poetry classification task with a multi-label classification layer. Overall, this model performs better than the baseline model, achieving .43 F1-macro, but fails to detect Annoyance, which was also problem of the baseline model (though Annoyance is not very frequent).

5.8.5 Supervised Fine Tuning on Meter

Our third transfer learning experiment uses meter/verse measure (the rhythm/rhythmic form of poetry), a domain which has been linked to the emotions poems elicited in readers (Obermeier et al., 2013). Meter is one of the formal constraints of poetry, which also contributes to the overall aesthetic of the art form. It describes the rhythmic structure of a poetry line by denoting the stress of syllables within it (+ denotes a stressed syllable, and − an unstressed syllable). Meter usually follows certain patterns that repeat multiple times in a stanza or the poems. The specifics of rhythm and meter in poetry will be discussed in the next Chapter (6), and also how we can learn verse measure labels. For this experiment, we use the for-better-for-verse dataset which is

also explained in more detail there. It contains 1,412 lines with meter annotation, which we transform to labels with regular expressions. For example, the pattern '-+|-+|-+|-+|-+' results in the label 'iambic.penta', since a iambic pattern '-+' is repeated five times (these labels we call 'verse measure', since they refer to the overall 'measure' of the line, and not the raw meter). See the Appendix for the used rules to derive the measure from the raw meter annotation.

See Table 5.9 for how well BERT-large learns verse measure labels from the text of the line. The result is underwhelming, and further experiments in Chapter 6 show that even with expensive uptraining methods BERT does not learn meter that well, only up to .6 F1-macro, which did not help learning emotions on a larger scale. The model here is not able to distinguish iambic lines from trochaic ones, but is probably picking up some signal from the length of the line, and through spurious correlations of meter with semantics.

Verse Measure Label	F1 macro score
iambic.di	0.46
iambic.tri	0.56
iambic.tetra	0.55
chol.iamb	0.10
alexandr.iambic.hexa	0.38
troch.tetra	0.00
iambic.penta	0.80
other	0.29
avg. F1-macro Score	0.39

Table 5.9: Learning Verse Measure Labels with BERT large

However, it should be noted that the dataset with the meter annotation here is a different dataset than with the emotion annotation. Thus, we can expect some transfer. Considering that the intermediate dataset is magnitudes smaller (1,400 lines vs. 30,000 stanzas), an overall result of .48 F1-macro is promising compared to the .50 F1-macro with the larger dataset.

A contributing factor to the increases in performance here is also that this experiment is decidedly in a low-resource setting. As sketched in the

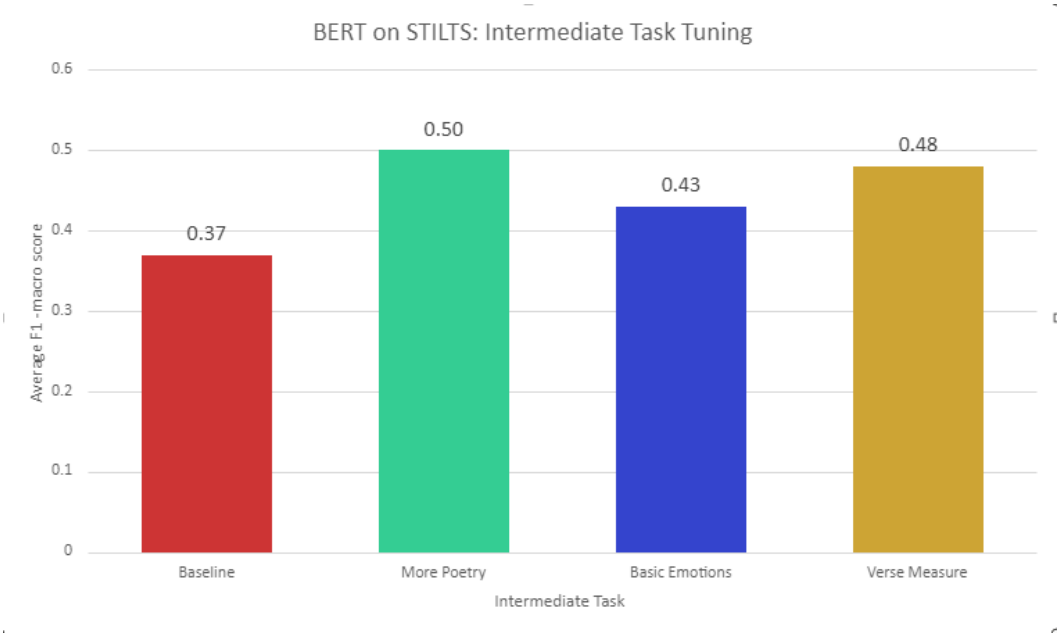

FIGURE 5.7: Intermediate Task Tuning BERT on STILTS for PO-EMO

next chapter, using meter information to improve a model that was already trained on significantly more (German) data is challenging. Still, these results suggest that there is a systematic relationship between the rhythm (verse measure) and the elicited emotions of poetry. Futher experiments are needed to determine when exactly the rhythm of a line interfers with the content of a line to elicit a different emotion. In cases like these, we would expect a change in the decision boundary in a model, such that the predicted emotion for a particular line changes if the model has access to only one signal or the other. First experiments with a logistic regression on a combination of features from word forms, BERT, and verse measure was promising in this regard, but the results are not fully conclusive yet. Future work should look into causal modeling, such that it can be determined which stimulus/signal (e.g., from rhythm or semantics) has an influence in which circumstances.

5.9 Concluding Remarks

We annotated German and English poetry with emotional and aesthetic reader response to reading poetry. We argued that basic emotions as proposed by psychologists (such as Ekman and Plutchik) that are often used in emotion analysis from text are of limited use for understanding poetry reception. We instead conceptualized aesthetic emotion labels and showed that a closely supervised annotation task results in substantial agreement—in terms of κ score—on the final dataset. The task of collecting reader-perceived emotion response to poetry in a crowdsourcing setting is not straightforward however. In contrast to expert annotators, who were closely supervised and reflected upon the task, the annotators on crowdsourcing platforms are difficult to control and may lack necessary background knowledge to perform the task at hand. However, using a larger number of crowd annotators may lead to finding an aggregation strategy with a better trade-off between quality and quantity of adjudicated labels. For future work, we thus propose to repeat the experiment with larger number of crowdworkers, and develop an improved training strategy that would suit the crowdsourcing environment.

The dataset presented in this chapter can be of use for different application scenarios, including multi-label emotion classification, style-conditioned poetry generation, investigating the influence of rhythm/prosodic features on emotion, or analysis of authors, genres and diachronic variation (e.g., how emotions are represented differently in certain periods), some of which we addressed with first experiments. Further, though our modeling experiments are still rudimentary, we propose that this data set can be used to investigate the intra-poem relations either through multi-task learning (Schulz et al., 2018), with the help of hierarchical sequence classification approaches, and, maybe most promising with causal models to learn more about the influence of rhythm and other formal features of poetry regarding the emotions it elicits, in contrast to its semantics.

Modeling Prosody 6

6.1 Introduction

Poetry as an oral art form likely predates written language (Beissinger, 2012). Metrical verse, lyric as well as epic, was already common in preliterate cultures, and to this day the majority of poetry across the world is drafted in verse (Fabb and Halle, 2008). To analyse oral traditions and records of performing arts such as poetry, literary scholars mainly study textual resources (rather than audio, as usually a speech recording is not available for old poems). The rhythmical analysis of written poetic verse is still widely carried out by example- and theory-driven manual annotation of experts, through so-called close reading (Carper and Attridge, 2020; Kiparsky, 2020; Attridge, 2014; Menninghaus et al., 2017). Fortunately, well-defined constraints and the regularity of metrically bound language aid the prosodic interpretation of poetry. A similar case applies to the analysis of rhyme. Many studies on rhyme have been decidedly small scale, where rhyme data was selected by hand (Knoop et al., 2019; Katz, 2015; Primus, 2011).

However, for projects that work with larger text corpora, close reading and extensive manual annotation are neither practical nor affordable. And

while the speech processing community explores end-to-end methods to detect and control the overall personal and emotional aspects of speech, including fine-grained features like pitch, tone, speech rate, cadence, and accent (Valle et al., 2020), applied linguists and digital humanists still rely on rule-based tools for the analysis of prosody (Plecháč, 2020; Anttila and Heuser, 2016; Kraxenberger and Menninghaus, 2016), where some of these tools have limited generality. For example, Navarro-Colorado (2018a); Navarro et al. (2016) rely on part-of-speech tags and some language specific rules to estimate the stress of syllables, or other tools were never properly evaluated (Bobenhausen, 2011), making the results of their application to literature questionable.

Other approaches to computational prosody are based on words in prose rather than syllables in poetry (Talman et al., 2019; Nenkova et al., 2007) (where the task is to assign rhythmic stress values to words in prose sentences), rely on lexical resources with stress annotation such as the CMU dictionary (Hopkins and Kiela, 2017; Ghazvininejad et al., 2016) (limiting applicability to out-of-vocabulary words and neglecting contextual effects), are in need of an aligned audio signal (Rosenberg, 2010; Rösiger and Riester, 2015) (extracting prosodic features from audio and aligning it to the text), or model only narrow domains such as iambic pentameter (Greene et al., 2010; Hopkins and Kiela, 2017; Lau et al., 2018) or Middle High German (Estes and Hench, 2016) (which makes it hard to judge the reliability of these systems in other, more general, domains).

To overcome these limitations in automatic prosodic analysis of poetry, we propose corpus driven neural models to predict the prosodic features of syllables from text directly. These models are evaluated against rhythmically heterogenous data, where the evaluation measures cover both the level of the syllable and also the whole line, i.e., whether the overall measure/sequence of stress values in a line is modeled correctly. We show that BiLSTM-CRF models with syllable embeddings outperform a CRF baseline and different BERT-based approaches. In a multi-task setup, particular beneficial task

relations illustrate the inter-dependence of poetic features. For example, a model learns foot boundaries better when jointly predicting syllable stress, aesthetic emotions and verse measures benefit from each other, and we find that caesuras are quite dependent on syntax and also integral to shaping the overall measure of the line.

Furthermore, to better understand rhyming, we carry out experiments in representation learning, where we train models to learn a similarity metric between phonologically similar rhyming words, and compare these models against baseline character overlap metrics. Such similarity metrics allow us to estimate the 'imperfectness' of a rhyme. Ideally, such as metric reflects the degree of deviations on particular phonological features (such as the locus of pronunciation in the mouth, or whether phonemes are voiced or voiceless, etc.). Hence, we implement unsupervised and supervised systems, i) a system based on character overlap ratio, ii) we test the system of Reddy and Knight (2011) based on Expectation Maximization (EM) on our dataset, and iii) we train and test Siamese Recurrent Networks.

Additionally, even though practically every culture has a rich heritage of poetic writing, large comprehensive collections with prosodic annotation are rare. For prosodic analysis, we present datasets in German and English, encompassing a varied sample of around 7000 manually annotated lines, and we automatically tag large corpora in both languages to advance computational work on literature and rhythm. This may include the analysis and generation of poetry, but also more general work on prosody, or even speech synthesis. Furthermore, for the analysis of rhyming, we created a diachronically balanced corpus and also annotated the previously mentioned smaller corpora for rhyme.

The contribution of this chapter are the documentation of (a) Experiments on the detection of rhyming and learning similarity metrics between phonologically similar words, (b) The annotation of prosodic features in a diverse sample of smaller corpora, including metrical and rhythmical features, mea-

suring their agreement across annotators and examining errors, particularly for foot boundaries. Also, we developed regular expressions to determine verse measure labels. (c) The development of sequence tagging models to jointly learn our prosodic annotations in a multi-task setup, highlighting the relationships of poetic features with each other. (d) With this approach we achieve a level of accuracy for the prediction of prosody that allows the analysis of large corpora. We show plots of how verse measures are distributed over time, and for the characterization of authors by their preferred verse forms. (e) We present first insight into the variation of poetic rhythm by calculating Normalized Pointwise Information (NPMI) for verse annotations w.r.t. the use of verse measures for aesthetic emotions and their prevalence in literary periods. Finally, (f) we examine the role of syntax for prosody by calculating accent ratios for part-of-speech with and without context and which verse measures stand out in the use of enjambement.

6.2 Learning to Rhyme

Rhyme is a pervasive style device in historical poetry, and has remained almost synonymous with poetry itself for centuries, before it saw an abrupt decline in the 20th and 21st centuries. Rhyme has stayed relevant in other art forms like the pop song or rap lyrics, with rap artists like 2Pac metonymically referring to their works as rhymes like Milton or Dante (Knoop et al., 2019).

By definition, rhyme occurs when two or more words are phonologically identical from at least the final stressed vowel onward (Kiparsky, 1973, p. 234), (Fabb, 1997). However, a narrow definition of so-called *perfect rhyme* disregards frequently used and accepted deviations from this convention, as seen in *imperfect rhyme* (Knoop et al., 2019; Primus, 2011; Berg, 1990) or related sonic devices such as half-rhyme, assonance, consonance and alliteration (McCurdy et al., 2015b). At the same time, phonological imperfections in rhyme, i.e., information on the phonological similarity/deviation of two

rhyme words, can be used to scrutinize phonological typology (Katz, 2015; Berg, 1990), for the analysis of sonic patterns (McCurdy et al., 2015b,a), or for the reconstruction of historical pronunciation (List et al., 2017). Unfortunately, for domain specific or historical data, obtaining precise pronunciation information is a challenge (Katz, 2015). Consequently, most studies on rhyme have been decidedly small scale, where rhyme data was mainly selected by hand. However, a system that can detect rhymes reliably across different domains without a reliance on grapheme to phoneme transliteration could open many new directions of such research. This research presents such a system and examines its advantages and shortcomings. Furthermore, a focus of this work is on exploring similarity metrics for rhyming.

We present the first supervised approach to rhyme detection with Siamese Recurrent Networks (SRN) that offer near perfect performance (97% accuracy) with a joint cross-lingual model for German, English and French. This class of Siamese Networks is particularly adept for the task of rhyme detection, as these models also learn a similarity metric on variable length (character) sequences that can be used as judgement on the distance of imperfect rhyme pairs. We find that learning rhyme similarity based on character representations gives us robust models that do not rely on previous steps of (potentially erroneous) grapheme-phoneme transliterations, but that these models learn phonological similarity simply on the basis of character strings (graphemes) from positive and negative examples of rhyme pairs. Unfortunately, it appears that cross-lingual models retain more idiosyncratic representations of phonological similarity and may even learn problematic phonotactics.

For training, we constructed a diachronically balanced rhyme goldstandard of New High German (NHG) poetry. For cross-domain testing, we sampled a second collection of NHG poetry and set of contemporary Hip-Hop lyrics, annotated for rhyme and assonance. We train several high-performing SRN models and compare them to simpler similarity measures and finally evaluate them qualitatively on selected sonnets.

6.2.1 Rhyme Annotation

In order to obtain a dataset that represents a real-world scenario, we create a rhyme schema goldstandard (DTA-RHYME) by drawing a diachronically balanced sample from DTA and then annotating the stanzas manually. To get an even temporal distribution, we divide the timeline by 20-year wide slots (1630 - 1650, ..., 1790 - 1810, ..., 1890 - 1913), and aim to sample to 500 stanzas per slot (allowing +/- 10%). An additional parameter is that an author needs to contribute enough poems within a std. deviation from the mean. No stanzas longer than one standard deviation over 12 lines (24) were allowed. We left the original poems intact, randomly sampling poems from DTA until the desired number of stanzas was fulfilled with complete poems. DTA-RHYME eventually contains 1,948 poems over 8,147 stanzas.

Corpus	Poems	Stanzas	Rhyme Pairs
ANTI-K	156	731	1,440
HIPHOP	116	789	2,489
DTA-RHYME	1,948	8,147	13,784

Table 6.1: Size of German Rhyming Corpora

In addition to DTA-RHYME, we also annotate ANTI-K (as discussed in Chapter 3) and a set of Hip-Hop texts. The size of these corpora is shown in Table 6.1. ANTI-K was was annotated by a competent student and re-checked. ANTI-K yields 1,440 rhyme pairs. We also collected 116 German Hip-Hop song texts and annotated them on rhyme and assonance (repetition of vowels). We retrieved the documents in plain text from `hiphoplyrics.de`, mainly covering the 90's and 2000's, with 1–4 texts per author. Hip-Hop differs from lyric poetry in the regard that it makes heavy use of internal rhyme and assonance. As the annotation of internal rhyme is very time consuming, we confine our analysis to end-rhyme. Yet, assonances and rhymes often form a complex schema, so we decided to mark assonances with capital letters in

the stanza level rhyme schema to extract them separately. We retrieve 2,489 rhyme pairs and 1,032 assonance pairs from the Hip-Hop data.

Three graduate students of literature/linguistics (thereafter called 'annotators') each annotated Georg Trakl's 1913 collection 'Gedichte', in total 251 stanzas in 51 poems for training and to calculate inter-annotator agreement. We then split the corpora among the annotators. We annotate rhyme schema for end-rhyme on stanza level, such as 'aabb', 'abab', or 'abba' for four liners, where matching indices ('aa') indicate a rhyme pair, and non-matching indices ('ab') indicate a non-rhyming pair. Annotators were instructed on the basic definition of rhyme, and discussed with the first author the concept of 'imperfect rhyme'. Strictly speaking, imperfect rhymes violate the requirement that rhyming words be fully identical in the last prominent vocalic nucleus and any following segments. However, they usually show partial identity between those segments at the level of phonological features and thereby often retain a high degree of similarity (Knoop et al., 2019). Generally, though, poetic traditions vary regarding what extent and type of similarity is required in a rhyme pair (Fabb, 1997). Annotation was carried out via silent reading. Certainly, in this scenario, without an audio reference, it cannot always be clear to annotators whether two words rhyme, depending on whether the intended pronunciation matches with their own pronunciation. Also, it can be unclear if the vernacular of an author would have permitted certain rhyming. Thus, if in doubt, annotators should annotate rhyme if two words are clearly not non-rhyme, and sound 'close enough'. They could look up the pronunciation (variants) of words, and they should consider the schema of other stanzas in the poem to follow a 'schema consistency'.

Schemas were annotated directly in inline TEI P5 XML, and we used the Oxygen XML editor to validate the XML.[1] Our annotators achieved .95–.97 Cohen κ on the Trakl poems, measured via stanza schema labels. Unfortunately, reading stanzas vertically and annotating (typing) schemas horizon-

[1]Against a prescriptive RELAXng schema for permissible XML.

tally can be prone to error, leading to false annotation (on average one or two non-matching stanza schemas for Trakl across two annotators).

We then extracted rhyme pairs (word pairs) from the schemas and find that, overall, DTA-RHYME yields 16,440 rhyme pairs. The list of these rhyme pairs was revisited by an annotator, and the first 1,500 pairs were annotated on the type of error. Out of these 1,500 pairs, 244 pairs did not rhyme, as judged through isolated word sounds (without considering the entire stanza or poem), amounting to 16% error rate. 17 further instances were considered assonances (where only the vowels match, but not necessarily the rest of the word), but not bona fide rhymes. In total, from DTA-RHYME, we retain 13,785 (from 16,440, amounting to 8% error rate) rhyme pairs that we use for further experiments.

$Chicago_{EN}$		$DTA-RHYME_{DE}$	
183	aa	354	aa
		223	**ab**
305	aaa	146	**abc**
83	aba	120	aab
19	aab	79	aba
		60	abb
		16	aaa
1417	abab	1639	abab
598	abcb	860	aabb
229	aabb	459	abcb
58	abca	424	abba
57	abba	223	**abcd**
26	aaba	12	abac
4	aaab	11	abcc
4	aaaa	11	aaaa
		8	abbc
		7	aabc
		5	abca
		5	aaab
		2	aaba
		1	aacc

Table 6.2: Rhyme Schema Frequency of 2,3, and 4-liners in English Chicago Corpus and DTA-RHYME

Other datasets that we use in this research originate from Reddy and Knight (2011) and Sonderegger (2011), who provide English and French poems annotated for rhyme schema. These corpora are called the English and French Chicago Rhyming Corpus. In Table 6.2 we compare the frequencies of rhyme schema from the English Chicago Corpus with DTA-RHYME (for couplets, triplets and quartets). Note that the frequencies of rhyme schemas across languages may differ substantially, e.g., 'aaa' is the most frequent schema for 3-line stanzas (triplets) in English, but the least frequent in German. More important however is the observation that the English corpus does not include any stanzas that do in fact *not* rhyme, as seen in the German schemas 'ab', 'abc', and 'abcd'. This may potentially impact any models that are trained on this data, such as the approach of Reddy and Knight (2011) as outlined below. Overall, the German DTA-RHYME corpus contains about a third of stanzas that do not rhyme, thus only two thirds of stanzas actually rhyme. For these reasons we believe that this corpus is of a quality that will allow further research on the distribution of rhyming patterns.

6.2.2 Learning to Rhyme: Experiments

A goal of this research is to investigate algorithms to automatically detect rhyme. Furthermore, we calculate similarity metrics for rhyme words to judge their phonological similarity. Such similarity metrics allow us to estimate the 'imperfectness' of a rhyme. Ideally, such as metric reflects the degree of deviations on particular phonological features (such as the locus of pronunciation in the mouth, or whether phonemes are voiced or voiceless, etc.). Hence, we implement unsupervised and supervised systems, i) a system based on character overlap ratio, ii) we test the system of Reddy and Knight (2011) based on Expectation Maximization (EM) on our dataset, and iii) we train and test Siamese Recurrent Networks.

Character Overlap Our character overlap metrics are based on the python package `difflib ratio`.[2] Basically, this algorithm looks at the characters in two sequences (two words) and calculates the ratio between the overlapping characters versus the non-overlapping characters. The system returns a measure of the sequences' similarity as a real number (float) in the range [0, 1]. Let T be the total number of elements in both sequences, and M the number of matches, then the ratio R is $R = 2 \cdot \frac{M}{T}$. Note that R is 1.0 if the sequences are identical, and 0.0 if they have nothing in common.[3] We use two variants of the algorithm, 1) comparing the unaltered words, and 2) where the words are cut to the same size (of the shorter word). Consider example output from variant 2) in Table 6.3 (where words are cut to same length from the end before calculating overlap).

Word 1	Word 2	Char. Ratio R	Transl. W1	Transl. W2
springen	springen	1.00	jump	jump
liegen	siegen	0.83	lie	win
kunst	dunst	0.80	art	mist
saus	maus	0.75	live high (idiom)	mouse
scheint	geweint	0.57	seems	cried
kalt	wald	0.50	cold	woods
ruh	der	0.33	calm	the
brod	gott	0.25	bread (arch.)	god
weiß	woll	0.25	white/know	want
umzogen	sternen	0.29	move/encompass	stars
keusch	rinnen	0.17	chaste	trickle
dran	kutte	0.00	attached	cloak

Table 6.3: Character Overlap Ratio of Rhyme Words, cut to same length

Identical pairs like (springen, springen) result in R=1.0, while pairs that don't have any characters in common like (dran, kutte) get R=0.0. The pair (kunst, dunst) has 80% overlap in characters, and (saus, maus) have 75%

[2] `https://docs.python.org/3/library/difflib.html`

[3] Note that this metric differs from Jaccard-Coefficient in that T in difflib ratio is the total number of characters from both sequences, while Jaccard uses the union set of characters. In practice (for this application) it should make little difference, but could be investigated in future work.

overlap. Note that the ratios for the latter two pairs are different, even though they only differ in a single character, but are of different length. We evaluate the two variants on the DTA-RHYME dataset with 13,785 pairs. We generate rhyming pairs from non-matching indices. For binary classification, variant 1) (Vanilla difflib) achieves 76% Accuracy, while variant 2) (cut to length difflib) achieves 83% Accuracy. This is also shown in Table 6.4.

Metric	Accuracy
Vanilla difflib Ratio	.76
Cut to length difflib Ratio	.83

Table 6.4: Rhyme Classification Accuracy of difflib ratio on DTA-RHYME Corpus.

6.2.3 Expectation Maximization

Reddy and Knight (2011) use a similar technique of character overlap, but they integrate it into an unsupervised learning framework with Expectation Maximization (EM) to predict (generate) the most probable schema (e.g., 'abba') of a stanza. This EM can be initialized uniformly (EM_uniform), only using information on the frequency of schemas, or also incorporate character overlap similarity between the words (EM_orthography). Basically, the algorithm learns the distribution of schemas for every stanza length from a training corpus, and EM_orthography additionally incorporates character overlap. We train an EM on our two German poetry corpora (ANTI-K and DTA-RHYME) with the code provided by Reddy and Knight (2011).[4]

For certain experiments on English, Reddy and Knight (2011) report accuracy up to 88% F1. We can only replicate these results on their dataset for certain train/test splits. For any train/test sets that are subsets of the full corpus, results are always lower, which confirms the importance of representative schema probabilities in the training set for this approach.

[4] `https://github.com/sravanareddy/rhymediscovery`

Table 6.5 lists the results for two German corpora. We train on ANTI-K and DTA-RHYME respectively, and evaluate both models on ANTI-K (thus, the first model is evaluated in-domain and the second model out-domain). When initialized uniformly (only schema distribution) the in-domain model achieves 63% Accuracy, but the uniform out-domain model only achieves 37%, which is only marginally above the majority baseline (selecting the most frequent schema per line length).

Train Corpus	Test Corpus	EM_uniform	EM_orthography
ANTI-K	ANTI-K	.63	.77
DTA-RHYME	ANTI-K	.37	.71

Table 6.5: Accuracy of EM on German stanzas

Incorporating character overlap (EM_orthography) improves the results significantly for both models, but in-domain Accuracy is yet 6 points higher (.77 vs. .71). These results indicate that the uniform EM mainly learns the majority schemas per stanza length, and is thus heavily dependent on the representativeness of train and test corpora. Compare Table 6.2, from which it becomes clear that schemas are distributed roughly along a normal distribution, where the majority schemas are rather prominent. And even though rhyme pair prediction is not directly comparable to rhyme schema prediction, the EM does not seem to add much benefit over a simple character overlap baseline, and an accuracy of 77% leaves considerable room for improvement.

6.2.4 Siamese Networks

In this work we introduce Siamese Recurrent Networks (SRNs) to the task of rhyme detection and to learn a distance metric between rhyme words. While the two approaches outlined above employ unsupervised algorithms, SRNs are supervised. Siamese recurrent networks (SRN) have been successfully applied in NLP applications to measure the distance of texts, both on the level of

characters and words. Neculoiu et al. (2016) use it for job title normalization, by learning the similarity of character embeddings. Mueller and Thyagarajan (2016) learn the similarity of sentences by using word embeddings, and Das et al. (2016) utilize SRNs the retrieval of similar questions on Quora. Rama (2016) used Siamese Convolutional Networks (not recurrent) for the detection of cognates (which is a similar task). Moreover, Siamese Recurrent Networks have become a best-practice approach to Natural Language Inference (NLI) and for learning sentence embeddings (Reimers and Gurevych, 2019).

The architecture of our Siamese Recurrent Network is illustrated in Figure 6.1. The network consists of two identical recurrent sub-networks that each encode a word (a sequence of characters) to learn a vector representation for each word. Each sub-network receives a word via integer indices of characters, effectively learning character embeddings.

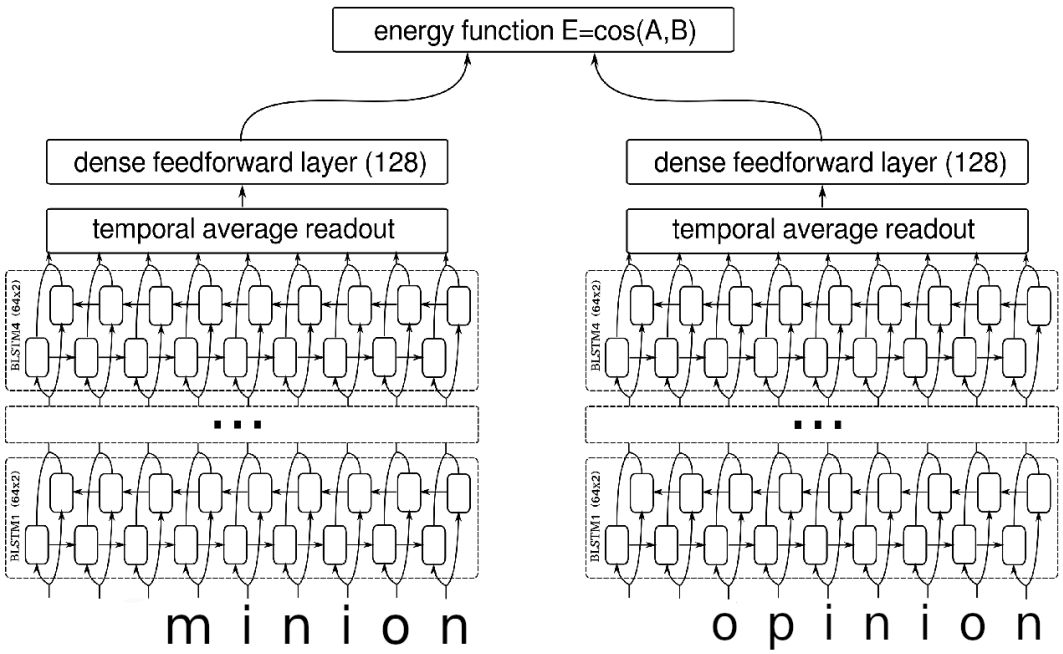

FIGURE 6.1: Architecture of our Siamese Recurrent Network.

These character embeddings are then encoded through several layers of bidirectional Long Short Term Memory (BiLSTM) Networks. The activations at each timestep of the final BiLSTM layer are averaged to produce a fixed-

dimensional output. This output is projected through feed-forward layers (a simple matrix multiplication, also called dense layer). We call the output vector of the first sub-network A, and the output of the second sub-network B. These outputs are then connected via an energy function. We want to make sure that the system learns to discriminate between the inputs in a predictable way. One such system is the energy based model. Learning is done by minimizing the system energy for desirable inputs and not minimizing it for the undesirable inputs. We define our energy function $E = cos(A, B)$ in terms of optimizing the cosine distance between A and B by learning through positive and negative examples (accuracy of a binary classification by using E=0.5) and optimizing against contrastive entropy loss. The energy of the model is then defined by the similarity between the embeddings of word 1 and word 2. Our architecture, as that of Neculoiu et al. (2016) uses a cosine distance as distance metric. In principle, any other vector distance metric would be applicable, like euclidian distance, city block distance, or learning a L2 norm. Furthermore, we use three (later four) stacked BiLSTM layers. After pilot experiments, we choose the following parameters: The size of the BiLSTM layers are set to 100 dimensions. Maximum word length per input is set to 30 characters. If a word is shorter, placeholders are used in front of the word (since we are interested in word endings). We experiment with the number of hidden units between the BiLSTM layers (20 – 100), settling on 50. We train for 100 epochs, use a batch size of 64, set dropout to 1.0 and L2 regularization to 0.0. Lastly, the system uses a random 80/20 train/dev split. The initial codebase for our implementation stems from `github.com/dhwajraj/deep-siamese-text-similarity`.

6.2.5 Binary Classification

First, we are concerned with learning a SRN model that can discriminate whether two words rhyme, i.e., a binary classification task. We experiment

with the ratio of positive to negative examples by selecting 5000 rhyme pairs (all positive) from DTA-RHYME and generate negative examples (non-rhyming pairs) from non-matching indices in rhyme schema.[5] We evaluate/test the final model on 2400 instances, of which 1200 are positive examples and the rest negative examples (as retrieved from DTA-RHYME in the previous cleaning step). At a ratio of 2:3 of positive:negative examples, we already achieve an Accuracy of 96%. Similar results were achieved for the French and English data. Neculoiu et al. (2016) reported an ideal ratio of 1:4 for job title normalization. However, despite the larger amount of data, we only get 93% Accuracy at this ratio. We then test a 1:1 ratio and achieve 97% Accuracy. These experiments were carried out with three BiLSTM layers. By increasing this to four layers, the 1:1 model gains another point to 98% Accuracy. Thus, we use a 1:1 ratio and four layers for further experiments.

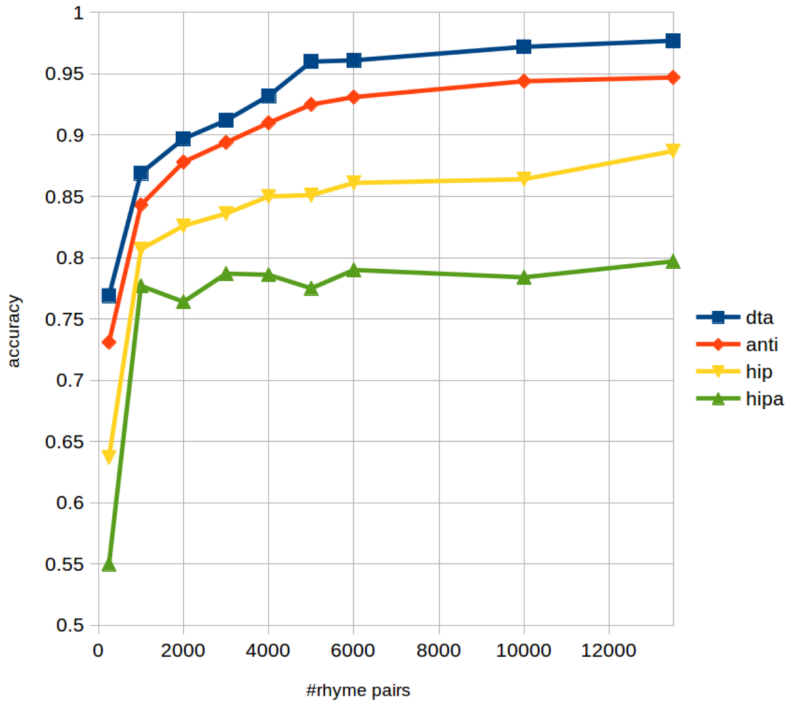

FIGURE 6.2: Learning rate of German rhyming models trained on subsets of DTA-RHYME with 1:1 ratio and 4 layers.

[5]Generating negative examples by random shuffling achieved similar results.

In a next step, we wanted to investigate the learning curve of a model. We tested models that were trained on varying amounts of rhyme pairs from poetry from the DTA-RHYME data (before post-correction of the training set, thus containing up to 16% noise). We test on the previously discussed datasets, i.e., the held out test set from DTA-RHYME (dta), ANTI-K (anti), a Hip-Hop dataset that only contains proper rhymes (hip), and a Hip-Hop dataset that additionally includes assonant word pairs that may not necessarily be proper rhymes, because they may only share same sounding vowels (hipa). The learning curves are plotted in Figure 6.2. The x-axis denotes the number of positive examples of rhyme pairs (thus there are an equal size of negative examples in the training set). We can see that 1000 rhyme pairs are already sufficient to train a model that achieves Accuracy in the high 80s on historical poetry rhymes (dta + anti), while 5000 rhyme pairs are needed for a well performing model with over 90% Accuracy, where triple the training data (15,000 rhyme pairs) does not result in a major performance increase anymore. The reason the ANTI-K test set lags behind the DTA-RHYME test set by 1–3 points is likely because the DTA-RHYME test set was cleaned in a post-correction step and thus is of higher quality. Performance on the Hip-Hop test set is 8–10% lower than on historical poetry. This can be attributed to slang words (that are not present in the poetry data and often in English) and code-switching (German word rhyming with an English word). Furthermore, we find that the model did not learn any representation of assonance, as seen in the low numbers on the Hip-Hop-Assonance (hipa) test set. Particularly recall was low on this test set, indicating that assonances are not detected. Since the pairs retrieved from DTA-RHYME included 16% noise, we decided to again manually correct the rhyme pair training set, but we did not see a performance increase. This indicates that our models can compensate for noise and thus an expensive cleaning step should not be necessary to train a well performing model for rhyme detection.

6.2.6 Language & Domain Dependence

As a next step, we were interested whether we can train a model that is independent of specific languages and if we can increase the out-of-domain performance on Hip-Hop. Learning a language-independent model for rhyme detection from characters can be a challenging task, because the pronunciation of (latin) graphemes may differ significantly across languages, and thus there may exist certain character sequences that sound similar in one language, but not in another. With German, English, and French rhyming data we cover two Germanic languages (German, English) and a Romance language (French). We find that monolingual models for either English or French perform similarly well as the previous German models with around 94–96% Accuracy, given a sufficient amount of training examples (over 5,000). We then train a model on all three languages at the same time (4indep). Each language contributes 10,000 positive examples and 10,000 negative examples, amounting to 60,000 training instances total (where the German examples come from DTA-RHYME only). See Table 6.6 for the evaluation of this model. The language independent model is on par with the monolingual models from previous experiments in terms of detection accuracy. However, as we will see in the qualitative evaluation below, this model appears to learn more idiosyncratic mappings than the monolingual models.

Test-Set	Accuracy	$F1_{micro}$	Precision	Recall
EN_Chicago	.965	.965	.959	.972
FR_Chicago	.963	.963	.960	.966
DE_DTA-RHYME_dev	.976	.976	.970	.981
DE_DTA-RHYME_test	.960	.960	.954	.967
DE_ANTI-K	.961	.960	.965	.956
DE_Hip-Hop	**.892**	.885	.948	**.830**
DE_Hip-Hop + HH in Train	**.920**	.920	.960	**.890**

Table 6.6: Language Independent Rhyme Models Performance across Languages

When testing on Hip-Hop, considering the large size of the training set, the model is on par or even slightly better than the monolingual German model, and we see that recall is similarly low, where certain rhyme pairs are simply not detected. To further try to minimize the error on Hip-Hop, we include 2,000+2,000 (pos.+neg. 1:1) pairs from Hip-Hop into the language independent training dataset and set aside the remaining 978+978 pairs for testing. We train on the same parameters (4indeph), and as documented in Table 6.6, see improvement for Hip-Hop (.92 F1, .96 prec, .89 recall) while all other testsets remain stable and even slightly improve. This shows that augmenting the training set with domain specific data improves the model on that domain, learning domain specific phonological mappings (such as code-switched rhyme).

6.2.7 Error Analysis of Rhyming Models

We have seen that Siamese Networks can learn rhyming very well. However, it remains an open question what exactly these models learn and why and when they are better than character-overlap metrics. So far, rhyme annotation has been viewed as a binary classification task. Thus, a model should learn some representation of rhyme that optimizes a decision boundary between rhyme and non-rhyme.

We test the German 4layer11 model against difflib and cut_difflib (where words are first cut to identical length before calculating the difflib character overlap ratio) on all word pairs in the ANTI-K test corpus. The ratio of rhyming to non-rhyming examples in this test set is 1:1, so that we should get a representative cross section across the spectrum of phonological similarity in rhyme. To see how our metrics represent/learn the similarity of rhyming words, we calculate the similarity with all three metrics.

First, we test the distribution of scores that the metrics determine for the similarity of two words. Second, we examine instances where the metrics

disagree and whether that leads to correct or incorrect predictions. Third, we test our different siamese models on a use case: We test how well they detect rhyme in three sonetts from different time periods, and which mistakes they make.

6.2.7.1 Siamese Networks vs. Naive Character Overlap:

Separation of Classes: First of all, to see how our metrics represent rhyme, we plot the density of the similarities of all ANTI-K rhyme pairs in Figure 6.3. If many instances are similar around a particular value, then the density at that point should be high, but density should be low when few instances accumulate at a particular value. A metric that learned rhyme should have high density at low similarities (many words do not rhyme) and at high similarities (many words do rhyme), but a low density in the middle (few words are neither rhymes or non-rhymes). The density plot shows that a siamese network (cossim) does in fact learn exactly that. I should be noted though that the maxima of the density is never at the extreme of 0 or 1.

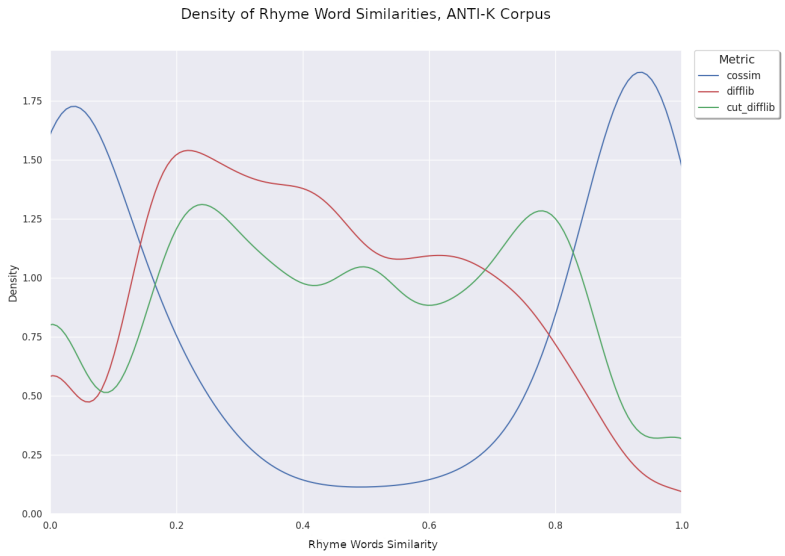

FIGURE 6.3: Similarity of Rhyme Pairs for Different Metrics, Density Plot, bandwidth=0.2

The character overlap ratios (difflib) do not represent rhyming that clearly. Still, it can be seen that cutting words to the same length leads to densities at 0.2 and 0.8 that are higher than at 0.5 (indicating a representation of rhyme), whereas not cutting the words leads to a more skewed distribution of the similarity of those words, overestimating the amount of distance (low similarity) in rhyming words (which is not surprising, since vanilla difflib measures the character overlap of words with potentially very different lengths). It is notable however that the difflib ratios find very few words that are either fully identical or don't share any characters at all. This shows that few word pairs in the dataset are either identical (word identity is not a very creative rhyme), nor fully distinct in characters (there is always overlap).

Examination of Similarity Measures: Next, we will examine German word pair instances from ANTI-K on the basis of our metrics, and how they represent similarity, especially in regards to classifying rhyme. The columns in the following tables are coded as follows: 'w1': First word, 'w2': Second word, 'cossim(w1,w2)': Similarity of the words as determined by the siames network, 'diff(w1,w2)': Difflib character overlap without cutting, 'cut_diff': Difflib overlap with cutting, 'pred.': Prediction of the siamese network whether the words rhyme (1) or not (0), and 'is_rhyme': The gold label whether the two words rhyme. Threshold for classification by metric is set to 0.5, where a number larger than 0.5 indicates rhyme, and vice versa.

Table 6.7 shows word pairs that were correctly classified by the siamese network model. In the first half, we see non-rhyming instances and in the second half, rhyming instances. Most examples in the first half are not very surprising w.r.t. their similarity. When few characters overlap, all metrics show a low similarity. In the first case (stehn, glut), only one character (t) overlaps, and thus similarity is low. Since difflib is agnostic to the position of characters, it measures around .25 overlap, while the siamese network learned that these words don't sound the same (.001). The instances (gra**uen**, q**uell**en) and

(**ei**ns, zw**ei**) stand out for their difflib scores, which are at the threshold of 0.5, even though the overlapping characters are at different positions in the word.

w1	w2	cossim(w1,w2)	diff(w1,w2)	cut_diff(w1,w2)	pred.	is_rhyme
stehn	glut	0.001	0.222	0.250	0	0
laufen	heizung	0.001	0.307	0.333	0	0
reif	wieder	0.003	0.200	0.250	0	0
wohin	komme	0.086	0.200	0.200	0	0
grauen	quellen	0.104	0.461	0.500	0	0
not	blut	0.156	0.285	0.333	0	0
eins	zwei	0.295	0.500	0.500	0	0
gebeten	hungersnöten	0.635	0.526	0.428	1	1
höllenfahrt	gewahrt	0.742	0.555	0.714	1	1
lüge	ziege	0.719	0.444	0.500	1	1
ziege	lüge	0.803	0.444	0.500	1	1
froh	so	0.832	0.333	0.500	1	1
verloren	geboren	0.879	0.666	0.714	1	1
fliege	züge	0.896	0.400	0.500	1	1
waren	gefahren	0.899	0.615	0.800	1	1
hervor	schlagmotor	0.923	0.352	0.333	1	1
beschmiert	erfriert	0.933	0.555	0.500	1	1
kopf	wassertopf	0.982	0.428	0.750	1	1
bist	ist	0.964	0.857	1.000	1	1
ein	sein	0.964	0.857	1.000	1	1

Table 6.7: Examples of Correct Detection of Siamese Network, with similarity metrics.

Regarding rhyming pairs in Table 6.7, such as (ein, sein), (bist, ist), and also (waren, gefahren) or (verloren, geboren) are obvious examples where significant character overlap leads to high similarity on all metrics. But in other examples we can see that the siamese network learned to ignore large portions of a word that are not relevant for rhyming, e.g., (kopf, wassertopf) only matches on 'opf', (beschmiert, erfriert) matches on 'iert', (höllenfahrt, gewahrt) matches on 'ahrt', (hervor, schlagmotor) matches on the final 'or',[6] and (froh, so) does in fact only match on the vowel 'o', where the final 'h' of 'froh' is silent. For many of these cases, cutting the words to same length, especially if one of both words is rather short (kopf, wassertopf), already helps the difflib metric.

[6](hervór, schlagmotór) are both iambic, and thus stressed on the final syllable.

Where the siamese network shines are imperfect rhymes such as (ziege, lüge) or (fliege, züge) where the vowel mapping (/i:/ → /y:/) (character mapping 'ie' → 'ü') is not identical, but close enough to form a rhyme.

w1	w2	cossim(w1,w2)	diff(w1,w2)	cut_diff(w1,w2)	pred.	is_rhyme
blick	stück	0.913	0.400	0.400	1	1
liebt	getrübt	0.973	0.500	0.400	1	1
eingeweide	räude	0.906	0.266	0.400	1	1
liebesblick	zurück	0.909	0.235	0.333	1	1
zieht	heldenlied	0.977	0.266	0.400	1	1
spiel	profil	0.882	0.545	0.400	1	1
schon	nation	0.949	0.363	0.400	1	1
elefant	bestand	0.967	0.428	0.428	1	1
schrei	mai	0.838	0.222	0.333	1	1
augenblick	zurück	0.937	0.375	0.333	1	1
behend	rennt	0.934	0.363	0.400	1	1
los	groß	0.955	0.285	0.333	1	1
syrakus	fuß	0.888	0.200	0.333	1	1

Table 6.8: Imperfect Rhymes, not detected by difflib ratio, but by siamese network.

Table 6.8 lists more examples of imperfect rhymes that are correctly identified by the siamese network, but not by difflib. We see again mappings from long /i:/ and short /i/ to long /y:/ and short /y/ in (blick, stück) and (liebt, getrübt) or long /i:/ to short /i/ in (spiel, profil). However, we also see other mappings from hard consonant plosives /t/ to soft consonant plosives /d/ in (elefant, bestand), (behend, rennt), (zieht, heldenlied) , or from /ai/ to /ɔi/ in (eingeweide, räude). We also find that the model learns phonological mappings that are not identical on the graphematic level, e.g., in (schrei, mai), or the mapping from /s/ to a sharp /ß/ in (los, groß) and (syrakus, fuß).

It is easy to see where the siamese network is better than difflib ratio, and that are cases where there is significant character overlap in non-rhyming words, as shown in Table 6.9. In all of these cases, most characters are present in the second word, but in different order. Thus, the position of characters plays a role in classifying rhyme. However, the siamese network model is by no means perfect, both for false positives and false negatives, as seen in Table 6.10. Unfortunately, it is hard to say where exactly the errors come from, i.e., what the model learned wrong. Most likely, the model didn't see

w1	w2	cossim(w1,w2)	diff(w1,w2)	cut_diff(w1,w2)	pred.	is_rhyme
verschwunden	grunde	0.470	0.555	0.666	0	0
belachen	sehen	0.286	0.615	0.600	0	0
geist	schreiten	0.067	0.428	0.600	0	0
geschaffen	gehn	0.041	0.571	0.500	0	0
wesentlich	entweicht	0.001	0.631	0.666	0	0
edelstein	einem	0.465	0.428	0.600	0	0
hochgeschmissen	sand	0.035	0.211	0.500	0	0
werden	studieren	0.027	0.533	0.666	0	0
stechen	hand	0.175	0.363	0.500	0	0
studieren	erden	0.033	0.571	0.800	0	0
see	rosen	0.044	0.500	0.666	0	0
bei	weib	0.117	0.571	0.666	0	0
klarheit	droht	0.228	0.461	0.600	0	0
bereiten	zerreißt	0.066	0.625	0.625	0	0
begehren	erben	0.143	0.461	0.800	0	0
fliehn	linken	0.011	0.666	0.666	0	0
heide	vinden	0.332	0.545	0.600	0	0

Table 6.9: Non-Rhymes, but detected by difflib ratio, with sim.

some character combinations and/or mappings during training, and thus can't predict certain rhyming words. On the other hand, false positives, i.e., words that do not rhyme but are predicted to do so, are probably matched on the final one or two characters. The matching characters in the second half of Table 6.10 thus are final 'r', 'er', 'en', and 't'.

w1	w2	cossim(w1,w2)	diff(w1,w2)	cut_diff(w1,w2)	pred.	is_rhyme
lenckt	denckt	0.486	0.833	0.833	0	1
briten	unbestritten	0.477	0.666	0.833	0	1
weisen	reisen	0.469	0.833	0.833	0	1
figuren	kreaturen	0.469	0.500	0.571	0	1
katzen	tatzen	0.460	0.833	0.833	0	1
pforte	worte	0.459	0.727	0.800	0	1
müssen	grüßen	0.356	0.500	0.500	0	1
pudel	häuserrudel	0.345	0.500	0.800	0	1
unzerstückelt	entwickelt	0.304	0.608	0.700	0	1
grausen	brausen	0.271	0.857	0.857	0	1
kiese	riese	0.146	0.8	0.800	0	1
schweif	nebelstreif	0.142	0.444	0.571	0	1
schwemmt	firmament	0.075	0.235	0.25	0	1
entferntes	erlerntes	0.023	0.736	0.666	0	1
straßen	hassen	0.011	0.461	0.500	0	1
mehr	wetter	0.807	0.4	0.5	1	0
zieht	tritt	0.783	0.4	0.4	1	0
nieder	tor	0.766	0.222	0.333	1	0
nieder	hervor	0.764	0.333	0.333	1	0
geister	nässer	0.723	0.461	0.5	1	0
geister	gewässer	0.720	0.666	0.571	1	0
fellen	wogen	0.704	0.363	0.4	1	0
quellen	bogen	0.659	0.333	0.4	1	0

Table 6.10: Examples of Incorrect Detection of Siamese Network, with sim.

6.2.7.2 Error of Siamese Networks on Selected Sonetts

We also conduct a small qualitative error analysis on three German sonnets from different literary periods. These include a variety of imperfect rhymes and orthographic deviations. With these poems we can approximate an evaluation in the wild, where each poem brings it's own challenges. The first poem is by Schlegel, the second by Andreas Gryphius, and the third is by Franz Xaver Kappus. The poems with the respective end words and the associated rhyme scheme can be found below. We generate pairs by all permutations (combination, but no duplicates) of end words and evaluate the following models:

1. Model **3token23**:
   This model has 3 BiLSTM layers and it was trained on the original dataset at a ratio 2:3 (positive:negative), still including duplicates in the dataset.
2. Model **3type23**:
   This model has 3 BiLSTM layers and it was trained on the original dataset at a ratio 2:3 (positive:negative), but with duplicates removed.
3. Model **4type11 (4layer11, as used in previous section)**:
   This model has 3 BiLSTM layers and it was trained at a ratio 1:1 (positive:negative), with duplicates removed.
4. Model **4clean11**:
   This model has 4 BiLSTM layers and it was trained at a ratio 1:1 (positive:negative), where the dataset was cleaned of non-obvious rhymes, i.e., when reading the word pair list, we removed word pairs that did not appear to be rhymes, consequently removing rhyme pairs that might have been licensed through the rhyme schema.
5. Model **3indep11**:
   This model has 3 BiLSTM layers and it was trained at a ratio 1:1 (positive:negative), joinly on the three languages German, English and French.

6. Model **4indep11**:

   This model has 4 BiLSTM layers and it was trained at a ratio 1:1 (positive:negative), joinly on the three languages German, English and French.

7. Model **4indephip11**:

   This model has 4 BiLSTM layers and it was trained at a ratio 1:1 (positive:negative), joinly on the three languages German, English and French, plus including Hip-Hop in the training set.

When not mentioned for a poem, a model delivers perfect performance. We only indicate problematic decisions of models, either false positives or false negatives.

Schlegel: Das Sonett 4clean11 and all indep models wrongly detect the mapping 'ei' → 'ä', therefore falsely identifying the following word pairs as rhymes: (Reihen, Gränzen), (Reihen, kränzen), (zweien, gränzen), (zweien, kränzen).

	Schlegel: Das Sonett	
a	Zwei Reime heiß' ich viermal kehren	**wieder**,
b	Und stelle sie, getheilt, in gleiche	**Reihen**,
b	Daß hier und dort zwei eingefaßt von	**zweien**
a	Im Doppelchore schweben auf und	**nieder**.
a	Dann schlingt des Gleichlauts Kette durch zwei	**Glieder**
b	Sich freier wechselnd, jegliches von	**dreien**.
b	In solcher Ordnung, solcher Zahl	**gedeihen**
a	Die zartesten und stolzesten der	**Lieder**.
c	Den werd' ich nie mit meinen Zeilen	**kränzen**,
d	Dem eitle Spielerei mein Wesen	**dünket**,
e	Und Eigensinn die künstlichen	**Gesetze**.
d	Doch, wem in mir geheimer Zauber	**winket**,
c	Dem leih' ich Hoheit, Füll' in engen	**Gränzen**.
e	Und reines Ebenmaß der	**Gegensätze**.

Table 6.11: Schlegel: Das Sonett, with annotated rhyme schema.

Gryphius: Tränen des Vaterlandes The model 3type23 does not detect the spelling variation of the diphtong [aw] (contemporary spelling: [au]) and consequently does not detect the following rhymes: (posawn, carthawn) and (zerhawn/schawn). After manually correcting the words to [au] (e.g., Posaun, Carhaun) the system does detect them. The model 3token23 also does not detect (schawn, zerhawn). Both indep models wrongly identify the following word pairs as rhymes: (blutt, todt), (blutt, noth), (flutt, todt), (flutt, noth), most likely matching on the 't's only.

	Gryphius: Tränen des Vaterlandes / Anno 1636	
a	Wir sind doch nunmehr ganz, ja mehr denn ganz	**verheeret**!
b	Der frechen Völker Schar, die rasende	**Posaun** (Posawn)
b	Das vom Blut fette Schwert, die donnernde	**Carthaun** (Carthawn)
a	Hat aller Schweiß, und Fleiß, und Vorrat	**aufgezehret**.
a	Die Türme stehn in Glut, die Kirch' ist	**umgekehret**.
b	Das Rathaus liegt im Grauß, die Starken sind	**zerhaun** (zerhawn),
b	Die Jungfern sind geschänd't, und wo wir hin nur	**schaun** (schawn)
a	Ist Feuer, Pest, und Tod, der Herz und Geist	**durchfähret**.
c	Hier durch die Schanz' und Stadt rinnt allzeit frisches	**Blut**.
c	Dreimal sind schon sechs Jahr, als unser Ströme	**Flut**
d	Von Leichen fast verstopft, sich langsam fort	**gedrungen**.
e	Doch schweig' ich noch von dem, was ärger als der	**Tod**,
e	Was grimmer denn die Pest, und Glut und	**Hungersnot**,
d	Das auch der Seelen Schatz, so vielen	**abgezwungen**.

Table 6.12: Gryphius: Tränen des Vaterlandes / Anno 1636, with annotated rhyme schema.

Kappus: Sonett The model 3type23 wrongly identifies (weh, Frage), (Blütenschnee, Frage), (Blütenschnee, wage), so it likely matches only on the final 'e's. 4type11 massively overgenerates on this poem, resulting in a precision of .34 (but perfect recall). This questions the overall sanity of this otherwise well performing model, since it learned to indiscriminatively match on the final 'e'. This was not apparent in the error analysis on a large dataset, but it turns problematic if many of the instances have this form. Using this model in production will require post-correction. The 4indephip11 model also wrongly assigns a high cosine similarity (>.9) to word pairs that end with 'e', 'eh' or 'ee', matching only on those 'e' variants: (weh, trübe), (weh, Liebe), (geh, trübe), (geh, Liebe), (Blütenschnee, trübe), (Blütenschnee, Liebe), (See, trübe), (See, Liebe).

	Franz Xaver Kappus: Sonett	
a	Durch mein Leben zittert ohne	**Klage**
b	Ohne Seufzer ein tiefdunkles	**Weh**.
a	Meine Träume reiner	**Blütenschnee**
b	Ist die Weihe meiner stillsten	**Tage**.
a	Öfter aber kreuzt die große	**Frage**
b	Meinen Pfad. Ich werde klein und	**geh**
b	Kalt vorüber wie an einem	**See**
a	Dessen Flu ich nicht zu messen	**wage**.
c	Und dann sinkt ein Lied auf mich, so	**trübe**
d	Wie das Grau glanzarmer	**Sommernächte**,
e	die ein Stern durchflimmert - dann und	**wann**
c	Meine Hände tasten dann nach	**Liebe**,
d	weil ich gerne Laute beten	**möchte**,
e	die main heißer Mund nicht finden	**kann**.

Table 6.13: Franz Xaver Kappus: Sonett, with annotated rhyme schema.

6.2.8 Rhyme Conclusion

In this section, we showed experiments on learning representations of rhyme, both with simple character overlap metrics, but also with siamese recurrent networks which learn a metric for the similarity of rhyme words. We have compiled three new rhyming corpora for German and discussed their annotation with rhyme schema. Siamese Recurrent Networks proved very useful for the detection of rhyme words, as they learn this task with near perfect accuracy across languages. They are especially useful to detect mappings in imperfect rhymes. When switching domains from poetry to Hip-Hop we lose 10 points, and assonances are not really detected. It is notable that these networks can apparently compensate a significant noise level in the training set. We have shown that a SRN can be trained on a dataset containing rhyme pairs of three languages and a model can be adapted to the Hip-Hop domain with little loss in performance. But even though we achieve over 96% Accuracy, each model exhibits individual errors in a qualitative error analysis, making it hard to determine an ideal model. While the independent models work well, they show some problems on particular character mappings, and also monolingual models tend to overgenerate. To work in a production environment, these models have to be checked whether they learn such idiosyncratic mapping, matching only on final characters like ‘e’ or ‘t’, and be corrected in a post-processing step. Overall, is might be advisable to use fewer BiLSTM layers (better 3 than 4) and possibly apply more dropout to prevent these models from being too eager to match single characters.

6.3 Manual Annotation of Prosody in Text

Traditionally, prosody is the study of measurable structures of sound in language. Linguistic prosody is concerned with describing and explaining the structure and function of the suprasegmentals in language. Prosody is the study of the elements of language that contribute toward acoustic and rhythmic effects, chiefly in poetry but also in prose (Gross, 2017). In English or German, the rhythm of a linguistic utterance is basically determined by the sequence of syllable-related accent values (associated with pitch, duration and volume/loudness values) resulting from the 'natural' pronunciation of a line, sentence or text by a competent speaker who takes into account the learned inherent word accents as well as syntax- and discourse-driven accents. Thus, lexical material comes with n-ary degrees of stress, depending on morphological, syntactic, and information structural context. The prominence (or stress) of a syllable is thereby dependent on other syllables in its vicinity, such that a syllable is pronounced relatively louder, higher pitched, or longer than its adjacent syllable.

```
msr iambic.pentameter
met -  + | -   + | -   + |  -  +| -  +  |
rhy 0  1   0   0   0   2 :  0  1  0  2  :
   My love is like to ice, and I to fire:

msr iambic.tetrameter
met  -   + |-   + | -   + | -   +  |
rhy  0   1  0   2   0   0   0   1  :
    The winter evening settles down

msr trochaic.tetrameter
met  +    - | +   - | +  -  | +
rhy  0    0   1   0   1  0    2  :
    Walk the deck my Captain lies,
```

FIGURE 6.4: Examples of rhythmically annotated poetic lines, with meter (`+/-`), feet (`|`), main accents (`2,1,0`), caesuras (`:`), and verse measures (msr). Authors: Edmund Spenser, T.S. Eliot, and Walt Whitman.

In this work, we manually annotate the sequence of syllables for metrical (meter, met) prominence (+/-), including a grouping of recurring metrical patterns with foot boundaries (|). We also annotate a more natural speech rhythm (rhy) by annotating pauses in speech, caesuras (:), that segment the verse into rhythmic groups, and in these groups we assign main accents (2), side accents (1) and null accents (0). In addition, we develop a set of regular expressions that derive the verse measure (msr) of a line from its raw metrical annotation. Figure 6.4 illustrates our annotation layers with three fairly common ways in which poetic lines can be arranged in modern English. A poetic line is also typically called *verse*, from Lat. *versus*, originally meaning to turn a plow at the ends of successive furrows, which, by analogy, suggests lines of writing (Steele, 2012).

We annotate these prosodic features in the two small poetry corpora that were previously collected and annotated for aesthetic emotions by Haider et al. (2020) and discussed in Chapter 3. Both corpora cover a time period from around 1600 to 1930 CE, thus encompassing public domain literature from the modern period. The English corpus contains 64 poems with 1212 lines. The German corpus, after removing poems that do not permit a metrical analysis, contains 153 poems with 3489 lines in total. Both corpora are annotated with some metadata such as the title of a poem and the name and dates of birth and death of its author. The German corpus further contains annotation on the year of publication and literary periods.

6.3.1 Related Work

A digital resource with annotation of poetic meter was missing for New High German. We develop the first resource of this kind here. For Middle High German, Estes and Hench (2016) annotated a metrical scheme for hybrid meter. Anttila et al. (2018) annotated main accents in political speeches. Agirrezabal et al. (2016a, 2019) used the English *for-better-for-verse* and the

dataset of Navarro et al. (2016), who annotated the stress in hendecasyllabic verse (11 syllables) in Spanish Golden Age sonnets. Algee-Hewitt et al. (2014) annotated 1700 lines of English poetry to evaluate their system.

Earlier work (Nenkova et al., 2007) found strong evidence that part-of-speech tags, accent-ratio (the ratio of how often a word form appears stressed vs. unstressed in a corpus) and local context provide good proxy signals for the prediction of word stress (if a whole word is pronounced more or less loud). Subsequently, architectures like MLP (Multi Layer Perceptron) (Agirrezabal et al., 2016a), CRFs (Conditional Random Fields) and LSTMs (Long Short Term Memorys) (Estes and Hench, 2016; Agirrezabal et al., 2019) and transformer models (Talman et al., 2019) have notably improved the performance to predict the prosodic stress of words and syllables from text. However, most of this work only evaluates model accuracy on syllable or word level, with the exception of Agirrezabal et al. (2019), who also evaluates the accuracy of lines.

6.3.2 Annotation Workflow

Prosodic annotation allows for a certain amount of freedom of interpretation and (contextual) ambiguity, where several interpretations can be equally plausible. The eventual quality of annotated data can rest on a multitude of factors, such as the extent of training of annotators, the annotation environment, the choice of categories to annotate, and the personal preference of subjects (Mo et al., 2008; Kakouros et al., 2016).

Three university students of linguistics/literature were involved in our manual annotation process. They annotated by silent reading of the poetry, largely following an intuitive notion of speech rhythm, as was the mode of operation in related work (Estes and Hench, 2016). The annotators additionally incorporated philological knowledge to recognize instances of poetic license, i.e., knowing how the piece is supposed to be read. Especially the annotation accuracy of metrical syllable stress and foot boundaries benefited from recog-

nizing the schematic consistency of repeated verse measures, license through rhyme, or particular stanza forms.

6.3.3 Annotation of Rhythmic Features

In this paper, we incorporate both a linguistic-systematic and a historically-intentional analysis (Mellmann, 2007), aiming at a systematic linguistic description of the prosodic features of poetic texts, but also using labels that are borrowed from historically grown traditions to describe certain forms or patterns (such as verse measure labels).

We evaluate our annotation by calculating Cohen's Kappa between annotators. To capture different granularities of correctness, we calculated agreement on syllable level (accent/stress), between syllables (for foot or caesura), and on full lines (whether the entire line sequence is correct given a certain feature).

6.3.4 Annotation of Main Accents & Caesuras

Caesuras are pauses in speech. While a caesura at the end of a line is the norm (to pause at the line break) there are often natural pauses in the middle of a line. In few cases the line might also run on without a pause. As can be seen in Figure 6.4, punctuation is a good signal for caesuras. Caesuras (csr) are denoted with a colon. We operationalize rhythm by annotating three degrees of syllable stress, where the verse is first segmented into rhythmic groups by annotating caesuras, and in these groups we assign primary accents (2), side accents (1) and null accents (0).

	Syllable		Whole Line	
	m.ac	caesura	m.ac	caesura
DE_{blind}	.84	.92	.59	.89
EN_{blind}	.80	.88	.66	.86

Table 6.14: Cohen Kappa Agreement for Main Accents and Caesura

Six German and ten English poems were annotated by two annotators to calculate the agreement for rhythm. Table 6.14 lists the agreement figures for main accents (m.ac) and caesuras. It shows that caesuras can be fairly reliably detected through silent reading in both languages. On the other hand, agreement on main accents is challenging. Figure 6.5 shows the confusion of main accents for German. While 0s are quite unambiguous, it is not always clear when to set a primary (2) or side accent (1).

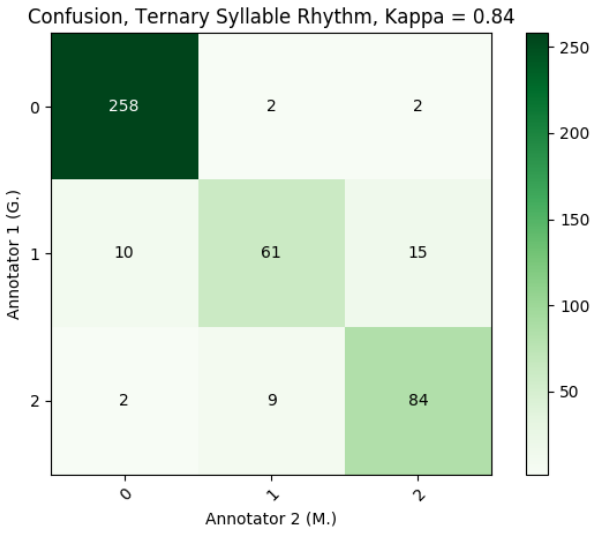

FIGURE 6.5: Confusion of German Main Accents

6.3.5 Annotation of Meter and Foot

In poetry, meter is the basic prosodic structure of a verse. The underlying abstract, and often top-down prescribed, meter consists of a sequence of beat-bearing units (syllables) that are either prominent or non-prominent.Non-prominent beats are attached to prominent ones to build metrical feet (e.g. iambic or trochaic ones). This metrical structure is the scaffold, as it were, for the linguistic rhythm. Annotators first annotated the stress of syllables and in a subsequent step determined groupings of these syllables with foot boundaries, thus a foot is the grouping of metrical syllables. The meter (or

measure) of a verse can be described as a regular sequence of feet, according to a specific sequence of syllable stress values.

	Syllable		Whole Line	
	meter	foot	meter	foot
$DE_{corr.}$	.98	.87	.94	.71
DE_{blind}	.98	.79	.92	.71
EN_{blind}	.94	.95	.87	.88

Table 6.15: Cohen Kappa Agreement for Metrical Stress and Foot Boundaries. Corr. is the agreement of the first version against the corrected version. Blind means that annotators did not see another annotation.

The meter annotation for the German data was first done in a full pass by a graduate student. A second student then started correcting this annotation with frequent discussions with the first author. While on average the agreement scores for all levels of annotation suggested reliable annotation after an initial batch of of 20 German poems, we found that agreement on particular poems was far lower than the average, especially for foot boundaries. Therefore, we corrected the whole set of 153 German poems, and the first author did a final pass. The agreement of this corrected version with the first version is shown in Table 6.15 in the row $DE_{corr.}$. To check whether annotators also agree when not exposed to pre-annotated data, a third annotator and the second annotator each annotated 10 diverse German poems from scratch. This is shown in DE_{blind}. For English, annotators 2 and 3 annotated 6 poems blind and then split the corpus.

Notably, agreement on syllables is acceptable, but feet were a bit problematic, especially for German. To investigate the sources of disagreement, we double annotated and calculated agreement on all 153 poems. See Figures 6.6 and 6.7 for boxplots of these results over all poems, once still including all poems with perfect agreement, and then only showing poems with disagreements to see the variation of only faulty annotations.

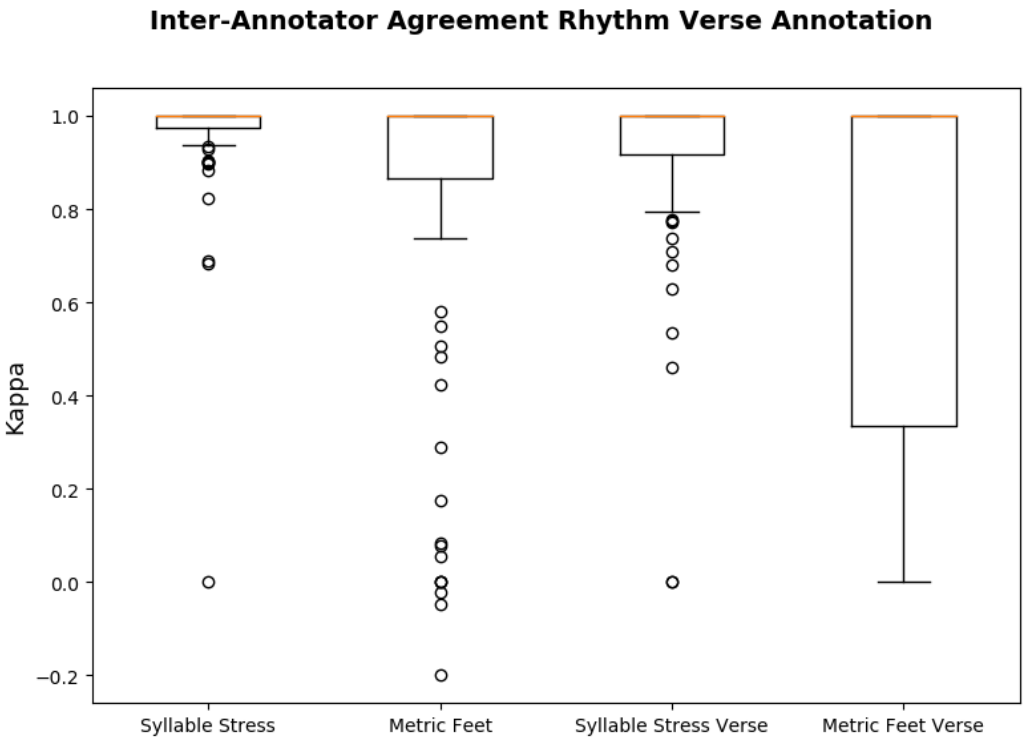

FIGURE 6.6: Meter: Kappas on poems with maximum scores 1.0

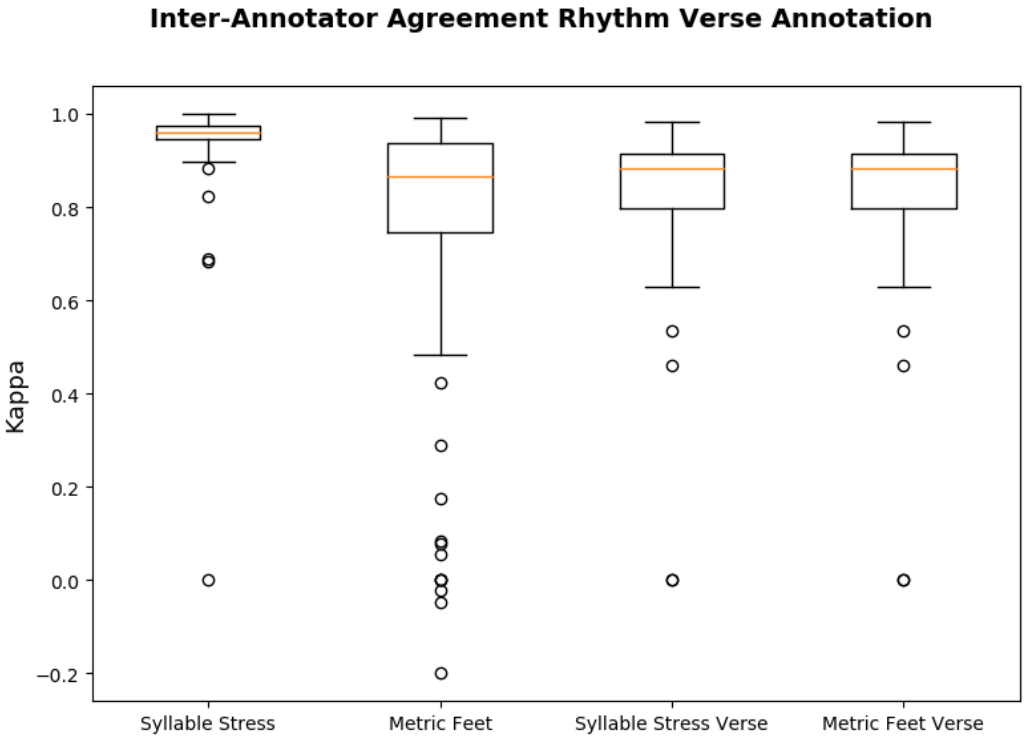

FIGURE 6.7: Meter: Kappas without maximum scores

Close reading for disagreement of foot boundaries revealed that poems with κ around .8 had faulty guideline application (annotation error). 14 poems had an overall $\kappa < .6$, which stemmed from ambiguous rhythmical structure (multiple annotations are acceptable) and/or schema invariance, where a philological eye considers the whole structure of the poem and a naive annotation approach does not render the intended prosody correctly. Particularly noteworthy are the outliers for metric feet ($\kappa < .6$), as listed in Table 6.16. We selected these instances for further inspection, and annotated the source of error in Table 6.16. In the following, we will examine cases of foot ambiguity and then ambiguity in metrical syllable stress.

κ feet	Author	Poem	Year	Error Source
-.20	Hofmannsthal	Regen i.d. Dämmerung	1892	ambiguous and schema
-.05	Trakl	An die Verstummten	1913	error
.02	Schiller	Die Bürgschaft	1798	ambiguous
.06	Eichendorff	Sehnsucht	1834	ambiguous
.08	Heine	Die schlesischen Weber	1844	ambiguous
.08	C. F. Meyer	Zwei Segel	1882	ambiguous
.17	Heym	Halber Schlaf	1912	schema and error
.29	Hoelderlin	Lebenslauf	1826	schema
.42	Stramm	Sturmangriff	1914	error
.48	Heine	Zur Beruhigung	1844	ambiguous
.51	Goethe	Grenzen der Menschh.	1813	schema
.55	Fontane	John Maynard	1886	ambiguous
.56	Rilke	Todeserfahrung	1907	error
.58	Schiller	Der Handschuh	1797	schema

Table 6.16: Poems with the lowest Kappa scores for metric feet, due to (1: schema) inconsistent annotation according to schema constraints, (ii: ambiguous) multiple valid options for feet boundaries and (iii: error) missing annotation of feet boundaries

6.3.5.1 Ambiguous Feet

When comparing the different versions in table 6.16 we found two poems very striking in their foot boundary deviations, (1) Friedrich Schiller's 'Die Bürgschaft' ($\kappa = .02$), and (2) Heinrich Heine's 'Die schlesischen Weber' ($\kappa = .08$). Both poems allow for a flexible positing of feet boundaries, while syllable stress is constant ($\kappa > .98$).

As an example for ambiguous foot boundaries, the following poem, Schiller's 'Bürgschaft' (Figure 6.8), can be set in either *amphibrachic* feet, or as a mixture of *iambic* and *anapaestic* feet. Such conflicting annotations were discussed by Heyse (1827), who finds that in the Greek tradition the *anapaest* is preferable, but a 'weak amphibrachic gait' allows for a freer rhythmic composition. This suggests that Schiller was breaking with tradition.

```
(Foot Boundary Example 1) Schiller, 'Die Bürgschaft'

(1) met="-+-|-+-|-+-|"
    Ich lasse |    den Freund   dir | als Bürgen, |
(2) met="-+|--+|--+|-"
    Ich las   | se den Freund | dir   als Bürg | en,
Transl.: I leave this friend to you as guarantor

(1) met="-+-|-+-|-+-|"
    Ihn magst   du, | entrinn'   ich, | erwürgen. |
(2) met="-+|--+|--+|-"
    Ihn magst | du,   entrinn' | ich,   erwür | gen.
Transl.: Him you may strangle if I escape.

(1) (amphibrach)
(2) (iambus / anapaest)
```

FIGURE 6.8: Foot Boundary Ambiguity: Schiller, *'Die Bürgschaft'*

Furthermore, in the following text by Heinrich Heine (Figure 6.9) we observed a third possibility. By isolating an initial syllable (Auftakt) and aligning word boundaries with foot boundaries one may annotate a mixed trochaic/-dactylic measure.

```
(Foot Boundary Example 2) Heine, 'Die schlesischen Weber'

(1) met="-+|-+-|-+-|-+-|"  Ein   Fluch | dem   Gotte, |     zu   dem   wir | gebeten |
(2) met="-+|-+|--+|--+|-"  Ein   Fluch | dem   Got    | te, zu   dem | wir   gebe | ten
(3) met="-|+-|+--|+--|+-|" Ein | Fluch   dem | Gotte,       zu | dem   wir   ge | beten |
    Translation:           A curse to the god we prayed to

(1) met="-+|-+-|-+|-+-|"   In   Win | terskälte |  und   Hun | gersnöten; |
(2) met="-+|-+|--+|-+|-"   In   Win | terskäl | te und   Hun | gersnö | ten;
(3) met="-|+-|+--|+-|+-|"  In | Winters | kälte    und | Hungers | nöten; |
    Translation:           In winter cold and famine;

(1) (amphibrach)
(2) (iambus / anapaest)
(3) (trochaeus / daktylus)
```

Figure 6.9: Foot Boundary Ambiguity: Heine, *'Die schlesischen Weber'*

6.3.5.2 Ambiguous Syllable Stress

Depending on semantic contrast (alternative semantics), a line in Schillers 'Der Handschuh' (Figure 6.10) invites opposing annotation of syllable prominence. Here, a narrow focus on the article 'den' (highlighting a "specific reward") comes along with stressing the initial syllable (+-) while a wide focus (focus on "Dank" meaning 'any reward') results in the syllable pattern -+.

In Heines satirical poem *'Zur Beruhigung'* (Figure 6.11), we find several examples for syllable stress that deviates from a conventional lexical stress reading. Two examples stand out here. First, as German is largely trochaic, the word '**Ta**bak' would be stressed on the first syllable (+-). But here a stress on the second syllable is licenced through the rhyme with 'Geschmack' (-+), resulting in a comical stress inversion (Ta**bak**). Furthermore, to emphasize the particular country his parody is on (Heine was largely an opponent of the 'young Germany' movement), 'dasjenige' can be pronounced differently.

```
    (Syllable Stress Example 1) Schiller, 'Der Handschuh'

    (Context)
    met="--+|--+|-+|-+|"    Und er wirft ihr den Handschuh ins Gesicht:
    Translation             And he throws the glove in her face

(1) met="+-|+--+|-+|"       »Den Dank, Dame, begehr ich nicht«,
(2) met="-+|+--+|-+|"       »Den Dank, Dame, begehr ich nicht«,
    Translation             »These thanks, lady, I do not desire"«

(1) (narrow focus)
(2) (wide focus)
```

FIGURE 6.10: Syllable Stress Ambiguity: Schiller, *'Der Handschuh'*

```
    (Syllable Stress Example 2) Heine, 'Zur Beruhigung'

(1) met="-+-|-+-|-+-|-+|"  Wir sind keine Römer, wir rauchen Tabak.
(2) met="-+|--+|--+|-+|-"  Wir sind keine Römer, wir rauchen Tabak.
    Translation            We are not Romans, we smoke tobacco

    (Context: Rhyme License)
    met="-+|-+|-+-|-+|"    Ein jedes Volk hat seinen Geschmack,
    Translation            Every nation has its taste

(1) met="-+|-+|-+|--+|"    Benennen wir dasjenige Land,
(2) met="-+|-+|+-|--+|"    Benennen wir dasjenige Land,
    Translation            Let us name that country

(1) (satire; rhyming license)
(2) (lexical stress)
```

FIGURE 6.11: Syllable Stress Ambiguity: Heine, *'Zur Beruhigung'*

6.3.5.3 Conclusion for Error Analysis of Meter and Foot

We showed a method of interpreting conflicting prosodic annotations on verse level with Cohen Kappa agreement scores. Here, we focused on meter for a close reading, but free speech rhythm and especially emotions will allow a far larger picture on multiple valid interpretations of poetry. But already by looking at the disagreement in meter, we found striking examples for different interpretations of the construction of poems that invite philological interpretation with a hermeneutic method.

We have shown that our method is an efficient way to identify disagreement in verse annotation, both for ambiguous interpretations and blatant errors, hence winning insight in the rhythmical structure of poetry. However, this method has its limitation, when both annotators annotate with the same philological lens and agree without considering the possibilities. This method works best to compare 'naive' (e.g., crowd or automatic) annotations with expert annotations.

6.3.6 Annotation of Verse Measures

We develop a set of regular expressions to determine the measure of a line from its raw metrical annotation. We orient ourselves with the handbook of (Knörrich, 1971). The 'verse measure' (msr) is a label for the whole line according to recurring metrical feet. We label the verse according to its dominant foot, i.e., the repetition of patterns like `iambus` (`-+`), `trochee` (`+-`), `dactyl` (`+--`), `anapaest` (`--+`), or `amphibrach` (`-+-`). Also, the rules determine the number of stressed syllables in the line, where *di-, tri-, tetra-, penta-*, and *hexameter* signify 2, 3, 4, 5, and 6 stressed syllables accordingly. Thus, `+-+-+-` is an example for a trochaic.trimeter and `-+-+-+-+` is a iambic.tetrameter, since the foot boundaries should look like this: `-+|-+|-+|-+|`. Typically, female (unstressed) line endings are optional (cadence). Additionally, we annotate labels for (i) `inversion`, when the first foot is inverted, e.g., the first foot in a

iambic line is trochaic: `+--+-+-+`, (ii) `relaxed`, if an unstressed syllable was inserted: `-+-+--+-+` (iambic.tetrameter.relaxed), (iii) and choliambic endings: `-+-+-+--+`. Besides these basic forms, we also implement historically important forms such as a strict `alexandrine`,[7] the `dactylic hexameter`,[8] conventionally known as 'hexameter', and some ode forms like the `asklepiadic` verse (`+-+--++--+-+`).

Table 6.17 lists the most frequent labels for each language without length, called short measure (smsr). The English data includes all datasets that are used in the experiments, as discussed in section 6.4.1. An account of fine grained verse measure labels (full measure, fmsr) in the whole DTA and the rules to derive measures from raw meter can be found in the Appendix. Also, in section 6.5 you'll find an overview of the most frequent fine-grained measures and how their frequency of use changed over time.

English		German	
freq.	smsr	freq.	smsr
2096	iambic	1976	iambic
490	trochaic	793	trochaic
306	anapaest	258	amphibrach
255	amphibrach	206	alexandrine
248	daktylic	76	daktylic
152	hexameter	72	anapaest
91	prosodiakos	26	asklepiade
52	other	17	pherekrateus
35	alexandrine	14	glykoneus

Table 6.17: Most frequent verse measures in small English and German corpora, without length.

[7] Alexandrine: `-+-+-+-+-+-+-?`
The symbol before ? is optional

[8] Hexameter: `+--?+--?+--?+--?+--+-`

6.4 Predicting Prosody with Multi-Task-Learning

In the following, we carry out experiments to learn the previously annotated features and determine their degree of informativeness for each other with a multi-task setup. We include two additional datasets with English meter annotation, and evaluate pre-processing models for syllabification and part-of-speech tagging.

6.4.1 Auxilliary Data and Format

```
# tok   met ft pos  syll csr main smsr    measure    met_line

1  Look   +  .  VB   0    .  1   iambic  i.penta.inv +--+-+-+-+
2  on     -  .  IN   0    .  0   iambic  i.penta.inv +--+-+-+-+
3  my     -  .  PRP$ 0    .  0   iambic  i.penta.inv +--+-+-+-+
4  works  +  :  NNS  0    :  2   iambic  i.penta.inv +--+-+-+-+
5  ye     -  .  PRP$ 0    .  0   iambic  i.penta.inv +--+-+-+-+
6  Might  +  :  NNP  1    .  1   iambic  i.penta.inv +--+-+-+-+
7  y      -  .  NNP  2    :  0   iambic  i.penta.inv +--+-+-+-+
8  and    +  :  CC   0    .  0   iambic  i.penta.inv +--+-+-+-+
9  de     -  .  VB   1    .  0   iambic  i.penta.inv +--+-+-+-+
10 spair'+  :  VB   2    :  1   iambic  i.penta.inv +--+-+-+-+
```

Figure 6.12: Tabular data format for experiments. Author of this line: Percy Blythe Shelley.

Figure 6.12 shows an example line in the data layout that is used for the experiments on prosody, including the 'measure' that was derived with regular expressions from the meter line. 'Syll' is the position of the syllable in a word, 0 for monosyllaba, otherwise index starting at 1. We removed punctuation to properly render line measures, even through punctuation is a good signal for caesuras (see Figure 6.4).

6.4.1.1 English Prosody Datasets

The annotated corpora in English for prosodic annotation include: (1) The for-better-for-verse (FORB) collection[9] with around 1200 lines which was used by Agirrezabal et al. (2016a, 2019), and (2) the 1700 lines of poetry against which `prosodic`[10] (Anttila and Heuser, 2016; Algee-Hewitt et al., 2014) was evaluated (PROS). We merge these with our own (3) 1200 lines in 64 English poems (EPG64). The first two corpora were already annotated for metrical syllable stress. However, FORB does not contain readily available foot boundaries, and in PROS foot boundaries are occasionally set after each syllable.[11] Table 6.18 shows the number of lines in each of our datasets and the number of lines that were incorrectly segmented by our best syllabification systems.

	German	EPG64	FORB	PROS
# correct lines	3431	1098	1084	1564
# faulty lines	58	114	49	173
% faulty lines	1.70	9.41	4.32	10.0

Table 6.18: Size of manually annotated corpora with meter. Faulty lines denotes the number of lines where our automatic syllabification failed. Correct lines are used for experiments, since only there the gold annotation aligns.

6.4.2 Learning Meter

To learn the previously annotated metrical values for each syllable, the task is framed as sequence classification. Syllable tokens are at the input and the respective `met` labels at the output. We test a nominal CRF (see section 3.8.1)

[9] `https://github.com/manexagirrezabal/for_better_for_verse/tree/master/poems`

[10] `https://github.com/quadrismegistus/prosodic`

[11] Additionally, FORB makes use of a `<seg>` tag to indicate syllable boundaries, so we do not derive the position of a syllable in a word. It also contains two competing annotations, `<met>` and `<real>`. The former is the supposedly proper metrical annotation, while the latter corresponds to a more natural rhythm (with a tendency to accept inversions and stress clashes). We only chose `<real>` when `<met>` doesn't match the syllable count (ca. 200 cases), likely deviating from the setup in (Agirrezabal et al., 2016a, 2019).

and a BERT model as baselines and implement a BiLSTM-CRF[12] with pre-trained syllable embeddings. These embeddings were trained by splitting all syllables in the large corpora described in Chapter 3, and training word2vec embeddings over syllables. This system uses three layers of size 100 for the BiLSTM and does the final label prediction with a linear-Chain CRF. Variable dropout of .25 was applied at both input and output. No extra character encodings were used (as these hurt both speed and accuracy).

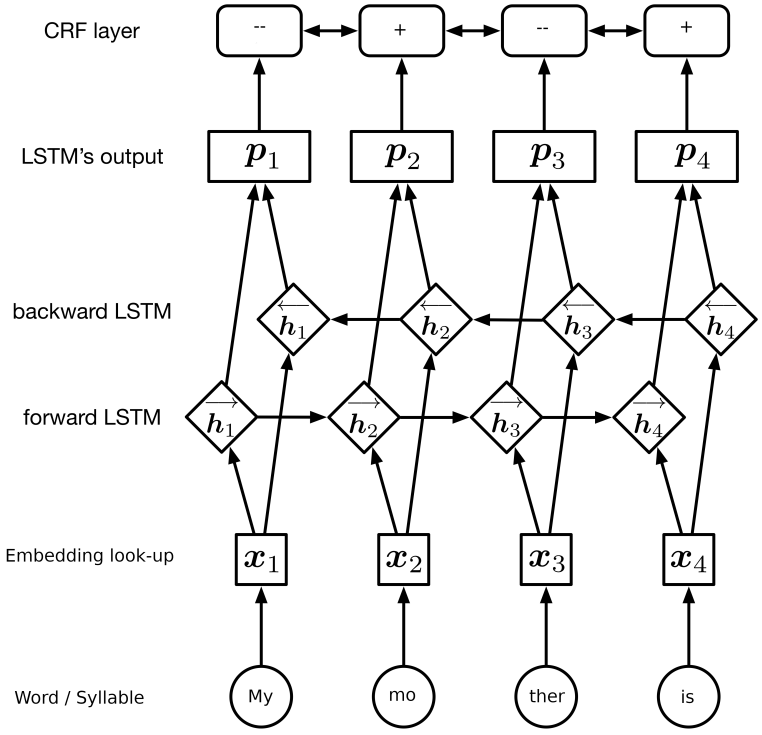

Figure 6.13: BiLSTM Architecture to Learn Meter

See Figure 6.13 for an illustration of the used BiLSTM architecture. We feed syllable tokens at the input and look up their embedding vector, to be propagated through a forward LSTM and a backward LSTM, to then pool the LSTM outputs and feeding these to a final CRF layer that overlays an-

[12] https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf

other structure between the output labels. Here, the labels indicate a iambic sequence, where `+` and `-` are predicted for syllable stress.

We do a three fold cross validation with 80/10/10 splits and average the results, reporting results on the test set in Table 6.19. We evaluate prediction accuracy on syllables and the accuracy of whether the whole line was tagged correctly (line acc.). Line accuracy is especially important if we want to classify poetic verse measures, because only correct line renders allow inference about the poetic form. A system that cannot reliably predict the correct meter of a line hampers large scale automatic annotation.

	English		German	
	syll. acc	line acc	syll. acc	line acc
CRF	.922	.478	.941	.553
BERT	.850	.371	.932	.498
BiLSTM-CRF	**.955**	**.831**	**.968**	**.877**
Agirrezabal (2019)	.930	.614	-	-
Antilla & Heuser (2016)	.894	.607	-	-

Table 6.19: Best Classifiers for Metrical Syllable Stress

Though not directly comparable (data composition differs), we include results as reported by Agirrezabal et al. (2019) for the English for-better-for-verse dataset. We also test the system '`prosodic`' of Anttila and Heuser (2016) against our gold data (EPG64), resulting in .85 accuracy for syllables and .44 for lines. When only evaluating on lines that were syllabified to the correct length (their syllabifier), 27% of lines are lost, but on the correctly syllabified subset the system achieves .89 syllable and .61 line accuracy.

Learning the sequence of metrical syllable stress with BERT cannot compete our other models, possibly resulting from an improper syllable representation, as the word-piece tokenizer segments word chunks other than syllables.

We also experiment with framing the task as document (line) classification, where BERT should learn the verse label (e.g., iambic.pentameter) for a given sequence of words. On the small English dataset, BERT only achieves around

.22 F1-macro and .42 F1-micro. We then tagged 20,000 lines of the large English corpus with a BiLSTM-CRF model and trained BERT on this larger dataset, reaching .48 F1-macro and .62 F1-micro. In this setup, BERT detects frequent classes like iambic.pentameter or trochaic.tetrameter fairly well (.8), but it appears that this model mainly picks up on the length of lines and fails to learn measures other than iambus and trochee like dactyl or anapaest or irregular verse with inversions. This might limit experiments with large scale transfer learning of verse measure knowledge.

6.4.3 Pairwise Joint Prosodic Multi-Task Learning

With the aim of learning the relationships between our different annotation layers, we performed experiments with a multi-task setup. We used the BiLSTM architecture from the previous experiment, where the sequence of syllable embedding vectors is at the input, and the respective sequence of labels at the output. We used the German dataset here, as the annotation is generally more reliable (e.g., POS). In this experiment we also try to learn the annotation of aesthetic emotions that was described for this dataset by Haider et al. (2020) and in Chapter 5. Each line was annotated with one or two emotions from a set of nine emotions. Here, we only used the primary emotion label per line.

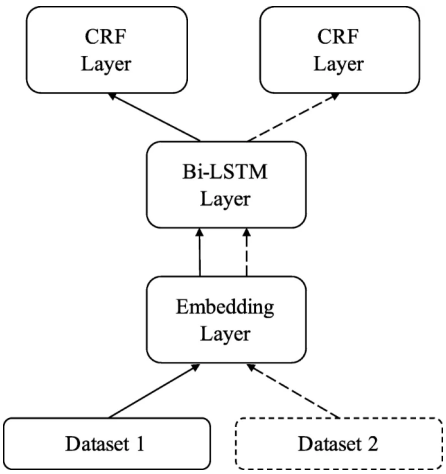

Figure 6.14: Illustration of Multi-Task Setup for Learning Syllable Sequence Tasks.

See Figure 6.14 for an illustration of the Multi-Task setup. We provide a 'Dataset 1' to lean the main task, and a second 'Dataset 2' to learn an auxilliary task. The BiLSTM layers are shared between the tasks, but the final prediction is done on two separate CRF layers.

First, we trained a single task model for each annotation layer, then all tasks jointly (`+all`), and finally pair-wise combinations (`+<auxiliary task>`). In Table 6.20, we report the accuracy on syllable level for each main task with their respective auxiliary tasks.

	met	feet	syllin	pos	csra	m.ac	emo
single	.964	.871	**.952**	.864	.912	.866	.328
+met	-	**.922**	.949	.856	.918	.869	.347
+feet	.961	-	.948	.853	.917	.863	.368
+syllin	.966	.900	-	.860	.919	.867	.330
+pos	.956	.879	**.953**	-	**.924**	**.879**	**.393**
+csra	.961	.886	.940	.855	-	.868	.364
+m.ac	.964	**.915**	.948	.865	.915	-	.354
+smsr	.965	.884	.942	.854	.918	.868	**.378**
+fmsr	**.968**	.899	.938	.858	**.926**	.868	**.395**
+m_line	.966	.882	.937	.853	.919	.868	**.398**
+all	.967	**.930**	.947	.790	.919	.870	.377

Table 6.20: Accuracy for Pairwise Joint Task Learning.

Note that learning syllable-level POS does not benefit from any other task, not even the syllable position in the word, while several tasks like caesuras, main accents and emotions benefit from additional POS information. Predicting meter also degrades from an additional POS task, which possibly interfers with the syllable embeddings. Meter might be also more contextual than suggested in Table 6.22.

However, meter tagging slightly benefits from fine-grained verse measure labels. Interestingly, learning foot boundaries heavily benefits from jointly learning syllable stress. In a single task setup, foot boundaries are learned with .871 accuracy, but in combination with metrical stress, feet are learned with .922 acc. and in combination with main accents at .915. This might

be expected, as foot groupings are dependent on the regularity of repeating metrical syllable stresses (though less dependent on main accents). However, our annotators only achieved Kappa agreement of .87 for feet. It is curious then, how the model overcomes this ambiguity. When learning all tasks jointly (`+all`), foot prediction even reaches .930, suggesting that feet are related to all other prosodic annotations.

We observe that the exchange between caesuras and main accents is negligible. However, caesuras benefit from POS (despite the absence of punctuation), syllable position (syllin) and global measures (msr), indicating that caesuras are integral to poetic rhythm and fairly dependent on syntax.

For emotions we find, despite the hard task (line instead of stanza), and only using syllable embeddings rather than proper word embeddings, that the single task setup is already better than the majority baseline. More importantly, we can see that jointly learning POS or verse measure benefits the emotion prediction (slightly the meter prediction itself: .97). This suggests that there might be a systematic relationship between meter and emotion. This further reinforces the finding in section 5.8, that meter contributes to the emotions that a verse elicits.

6.5 Characterizing Authors and Periods

The previous sections have established that we can learn robust models to predict verse measures from text, allowing us to use them for large scale annotation. In this section, we track the use of said measures over time and across authors. We use the best models for English and German, with around 83–88% Accuracy per correct line render. We derive the measure label from the raw meter annotation of the models with regular expressions as they are shown in the Appendix. On a sidenote: Testing these models against measure labels (e.g., 'iambic.trimeter') and not raw metrical lines (e.g., '`-+-+-+`) actually improves the Accuracy numbers, since the labels are not as fine grained.

Table 6.21 lists the 20 most frequent measure labels in DTA. We find that the majority of lines is set in a strict alexandrine (or iambic.hexameter, since we do not check for caesura positions). There are also many iambic.tetrameters, which indicates the German stanza form 'Volksliedstrophe', which, according to Frank (1980) is among the most frequent stanza forms in German.

Measure	Rel. Freq.	Abs. Freq.
all lines	1	494520
alexandrine.iambic.hexa	0.270	133297
iambic.tetra	0.172	84956
trochaic.tetra	0.097	48060
iambic.tri	0.081	39854
iambic.penta	0.079	39265
iambic.penta.relaxed	0.020	9865
iambic.tetra.relaxed	0.016	7794
iambic.di	0.015	7514
iambic.tri.relaxed	0.015	7365
iambic.hexa.relaxed	0.014	7141
trochaic.di	0.013	6261
trochaic.penta	0.010	5114
trochaic.tri	0.010	4968
trochaic.single	0.009	4696
amphibrach.single	0.008	4143
anapaest.di.plus	0.008	3907
single.up	0.008	3901
amphibrach.di.relaxed	0.007	3453
amphibrach.tri.plus	0.007	3217
trochaic.hexa	0.006	3047
iambic.single	0.006	3006
trochaic.hexa.relaxed	0.006	2784
amphibrach.tetra	0.006	2781
trochaic.septa	0.006	2750

Table 6.21: Most Frequent Verse Measures of Lines by Frequency in DTA Determined with Automatic Annotation. Full Table in Appendix.

6.5.1 Frequency of Verse Measures over Time

The temporal distribution of these verse measures is shown in the Figures 6.15 (DTA-Lyrik: Deutsches Textarchiv). We can see the the alexan-

drine is the dominant verse form in pre-romantic times (before 1750), but it loses importance in later times. Figure 6.16 shows the same for the German Poetry Corpus DLK, though the alexandrine is again present in the time slot 1800–1850. Inspection of the corpus revealed that this sudden renewed interest in this form is only attributable to the five volumes of 'Die Weisheit des Brahmanen' by Friedrich Rückert, which are entirely set in alexandrine verse. Furthermore, both graphs show that iambic.tetrameter, trochaic.tetrameter and iambic.pentameter have enjoyed continuous popularity over the whole time span.

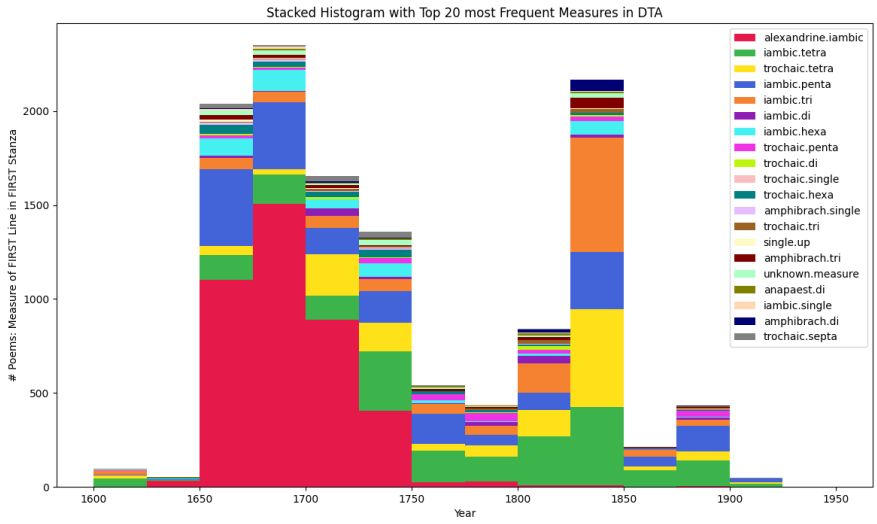

Figure 6.15: DTA Verse Measures over Time 1st Lines in 1st Stanza

Finally, Figure 6.17 shows the distribution of the most frequent verse measures in the English Poetry Gutenberg (EPG) corpus. This plot shows that the most frequent measures in English have been used in relatively equal number across the whole time span, where iambic.tetrameter is the most frequently used, followed by iambic.pentameter, iambic.trimeter and trochaic.tetrameter.

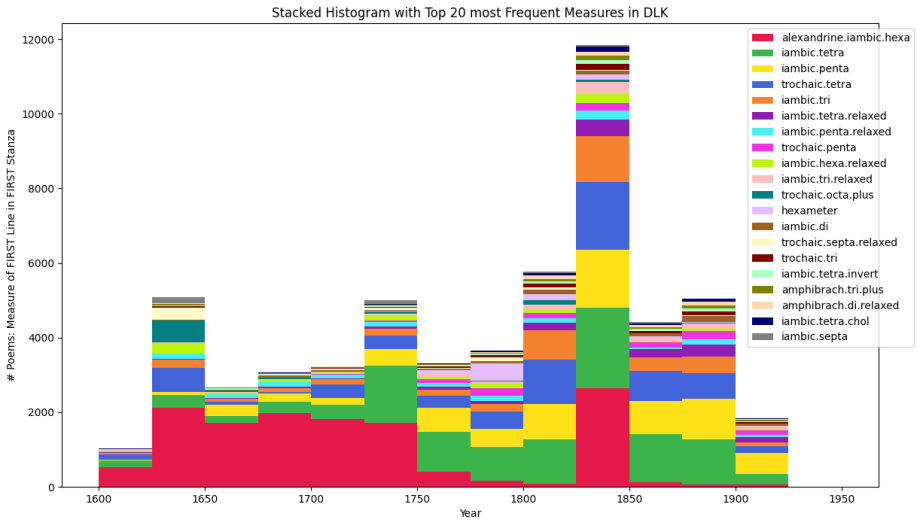

FIGURE 6.16: DLK Verse Measures over Time 1st Lines in 1st Stanza

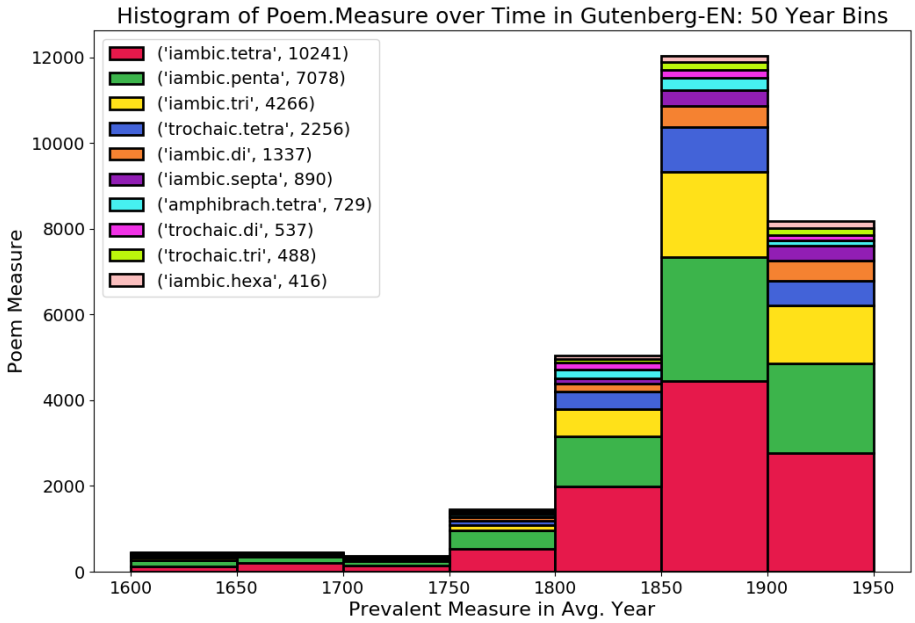

FIGURE 6.17: Histogram of Measures of Poems over Time in Gutenberg-EN

6.5.2 Frequency of Verse Measures over Authors

At the same time, tagging verse measure large scale also allows us to characterize authors by their preferred verse forms. In Figure 6.18 we can see that Klopstock had an interest to write in trochaic verse and also in in the epic verse form of trochaic hexameter. Figure 6.19 shows the verse forms of the translated poems of Charles Baudelaire in the corpus. We find that besides the iambus, the alexandrine, which is a popular form in French poetry, is also present in the translation. However, as seen for example in Figures 6.20 and 6.21 the most popular measures are the iambic and trochaic. More of these plots are found in the Appendix.

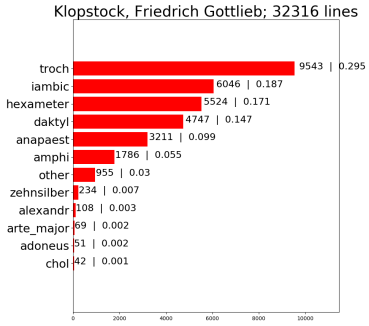

Figure 6.18: Measures of Klopstock: Trochäus, Hexameter

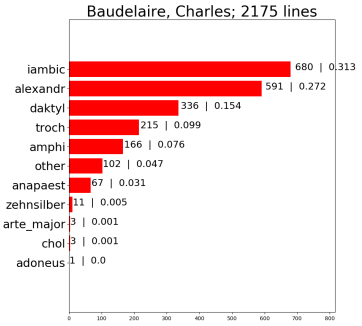

Figure 6.19: Measures of Baudelaire Transl.: Alexandrine

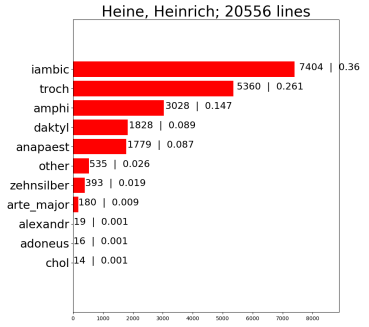

Figure 6.20: Measures of Heine: Iambus, Trochäus

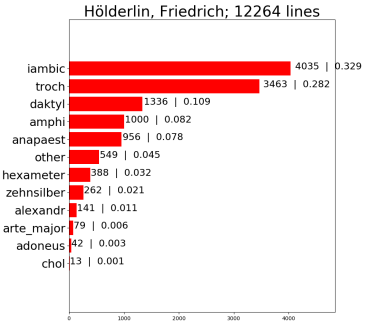

Figure 6.21: Measures of Hölderlin: Iambus, Trochäus

6.5.3 NPMI: Verse Measures in Lit. Periods and over Emotions

In the following, we illustrate which **verse measures and emotions** occur with each other for a line, and also which **verse measures are prevalent in particular literary periods**. To that end, we use the association measure Normalized Pointwise Mutual Information. These calculations are done on the German ANTI-K corpus, and not on the large corpora, since annotation for emotions and time periods is not available in the large corpus, and so far we have not developed methods that would allow a robust large scale annotation. As was shown in section 5.5, NPMI is calculated as follows:

$$PMI(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)} \tag{6.1}$$

$$NPMI(x,y) = \frac{PMI(x,y)}{h(x,y)} \tag{6.2}$$

where

$$h(x,y) = -log_2 p(x,y) \tag{6.3}$$

Pointwise Mutual Information (PMI) is the joint probability of two labels x and y, divided by the individual probabilities of x and y given all labels (here also with log_2). NPMI is PMI divided by the Mutual Information (h). This results in real values in the interval $[-1;+1]$, where $+1$ means that two labels occur exclusively with each other, 0 means that two labels have a chance occurrence with each other (random distribution), and -1 means that two labels never occur with each other. As mentioned before, PMI is a discriminative measure and shows the signal that two labels have for each other. It is not a co-variance method like factor analysis. Therefore, if a label (like Beauty/Joy) occurs with many other labels, but is evenly distributed across those, the PMI score will nevertheless stay low. Only if there is a non-even distribution, PMI will show with which other labels it is more associated and with which less.

In Figure 6.22 we see the NPMI scores between fine grained verse measure labels and aesthetic emotions. Although the plot is very large, we can point out that the trochaic.tetrameter is associated with the emotion Energy/Vitality, possibly attributable to e.g., the satirical verse of Heine. We can also see that Suspense works with certain measures like chol.iambus, iambic.septameter, amphibrach, and anapaest, but it is only distributed randomly across shorter iambic forms, and trochaic forms appear to be not suspenseful. Finally, we can see that the alexandrine elicits emotions of Awe/Sublime and Sadness, which could be also an effect of the Baroque time period (e.g., we can assume that Rückerts alexandrines from the later period are neither sad nor sublime, but mostly humorous). This effect is seen in section 5.5.

Figure 6.23 we see the use of verse measures in literary periods. We can see that e.g., Expressionism is associated with longer iambic and trochaic forms, but not the shorter variants, and of course it is negatively associated with the alexandrine. Expressionism also does not make use of licensed form like choliambus, but we do see some relaxed verse (where an unstressed syllable can be inserted), e.g., in the poems of Heym. Of course we see that the alexandrine has a strong association with the Baroque period. We can also that the 'asklepiade', as part of ode stanzas is found in Sturm und Drang and also the Romantic/Weimar classicism poems. The iambic.pentameter was popular in many periods, but it was avoided in Naturalism, Vormärz poetry, high Romanticism, and also in Barock.

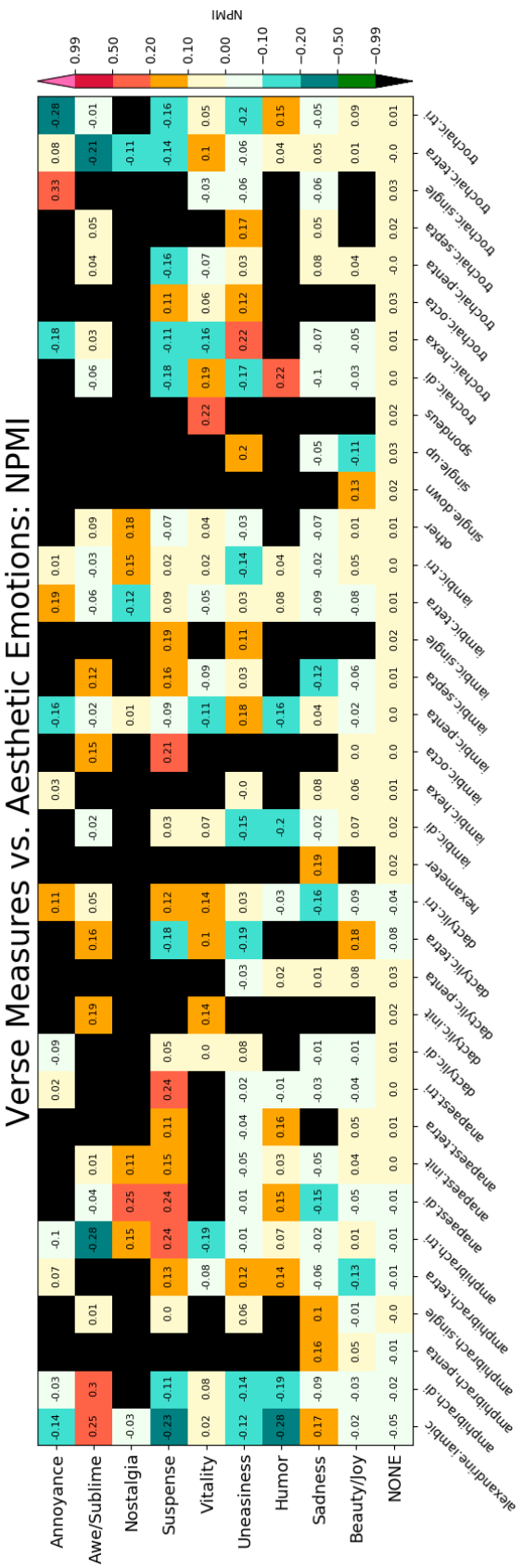

FIGURE 6.22: NPMI Association b/w Verse Measures and Emotion

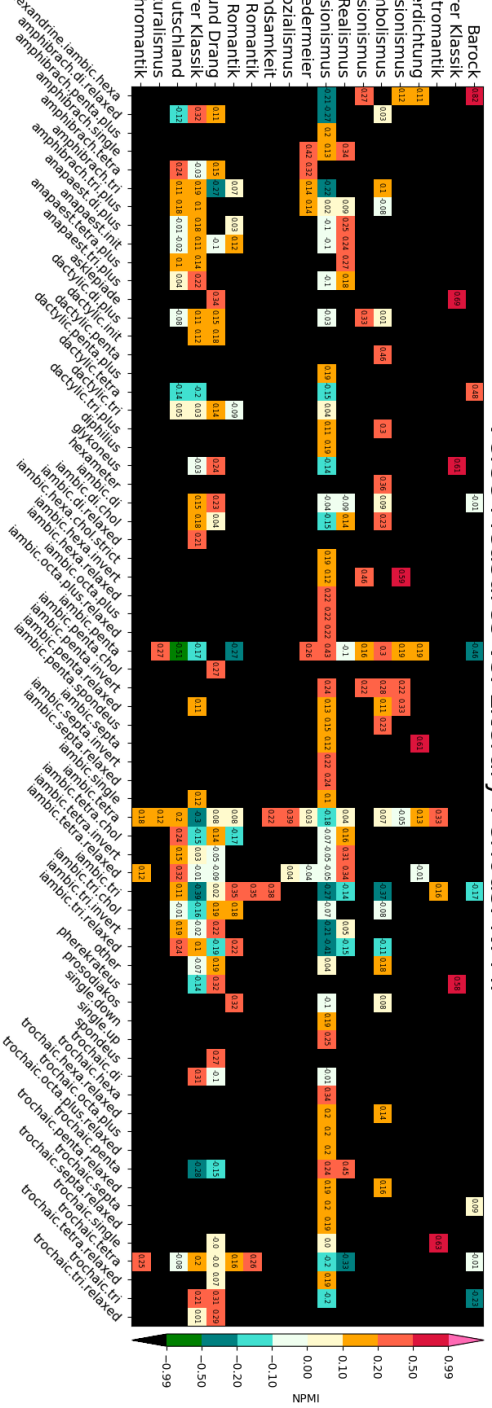

Figure 6.23: NPMI Association b/w Verse Measures and Time Periods

6.6 Prosody and Syntax

The verbal rendering of thought requires the choice of appropriate lexical items and their ordering according to the rules of syntax. Syntax, however, does not fully determine word order: speakers and writers can often choose among possible syntactic constructions when formulating their message. Semantic, pragmatic, as well as phonological constraints are known to affect wording. In spontaneous language production, semantic constraints presumably control sentence structure more immediately and to a stronger degree than phonological constraints. This follows from the logical directionality of language production, in which the semantic content of the message governs lexical choice and the assignment of syntactic function; phonology and rhythm can exert their role and endow the structure with sound only once a syntactic scaffold has been constructed (Levelt, 1993).

In contrast, the diction of poetry can be strongly governed by prosodic regularities, not only for aesthetic effect by endowing a musical quality (Menninghaus et al., 2017, 2018), but also through deliberate deviation from a regular form, consequently highlighting the relevant utterance (Attridge, 2014; Blohm et al., 2018). To enforce a particular prosodic form may thus affect the choice of syntactic constructions and the order of constituents within a sentence, attesting the influence of prosody on syntax in lyric poetry.

6.6.1 Stress Hierarchy

Previous research has noted that part-of-speech annotation provides a good proxy signal to determine the prosodic stress of words (Nenkova et al., 2007; Greene et al., 2010; Navarro-Colorado, 2018a) (nouns are typically stressed, while articles are typically unstressed). To test this, we calculate the pos-accent ratio of monosyllabic words in our German ANTI-K annotation by dividing how often a particular part-of-speech appears stressed (+) in the corpus by how often this part-of-speech occurs in the corpus. For our experiments,

we simplify the tagset. We restrict our analysis to monosyllabic words, as polysyllabic words typically have a lexical stress contour, e.g., German words with two syllables are usually trochaic, where the first syllable is stressed and the second syllable is unstressed (the notable exceptions being alTAR and naTUR). For words with three syllables, we found that nouns are more likely to follow the (+−+) pattern, while verbs prefer (−+−). Furthermore, we only measure the prominence of monosyllabic words, as these display the most ambiguity, as the stress of monosyllabic words is mostly determined by their context. The result is a hierarchy of stress that we report in Table 6.22. At the ends of the spectrum, we see that nouns are usually stressed, while articles are seldom stressed.

POS	Noun	Adj.	Full V.	Adverb	Modal V.	Interj.	Pron.	Prep.	Konj.	Art.
Abr. POS Tag	NN	ADJ	VV	ADV	VM	ITJ	P	AP	KO	AR
Accent Ratio	.97	.89	.84	.75	.73	.55	.4–.015	.27	.23	.06

Table 6.22: Accent ratio for part-of-speech of German monosyllabic words (ratio of metrical stress), from gold data.

To determine the likelihood of a word belonging to a certain POS class being stressed or unstressed, we also iterate over our lines of poetry from DLK, using the models to annotate the corpus for meter and part-of-speech. We then count how often a POS tag falls into a metrically stressed or unstressed syllable.

Anttila et al. (2018) also determined a stress hierarchy, only for sentential stress in political speeches and not for syllables in poetic lines. They are able to establish a stress hierarchy of POS-tags, such that NOUN > ADJ > VERB > FUNC. This shows that functions words (FUNC, e.g. KONJ, ART, APPR, etc) are seldom the main accent of a sentence, while nouns are usually stressed. Based on our corpus, we determined the following hierarchy for monosyllabic word forms:

$$NOUN > VERB\_modal > VERB\_full > ADJ > ADV > FUNC.$$

Word Class	Label	Stress Ratio	Examples
Noun	**'NN'**	16.8	Baum, Welt, Herz, Scherz
Modal Verb	**'VM'**	3.81	muss, kann, soll
Full Verb	**'VV'**	3.67	springt, läuft, sitzt, rennt
Truncated Verb Suppl.	'TR'	3.00	an-, ab-, auf-
Auxilliary	**'VA'**	2.23	hat, ist
Adjective	**'ADJ'**	1.85	rot, hoch, rund, schnell
Named Entity	'NE'	1.44	Hans, Lutz, Max, Liz
Interjection	**'IT'**	1.29	O, Oh, Ach
Particle	**'PT'**	1.28	zu, nicht
Adverb	**'ADV'**	1.14	schon, bald, doch
Cardinal Number	'CA'	1.04	3, 5, 8, 9
Not a word	'XY'	0.85	3:7, H2O, D2XW3
Foreign Material	'FM'	0.79	A, big, fish
Reflexive/Relative Pronoun	'PR'	0.75	sich, dich, mir, der
Pronominal Adverb	'PA'	0.72	drin, dabei
Interrogative Pronoun	'PW'	0.70	wer, was, wo, wann
Indefinite Pronoun	'PI'	0.62	man, kein
Personal / Possessive Pronoun	**'PP'**	0.34	ich, er, ihm, mich, dir, eins
Preposition	'AP'	0.26	im, zur, am
Konjunction	**'KO'**	0.24	um, weil, ob, und, als, wie
Demonstrative Pronoun	'PD'	0.09	dies
Article	**'AR'**	0.06	der, die, das, ein

Table 6.23: POS tag stress hierarchy from automatic CRF prediction. Stress ambiguous monosyllaba. Ratio stressed / unstressed syllables, e.g. 16:1, 4:1, 2:1, .5:1, etc.

This hierarchy reflects the ratio r of stressed to unstressed syllables. The ranking of all POS with their respective r can be seen in Table 6.23. When a POS class is equally likely to be stressed or unstressed, r will be 1.0, or 1:1. For a ratio r = 16.0 (16:1), the word class is 16 times more likely to be stressed. We found it striking that modal verbs are stressed so strongly (3.8:1). We also determined that monosyllabic verbs are more likely in metrically strong positions than monosyllabic adjectives, which differs from what Anttila et al. (2018) found. However, the ends of the hierarchy (nouns and function words) are the same. Plus, their study does not distinguish between adverbs and adjectives and also has only one verb class. We also find that this automatically determined stress hierarchy differs a bit from the one derived from the manual annotation, mainly in that in the manual annotation adjectives are

more stressed than verbs, while the automatic annotation finds verbs more stressed.

6.6.1.1 Contextual Stress Ambiguity

As words are heavily dependent on their context regarding their stress, we look at the immediate left and right context of POS tags, i.e., which POS tag occurs next to it. For brevity, we only show the left context. We retrieve the stress ratio for particular monosyllabic word classes dependent on their context. Context words can be multisyllabic. See Table 6.24 for an overview of nouns, modal verbs, adjectives, adverbs and demonstrative pronouns, and their respective left contexts.

Nouns		Mod. Verbs		Adj.		Adv.		Demonstr. Pron.	
Ratio	Context	Ratio	Context	Ratio	Context	Ratio	Context	Ratio	Context
42.0	KO_NN	30.33	KO_VM	8.01	KO_ADJ	3.78	AR_ADV	2.0	PD_PD
33.7	ADJ_NN	15.85	PD_VM	4.17	PT_ADJ	3.05	KO_ADV	0.75	PA_PD
31.0	AR_NN	10.31	VV_VM	3.44	PW_ADJ	2.19	PD_ADV	0.67	AR_PD
24.6	PD_NN	5.71	AR_VM	3.15	PR_ADJ	1.78	PI_ADV	0.62	KO_PD
21.0	AP_NN	5.54	PP_VM	2.98	PP_ADJ	1.75	PP_ADV	0.5	FM_PD
17.7	PP_NN	4.44	VA_VM	2.82	ADV_ADJ	1.64	PR_ADV	0.42	PI_PD
16.5	PR_NN	4.42	PI_VM	2.35	PI_ADJ	1.54	AP_ADV	0.36	PR_PD
16.0	FM_NN	4.36	PW_VM	2.23	VV_ADJ	1.45	XY_ADV	0.29	PW_PD
14.0	PI_NN	4.28	ADJ_VM	2.15	VA_ADJ	1.43	FM_ADV	0.29	CA_PD
9.91	NN_NN	3.5	FM_VM	1.79	NN_ADJ	1.24	PW_ADV	0.26	PP_PD
9.56	PW_NN	3.33	XY_VM	1.64	FM_ADJ	1.04	ADJ_ADV	0.24	ADJ_PD
8.94	VA_NN	2.96	PA_VM	1.62	PA_ADJ	1.02	ADV_ADV	0.18	IT_PD
8.94	PT_NN	2.58	NE_VM	1.49	PD_ADJ	0.99	NN_ADV	0.17	NN_PD
8.01	ADV_NN	2.21	ADV_VM	1.15	AR_ADJ	0.84	VV_ADV	0.17	ADV_PD
7.03	VV_NN	1.41	NN_VM	1.0	IT_ADJ	0.77	NE_ADV	0.15	VV_PD
6.57	IT_NN	1.28	PT_VM	0.88	VM_ADJ	0.76	PA_ADV	0.15	AP_PD
6.56	NE_NN	1.25	AP_VM	0.75	AP_ADJ	0.73	VM_ADV	0.14	VM_PD
4.12	VM_NN	1.12	PR_VM	0.61	ADJ_ADJ	0.67	VA_ADV	0.13	XY_PD
3.83	XY_NN	1.0	IT_VM	0.36	XY_ADJ	0.6	CA_ADV	0.11	VA_PD
2.83	PA_NN	0.17	VM_VM	0.24	NE_ADJ	0.54	PT_ADV	0.11	NE_PD
1.67	CA_NN			0.12	CA_ADJ	0.41	IT_ADV	0.05	PT_PD

Table 6.24: Context dependence of monosyllabic word stress.

We can see that the hierarchy from Table 6.23 reiterates for contextual dependence. If a word is preceded by a conjunction (KO), then the likelihood of stress is higher. However, nouns never lose their prominence ($r > 1$), regardless of context. Most interestingly, adverbs, which are quite balanced, also show a balanced context dependence, while modal verbs are still mostly stressed, except when they are preceded by another modal verb. We acknowledge that

this table can be problematic, such that some of these contexts seem atypical for particular word classes. Future research should investigate the frequency of particular contexts, and how significant they are. Lastly, the models may also introduce a systematic error that disproportionately affects certain POS classes.

6.6.1.2 Verse Measure vs. Enjambement

Our setup also allows us to get an impression of the interaction of enjambement with verse measures and also POS transitions between lines. Enjambement is an integral part of many poetic lines. It typically signifies incomplete syntax at the end of a line, such that the end of the line encourages a pause in speech, but the sentence, or clause, or phrase, or word is not yet finished. We use a simple way to operationalize enjambement, by assigning enj+ to lines that do not end on a punctuation mark, and enj- to lines that do. Beyond obvious cases (ART_NN does not cross punctuation/clause boundaries), we could not identify clear preferences of enjambement for particular POS transistions. We also conducted research to annotate a more fine grained version of enjambement, as it is sketched in the glossary in the appendix, but the experiments are not yet finished.

We use our verse measure prediction on DLK (here, we use I for stressed and o for unstressed syllables, while the symbol before '?' is optional and '$' is the end of the line). Unsurprisingly, we find that lines with fewer stressed syllables prefer enjambement more. However, for measures with six stressed syllables, the 'running measure' hexameter (Ioo?Ioo?Ioo?Ioo?IooIo$) is more sympathetic with enjambement than any other measure with a probability of p(enj+) = .41, while the alexandrine (oIoIoIoIoIoIo?$) dislikes it, ranking as most unlikely with p(enj+) = .16, compared to all other measures.

6.7 Concluding Remarks

We created large poetry corpora for English and German to support computational literary studies and annotated prosodic features in smaller corpora. Furthermore, we annotated a number of features (part-of-speech, syllable boundaries, and meter) on the large corpora for large scale analysis.

Our evaluation shows that a multitude of prosodic features can be reliably annotated through silent reading, including meter, main accents and caesuras. Still, foot annotation can be challenging. We examined cases of prosodic ambiguity in poetry annotation, working out different sources of error through measuring agreement and inspecting the data.

Finally, we performed first experiments with a multi-task setup to find beneficial relations between certain prosodic tasks. For example, learning metrical annotation, including feet and caesuras, largely benefits from a global verse measure label, while foot boundaries also benefit from any joint learning with syllable stress and all features altogether, even surpassing the human upper bound.

We have presented preliminary results on the informativity of verse measure for elicited emotions and the prevalence of verse forms in literary periods, and also showed large scale analysis of the temporal distribution of verse forms and showed the viability of author profiles with regard to their preferred verse forms.

Finally, we have also shown experiments on the intersection of syntax and speech rhythm, outlining stress hierarchies with and without context, and how historically grown verse forms interact with enjambement.

Final Conclusion 7

This thesis presents a comprehensive investigation into the use of computational stylistics for the analysis of poetry with a focus on German and English texts, but also incorporating poetry from other languages where appropriate. The included studies adopt a distant reading approach, which rests on the examination of large language corpora and facilitates the identification of patterns and trends regarding literary reading and literary history, encompassing a range of methodologies and techniques, including manual annotation workflows, descriptive and predictive data modeling, and the use of extensively curated poetry corpora. These approaches are informed by ideas from distant reading and stylistics, which allow for exploration and discovery while also incorporating insights and hypotheses derived from literary scholarship. Overall, this thesis aims to provide a better understanding of poetry, its evolution, the emotions it elicits, the way prosody works, and the relationships between its prosody, aesthetics, and historical context.

A major contribution of the thesis is the creation of large poetry corpora for German and English, and augmenting them with reliable automatic annotation, such as part-of-speech tags, syllabification, metrical syllable stress, and verse measures. On the other hand, smaller corpora build the backbone of

the research, providing insight on how poetic features can be annotated (hopefully helping future research in literary scholarhip), and providing the basis to build models. Models allow us to look at the inter-relatedness of (poetic) devices and features, and they are used here in a multitude of ways: Modeling the co-variance of an assortment of lexicon-based text features enables us to study the boundaries and characterization of literary genres. Unsupervised methods from distributional semantics allow us to get insight on the themes that are associated with literary periods and tracking the emergence of poetic tropes. A model based on contextualized embeddings allows us learn linguistic representations of aesthetic emotions and a transfer learning setup allows us to see that there is a systematic relationship between aesthetic emotions and meter. A recurrent neural network based on syllable embeddings in a multi-task learning setup allows us to improve the state-of-the-art in metrical tagging and gauge the inter-relatedness of prosodic devices in poetry.

The use of Latent Dirichlet Allocation for a visualization of topic trends in a mono-lingual and cross-lingual setting has allowed us to illustrate the similarities and disparities between different poetic traditions and literary periods. The method used is largely based on reading and translating topic distributions and interpreting the trajectories of relative topic importance against the backdrop of literary history. However, the method is not without flaws, as over- or underrepresentation of certain authors or the bias towards certain literary periods (like romanticism) can lead to corpus imbalance, impacting the measure to calculate the relative importance of a topic given a certain time stamp. This should be addressed in future work.

A similarity analysis of words in diachronic word embeddings has allowed us to reconstruct literature period transitions and extract emerging and vanishing poetic tropes based on the co-variance of time trajectories of the semantic distance in word pairs. Although the dataset used was large, it was still somewhat sparse in the distribution of words over all time slots, partially because many word forms simply emerge or vanish at a certain point. This

calls for more data and a more robust model for confident analysis of the laws of semantic change or to get a broader view of poetic metaphors and motifs.

Furthermore, we have shown that and how it is possible to annotate aesthetic emotions in small corpora and developed models that can identify these emotions from text, also in a transfer learning setup. The annotated emotional and aesthetic reader response to reading poetry has provided insight into the experience of reading poetry. The research has shown that basic emotion terms are of limited use for understanding poetry reception and that a closely supervised annotation task results in substantial agreement on the final dataset. However, the task of collecting reader-perceived emotion response to poetry in a crowdsourcing setting is not straightforward and requires an improved training strategy and a larger number of crowdworkers to achieve better results. Modeling experiments showed that more data would improve the detection of our emotion categories from text, but also small data already allows us to learn the more frequent emotion fairly reliable (like Beauty/Joy and Sadness), and that there is a relationship between emotions and meter.

We have also shown the viability of annotating prosodic features in poetic text and how to reliably detect them with computational models that, in a multi-task setup, also show the inter-dependence of these features. We built a balanced rhyming corpus for German and used machine learning techniques to study the similarity of rhymes. We have used dis-agreements as departure to further closely investigate the texts in question, encountered e.g., in the annotation of foot boundaries, where closer examination revealed that there are ambiguous patterns at work. Moreover, the examination of prosody and its relationship with syntax and emotional impact has offered insights into the technical aspects of poetry, and highlighting the importance of rhythm and meter in shaping the emotional tone and aesthetic appeal of the genre. We have also shown stylistic variation with information theoretic measures and counting, for example the variation of emotions and verse measures across literary periods.

Overall, the developed methods on the one hand showed some promise and future directions for this research became apparent, but we could also identify shortcomings. First, we will discuss the shortcomings, by chapter, and then directions for future research. Overall, this research provides a foundation for further studies on the use of computational tools for analyzing the variation of stylistic features and changes in language use in literature. Further research should investigate the robustness of these methods across other datasets and the impact of corpus imbalance on the results. For example, it could explore how tropes and metaphors form in other text genres and the emotional and aesthetic reader response to reading poetry in larger and more diverse datasets.

7.1 Shortcomings of this Research

7.1.1 Representativeness of Corpora

While we have compiled a large corpus of New High German poetry, it is restricted by several factors: We are dependent on the current availability of high-quality digitized corpora under free license. For example, we excluded the German Project Gutenberg collection, even after investing considerable work, simply because (a) the quality of annotation (e.g., segmentation of stanzas etc.) and the metadata are problematic, and (b) the project is under private curation, hampering the free distribution of processed data. Yet, as more poetry will be digitized, future researchers will be able to get a more complete picture of the history of New High German poetry. For other languages, like Spanish, it would be desirable if more poetry would be digitized, to get a more complete picture beyond e.g., sonnets and the Siglo de Oro, or for English to have more poetry before Romanticism. Still, much will depend on the quality of future resources. One might argue that the quality of the data is actually more important than quantity.

German law on when literature becomes public domain is fairly restrictive, as we can only work with texts where the author has been deceased for 70 years. This excludes most texts from the 20th century. At the same time, there is the corpus of `lyrikline.org`, which includes contemporary poetry, but it is still comparatively small (a few thousand poems), and not standardized (e.g., in XML), and problematic w.r.t. licensing. But first research on some of these poems was conducted by Baumann et al. (2018).

The corpora are quite underrepresented in pre-romantic times, and the annotation (e.g., of publication date) in that era is still a bit inconsistent (where we often took the average year over the life-span of the author). It is hard to tell how complete the corpora are, especially w.r.t. authors, whether all of, most of, or just parts of the oeuvre of an author is represented in the corpus. Research that might want to investigate the evolution of the language of single authors will still encounter obstacles in this corpus. Furthermore, it needs to be determined what actually constitutes a representative corpus of German poetry, what constitutes the canon, and whether and to what extent 'non-canon' is desirable.

7.1.2 Diachronic Variation

The topic models and diachronic word embeddings revealed interesting patterns, despite their apparent simplicity. However, simply calculating the average probability of a topic per time slot is (a) heavily dependent on the quality and the representativeness of the underlying corpus, and (b) relies upon finding relevant topics with a lateral topic model (that was trained on the whole corpus, rather than looking at individual time slots). Furthermore, the division of time into fixed time slots obscures the continuous nature of time. Also, the topics and tropes shown were selected based on manual inspection, with the exception of topics that were relevant for temporal classification (identified with information gain of the classifier). A similar problem arose with translat-

ing topics across languages. We relied on a manual translation and mapping across the languages, presenting topics that we deemed relevant for literary history, rather than using automatic methods for translation. Future work should look into cross-lingual alignment methods, e.g., through multi-lingual embeddings or poly-lingual topic models without parallel data.

The diachronic embedding method aligned embeddings by first computing a MAIN embedding and then concatenating temporal representations. This is helpful to deal with aligning embeddings over time, but it is problematic when words are not available in certain time slots. In these cases, the model reverts to the MAIN embedding, and thus only represents the embedding of the word over the whole corpus, rather than the embedding of the word at a certain time slot (of which there would be none if the word is not available).

7.1.3 Aesthetic Emotions

How exactly to measure emotional impact from texts or other stimuli is still an active research area (Fayn et al., 2021; Schindler et al., 2017b). In this work we have shown that methods from corpus linguistics are suitable to label poetry with the emotions it elicits, providing an alternative to psychometrics. Still, our research did not look into personal differences of people, but focused on conceptualizing a fitting label set that is suitably discriminative and able to explain the emotional states that can be experienced when reading poetry. The end result was corpora in different languages that are annotated with these aesthetic emotions in a fine grained manner. Since publication of the paper we also annotated further languages, that should allow research of aesthetic experience over different languages. Yet, we cannot claim that our labelset is exhaustive. There might be other experiences of emotions (like melancholia) or other judgements that we did not include, but that might be of interest for specific research questions. Furthermore, we did not measure any emotional intensities (like different degrees of Sadness), and we focused on designing

a discrete labelset, annotating only whether the emotion is present, and not measuring on a scale, as is typically done with dimensions like valence, arousal or dominance.

We also found that using crowd workers for this annotation is not straightforward, and does not lead to sensible results, since the data is fairly inconsistent, where the crowd workers did not really agree on more subtle emotions. This reveals that training annotators (as was the case with our experts) is quite necessary, to have a shared conceptualization of the used emotion terms. Still, we saw that there is something like a 'crowd wisdom', where aggregating the annotations of more crowd workers leads to more agreement with the experts. A future crowd working experiment for this task should however focus on improving the training of the crowdworkers, such that they do actually have a better concept of the emotion categories.

Finally, the experiments on modeling revealed that a transformer model based on BERT does in fact learn something about emotions from text. Also, that transfer learning is useful, especially in a low resource setting. However, the model does need some larger context like a stanza, as it is difficult to classify emotions only on the basis of independent lines. But the human annotators also had access to the whole poem (or stanza in the case of the crowd workers), and had trouble with identifying an emotion when they only read an independent line. Still, agreement of the experts was consistently good across all emotions, while the computational model struggled with not-so-frequent emotions. A hierarchical model might be able to better model the emotion arcs that we found in our poetry. Also, we found that information on the meter (verse measure) of a line is helpful for a computational model to detect the elicited emotions. A model does not need to 'read' as much poetry if it has knowledge about the meter of the poetry, to achieve similar performance in simulating the emotional effect of it. However, our results so far can only show some effects in a low resource setting. When training on more metrical annotation or on more emotion annotation does not substantially increase the

detection accuracy for emotions. Achieving .6 F1-macro for emotion detection is still out of reach with the methods we presented here (where the human upper bound is in the high 70s).

7.1.4 Modeling Prosody

We have shown that it is quite possible to annotate a number of prosodic features in poetry through silent reading, resulting in good agreement scores across the board. However, we also identified some problems with ambiguous prosody (especially w.r.t. foot boundaries and main accents), that might deserve more in-depth close reading than was done here, especially for border cases and poetic license.

Besides rhythmic features, we have also shown that annotating rhyme is feasible. Still, it is of course problematic for contemporary annotators to detect all possible historic and dialectal variations. This hinders this dataset to be used for the reconstruction of pronunciation (List et al., 2017), just as much as dealing with literary data for linguistic questions can be problematic due to poetic license and the wide stylistic variation. The annotation of rhyme as schemes can be problematic in sensitive settings, because mistakes in the annotation amplify with broken indices, leading to wrong mappings between words.

We have presented high performing models for the detection of rhyme and for prosody, and we have shown that learning such models from text can result in reliable models that do not need any other modalities like audio (which brings its own set of problems). Our rhyming models learned apt representations of phonological similarities between words only from characters, and a representation of the cosine distance/similarity of those words allowed us to estimate the 'imperfectness' of rhyme. However, training neural networks for this task brings the problem that it is hard to interpret what exactly such a model has learned. We found that some models learned a mapping based only

on isolated characters (like ‘t’ or ’e’), and not always an adequate representation of rhyming itself. And it is not obvious during training when and why a model will learn an improper representation, and it is not trivial to find the source of error, besides testing the model post-hoc.

Multi-task-learning (MTL) improved the results for meter (and especially foot) detection, bringing the correct labeling of verse measures (for example ‘iambic.pentameter’) per line from around 60% into the mid-to-high 80s. This allows us for the first time to use such models to accurately label the verse measure in large corpora. Furthermore, MTL helped us better understand the interdependence of prosodic features in poetry. However, so far our models are based on BiLSTM architectures, and we found that transformer models (Devlin et al., 2019), despite their wide success, don’t work out of the box for such specific linguistic units such as syllables. To further work on models that illustrate stylistic variation and especially the relation of rhythmic features and emotions, better methods to represent non-standard linguistic units (like syllables) need to be found.

7.2 Directions for Future Research

Overall, this thesis spent considerable time on fundamental research to conceptualize issues of stylistic variation and to build predictive models for large scale poetry research. We identified a number of problems that might look trivial at first, but that pose major problems when evaluated in a rigorous manner. Examples of this include accurate part-of-speech tagging or the prediction of metrical stress. We hope this research will raise some awareness that off-the-shelf tools might not be immediately adequate for such specific domains such as poetry.

We have created a host of resources that hopefully will be helpful for future research, we have shown first experiments that can be carried out with these resources, and we have investigated the variation of style features in

poetry w.r.t. their inter-dependence among each other, and also for genre, emotions, and time. The methods to do this were varied and offered different degrees of insight and confidence in the results. Especially the results based on NPMI are limited in their credibility without looking into more detail of the significance of certain numbers. However, to get a first feeling of stylistic variation, they were quite insightful. And despite the interesting results in stylistic variation that was revealed through transfer learning and multi-task-learning, these models have remained restricted to low resources scenarios. Future research should definitely look into causal models and decision boundaries to adequately find out which stylistic or linguistic modality is responsible to elicit certain emotions or aesthetic judgements on particular instances. Also, detecting the more frequent stanza forms such as sonetts, 'Volksliedstrophe' and odes should be straightforward on our large annotated corpora.

Appendix 8

8.1 A Poem in TEI P5: Annotation Glossary

Here, we will illustrate annotation layers that occur in our corpora. The annotation layers are coded in inline TEI P5 XML, and can be understood as suggestion for future archivars. As example, we will use the poem 'The Mystery of Pain' by Emily Dickinson (1830–86) that was published posthumously in 1924.

THE MYSTERY OF PAIN.

Pain has an element of blank;
It cannot recollect
When it began, or if there were
A day when it was not.

It has no future but itself,
Its infinite realms contain
Its past, enlightened to perceive
New periods of pain.

from: Emily Dickinson (1830–86), published posthum. 1924

Lines and Title

The most basic unit of a poem is the line with the tag `<l>`. A poetic line is also typically called *verse*, from Lat. *versus*, originally meaning to turn a plow at the ends of successive furrows, which, by analogy, suggests lines of writing (Steele, 2012). The title of a poem is annotated with a `<head>` tag.

```
<head>THE MYSTERY OF PAIN.</head>

    <l>Pain has an element of blank;</l>
    <l>It cannot recollect</l>
    <l>When it began, or if there were</l>
    <l>A day when it was not.</l>


    <l>It has no future but itself,</l>
    <l>Its infinite realms contain</l>
    <l>Its past, enlightened to perceive/l>
    <l>New periods of pain.</l>
```

Line Groups

Lines typically build groups to form line groups `<lg>`. These line group tags typically feature an attribute `type`, that either designates a stanza (`type="stanza"`) or a whole poem (`type="poem"`).

```
<lg type="poem">
    <head>THE MYSTERY OF PAIN.</head>
    <lg type="stanza">
        <l>Pain has an element of blank;</l>
        <l>It cannot recollect</l>
        <l>When it began, or if there were</l>
        <l>A day when it was not.</l>
    </lg>
    <lg type="stanza">
        <l>It has no future but itself,</l>
        <l>Its infinite realms contain</l>
        <l>Its past, enlightened to perceive/l>
        <l>New periods of pain.</l>
    </lg>
</lg>
```

Rhyme

The most basic definition of rhyme is 'the repetition of identically sounding word segments from the last accented vowel to the end of the word'. Our rhyming corpora are discussed in Section 3.6 and the modeling of rhyme is discussed in section 6.2.1. In this thesis, we we annotate rhyme through end-rhyme schemas in the stanza. Note that this poem only has the pair (contain, pain) in end-rhyme position, thus the 4-line stanza is annotated with the scheme `abcb`, where the `b` lines rhyme. Yet, annotating schemes as indices has the drawback that mistakes in the annotation can lead to broken mappings. A line based or token based annotation will be more robust.

```
<lg type="poem">
    <head>THE MYSTERY OF PAIN.</head>
    <lg type="stanza" rhyme="abcd">
        <l>Pain has an element of blank;</l>
        <l>It cannot recollect</l>
        <l>When it began, or if there were</l>
        <l>A day when it was not.</l>
    </lg>
    <lg type="stanza" rhyme="abcb">
        <l>It has no future but itself,</l>
        <l>Its infinite realms contain</l>
        <l>Its past, enlightened to perceive/l>
        <l>New periods of pain.</l>
    </lg>
</lg>
```

Meter

In poetry, meter is the basic prosodic structure of a verse. The underlying abstract, and often top-down prescribed, meter consists of a sequence of beat-bearing units (syllables) that are either prominent or non-prominent. Non-prominent beats are attached to prominent ones to build metrical feet (e.g. iambic or trochaic ones). This metrical structure is the scaffold, as it were, for the linguistic rhythm. Meter is discussed in Chapter 6.

After a annotation of metrical syllables, we determine groupings of these syllables with foot boundaries, thus a foot is the grouping of metrical syllables. The meter (or measure) of a verse can be described as a regular sequence of feet, according to a specific sequence of syllable stress values. Foot boundaries are denoted with the pipe symbol `|`.

```
<lg type="poem">
    <head>THE MYSTERY OF PAIN .</head>
    <lg type="stanza" rhyme="abcd">
        <l met="-+|-+|-+|-+|">
        Pain has an element of blank ;</l>
        <l met="-+|-+|-+|">
        It cannot recollect</l>
        <l met="-+|-+|-+|-+|">
        When it began , or if there were</l>
        <l met="-+|-+|-+|">
        A day when it was not .</l>
    </lg>
    <lg type="stanza" rhyme="abcb">
    [...]
    </lg>
</lg>
```

Verse Measure

As discussed in Chapter 6, we developed a set of regular expressions to derive the so-called 'verse measure' label of a line from its raw metrical annotation. We make the distinction between the terms 'measure' and 'meter' here, where 'meter' denotes the raw sequence of syllable stress values (like -+-+-+-), whereas 'measure' refers to a label for a line that denotes the most likely foot pattern ('iambic') and the number of these feet ('tri'), plus certain deviations from the norm like 'inversion' or 'choliambus'. We orient ourselves with the handbook of (Knörrich, 1971). Note that the first line contains a so-called 'inversion', or 'foot inversion' where 'Pain' is stressed harder than 'has', thus breaking the regular meter of the line so that the first iambic foot is inverted to a trochaic foot.

```
<lg type="poem">
    <head>THE MYSTERY OF PAIN .</head>
    <lg type="stanza" rhyme="abcd">
        <l met="+--+|-+|-+|"   measure="iambic.tetra.invert">
        Pain has an element of blank ;</l>
        <l met="-+|-+|-+|"     measure="iambic.tri">
        It cannot recollect</l>
        <l met="-+|-+|-+|-+|"  measure="iambic.tetra">
        When it began , or if there were</l>
        <l met="-+|-+|-+|"     measure="iambic.tri">
        A day when it was not .</l>
    </lg>
    <lg type="stanza" rhyme="abcb">
    [...]
    </lg>
</lg>
```

Rhythm

The rhythm of a linguistic utterance is determined by the sequence of syllable-related accent values (associated with pitch, duration and volume/loudness values) resulting from the 'natural' pronunciation of a line.

Caesura are pauses in speech. While a caesura at the end of a line is the norm (to pause at the line break) there are often natural pauses in the middle of a line. In few cases the line might also run on without a pause. As can be seen in Figure 6.4, punctuation is a good signal for caesuras. Caesuras (csr) are denoted with a colon. We operationalize a more free system of rhythm (as opposed to binary metrical syllable stress) by annotating three degrees of syllable stress, where the verse is first segmented into rhythmic groups by annotating caesuras, and in these groups we assign primary accents (2), side accents (1) and null accents (0).

```
<lg type="poem">
    <head>THE MYSTERY OF PAIN .</head>
    <lg type="stanza" rhyme="abcd">
        <l met="+--+|-+|-+|"   rhythm="20010002:">
        Pain has an element of blank ;</l>
        <l met="-+|-+|-+|"     rhythm="000001">
        It cannot recollect</l>
        <l met="-+|-+|-+|-+|"  rhythm="0002:0001">
        When it began , or if there were</l>
        <l met="-+|-+|-+|"     rhythm="02:0002:">
        A day when it was not .</l>
    </lg>
    <lg type="stanza" rhyme="abcb">
    [...]
    </lg>
</lg>
```

Enjambement

Enjambement describes the interaction of a line break with syntactic or semantic units of the sentence or poem. We use 5 levels of granularity. It should be noted that enjambement is not included in this thesis.

`Level 1`: Denotes a syllable boundary over the line (Morpheme or syllable unit (mostly) preserved/word unit separated)

`Level 2`: Word unit kept/constituent separated (in NP; AP; PP; AdvP; the closest broken phrase).

`Level 3`: Constituent (Phrase) kept/clause separated

`Level 4`: Clause kept/sentence separated

`Level 5`: Verse and clause/sentence unit coincides/no enjambment.

`Cross Clause: cc_4 or cc_5`: If the reference noun is separated from the attributive clause by the verse boundary.

```
<lg type="poem">
    <head>THE MYSTERY OF PAIN .</head>
    <lg type="stanza" rhyme="abcd">
        <l met="+--+|-+|-+|" enj="5">
        Pain has an element of blank ;</l>
        <l met="-+|-+|-+|" enj="4">
        It cannot recollect</l>
        <l met="-+|-+|-+|-+|"  enj="3">
        When it began , or if there were</l>
        <l met="-+|-+|-+|" enj="5">
        A day when it was not .</l>
    </lg>
    <lg type="stanza" rhyme="abcb">
    [...]
    </lg>
</lg>
```

Part-Of-Speech

Part-of-Speech (POS) annotation is straightforward, where the sequence of POS is simply aligned to the tokens in the text. There are better practices to annotate POS in TEI. So this form is just for visual aid and illustration.

```
<lg type="poem">
    <head>THE MYSTERY OF PAIN .</head>
    <lg type="stanza" rhyme="abcd">
        <l met="+--+|-+|-+|"   pos="NNP VBZ DT NN IN NN/JJ ;">
        Pain has an element of blank ;</l>
        <l met="-+|-+|-+|"     pos="PRP MD RB VB">
        It cannot recollect</l>
        <l met="-+|-+|-+|-+|"  pos="WRB PRP VBD , CC IN EX VBD">
        When it began , or if there were</l>
        <l met="-+|-+|-+|"     pos="DT NN WRB PRP VBD RB .">
        A day when it was not .</l>
    </lg>
    <lg type="stanza" rhyme="abcb">
    [...]
    </lg>
</lg>
```

Aesthetic Emotion

As discussed in Chapter 5, a line of poetry can have annotation for one (main) emotion, and also annotation for a second (secondary) emotion. The labelset we found to be effective is the following, in order of frequency (from most frequent to least frequent): *Beauty/Joy, Sadness, Uneasiness, Energy/Vitality, Suspense, Awe/Sublime, Humor, Annoyance*

```
<lg type="poem">
    <head>THE MYSTERY OF PAIN .</head>
    <lg type="stanza" rhyme="abcd">
        <l met="+--+|-+|-+|"   emotion="Sadness">
        Pain has an element of blank ;</l>
        <l met="-+|-+|-+|"     emotion="Sadness">
        It cannot recollect</l>
        <l met="-+|-+|-+|-+|"  emotion="Sadness">
        When it began , or if there were</l>
        <l met="-+|-+|-+|"     emotion="Sadness">
        A day when it was not .</l>
    </lg>
    <lg type="stanza" rhyme="abcb">
    [...]
    </lg>
</lg>
```

8.2 Genres in DTA

# of Documents	Genre Label
155	Roman
128	Prosa
128	Lyrik
92	Leichenpredigt
78	Philosophie
76	Recht
70	Drama
46	Technik
44	Medizin
44	Historiographie
40	Geographie
40	Biologie
34	Psychologie
33	Gesellschaft
31	Ökonomie
27	Theologie
27	Philologie
25	Gesellschaftswissenschaften
24	Sprachwissenschaft
23	Physik
23	(Auto)biographie
21	Mathematik
19	Kunst
19	Erbauungsliteratur
18	Politik
17	Verordnung
16	Pädagogik
15	Gartenbau
13	Zoologie
12	Militär
12	Anstandsliteratur
11	Naturwissenschaft
10	Reiseliteratur
9	Chemie
7	Landwirtschaft
7	Kunstgeschichte
7	Handbuch
6	Sonstiges
5	Verslehre

5	Musik
5	Alchemie
4	Hausväterliteratur
4	Flugschrift
4	Buchkunde
4	Bergbau
4	Amtsdruckschrift
3	Sport
3	Novelle
3	Gelegenheitsschrift:Tod
3	Epos
3	Altertumskunde
2	Tierheilkunde
2	Schäferdichtung
2	Poetik
2	Libretto
2	Kochbuch
2	Geschichte
2	Geologie
2	Buchwesen
2	Brief
2	Astronomie
2	Architektur
1	Vertrag
1	Streitschrift
1	Satire
1	Rhetorik
1	Reformschrift
1	Rede
1	Pflanzenbuch
1	Ordensliteratur:Jesuiten
1	Musikwissenschaft
1	Literaturwissenschaft
1	Lexikon
1	Kolportageliteratur
1	Kinderliteratur
1	Katechismus
1	Kameralwissenschaft
1	Kalender
1	Humboldts
1	Grammatik
1	Glasherstellung
1	Gebrauchsliteratur

1	Fest
1	Epigramm
1	Briefsteller
1	Biographie

Table 8.1: Genre Labels in DTA

8.3 Poems in ANTI-K

Table 8.2: Poems in Antikoerperchen (ANTI-K) Corpus with Publication year and Author Name

1624_Ach_Liebste_lass_uns_eilen_Martin_Opitz
1624_Carpe_diem_Martin_Opitz
1636_Aennchen_von_Tharau_Simon_Dach
1636_Traenen_des_Vaterlandes_Anno_1636_Andreas_Gryphius
1637_An_die_Welt_Andreas_Gryphius
1637_Menschliches_Elende_Andreas_Gryphius
1640_Traenen_in_schwerer_Krankheit_Andreas_Gryphius
1641_An_Sich_Paul_Fleming
1650_Ebenbild_unseres_Lebens_Andreas_Gryphius
1650_Morgensonett_Andreas_Gryphius
1658_Einsamkeit_Andreas_Gryphius
1662_Auf_meinen_bestuermeten_Lebens-Lauff_Catharina_Regina_von_Greiffenberg
1663_Es_ist_alles_eitel_Alles_ist_eitel_Andreas_Gryphius
1670_Vergaenglichkeit_der_Schoenheit_I_Christian_Hoffmann_von_Hoffmannswaldau
1670_Vergaenglichkeit_der_Schoenheit_II_Christian_Hoffmann_von_Hoffmannswaldau
1695_Venus_Denn_lieben_ist_nichts_mehr_-_als_eine_schifferey_Daniel_Casper_von_Lohenstein
1753_Das_Rosenband_Friedrich_Gottlieb_Klopstock
1771_Maifest_Mailied_Johann_Wolfgang_von_Goethe
1774_An_Schwager_Kronos_Johann_Wolfgang_von_Goethe
1774_Faust_I,_Der_Koenig_in_Thule_Johann_Wolfgang_von_Goethe
1774_Ganymed_Johann_Wolfgang_von_Goethe
1774_Kuenstlers_Abendlied_Johann_Wolfgang_von_Goethe
1774_Prometheus_Bedecke_deinen_Himmel,_Zeus_Johann_Wolfgang_von_Goethe
1775_Auf_dem_See_aufm_Zuerichersee_Johann_Wolfgang_von_Goethe
1775_Das_Landleben_Ludwig_Christoph_Heinrich_Hoelty
1775_Neue_Liebe,_neues_Leben_Johann_Wolfgang_von_Goethe
1776_Rastlose_Liebe_Johann_Wolfgang_von_Goethe
1776_Wanderers_Nachtlied_Johann_Wolfgang_von_Goethe
1778_An_den_Mond_Johann_Wolfgang_von_Goethe
1778_Erlkoenig_Wer_reitet_so_spaet_durch_Johann_Wolfgang_von_Goethe

1779_Abendlied_Der_Mond_ist_aufgegangen_Matthias_Claudius
1782_Im_Winter_Matthias_Claudius
1783_Das_Goettliche_Johann_Wolfgang_von_Goethe
1789_Heidenroeslein_Heideroeslein_Johann_Wolfgang_von_Goethe
1789_Willkommen_und_Abschied_Es_schlug_mein_Herz_Johann_Wolfgang_von_Goethe
1795_Die_Teilung_der_Erde_Friedrich_Schiller
1795_Naehe_des_Geliebten_Johann_Wolfgang_von_Goethe
1797_Der_Handschuh_Friedrich_Schiller
1797_Der_Zauberlehrling_Johann_Wolfgang_von_Goethe
1798_Andacht_Ludwig_Tieck
1798_An_die_Parzen_Friedrich_Hoelderlin
1798_Die_Buergschaft_Friedrich_Schiller
1800_Wenn_nicht_mehr_Zahlen_und_Figuren_Novalis
1802_Der_Spinnerin_Nachtlied_Es_sang_vor_langen_Jahren_Clemens_Brentano
1802_Fruehling_Clemens_Brentano
1802_Kassandra_Friedrich_Schiller
1803_Wenn_die_Sonne_weggegangen_Clemens_Brentano
1805_Der_Kuss_im_Traume_Karoline_von_Guenderrode
1805_Haelfte_des_Lebens_Friedrich_Hoelderlin
1806_Lass_rauschen_Lieb,_lass_rauschen_Achim_von_Arnim
1807_Maechtiges_Ueberraschen_Johann_Wolfgang_von_Goethe
1810_Abschied_Joseph_von_Eichendorff
1810_Gefunden_Johann_Wolfgang_von_Goethe
1813_Das_zerbrochene_Ringlein_Joseph_von_Eichendorff
1813_Grenzen_der_Menschheit_Johann_Wolfgang_von_Goethe
1815_Das_Maedchen_Joseph_von_Eichendorff
1815_Nachtlied_Joseph_von_Eichendorff
1815_Waldgespraech_Joseph_von_Eichendorff
1815_Zwielicht_Joseph_von_Eichendorff
1817_Der_Abend_Joseph_von_Eichendorff
1818_Die_zwei_Gesellen_Joseph_von_Eichendorff
1822_Aus_alten_Maerchen_winkt_es_Heinrich_Heine
1823_Der_frohe_Wandersmann_Joseph_von_Eichendorff
1824_Die_Lore-Ley_Ich_weiss_nicht_was_soll_es_bedeuten_Heinrich_Heine
1826_Lebenslauf_Friedrich_Hoelderlin
1826_Nachts_Joseph_von_Eichendorff
1827_Mein_Herz,_mein_Herz_ist_traurig_Heinrich_Heine
1827_Um_Mitternacht_Eduard_Moerike
1830_An_die_Geliebte_Eduard_Moerike
1833_In_der_Fremde_Joseph_von_Eichendorff
1833_Liebe_und_Fruehling_Hoffmann_von_Fallersleben
1834_Lockung_Joseph_von_Eichendorff
1834_Sehnsucht_Es_schienen_so_golden_die_Sterne_Joseph_von_Eichendorff
1835_Wuenschelrute_Joseph_von_Eichendorff
1837_Im_Abendrot_Joseph_von_Eichendorff

1837_Mondnacht_Joseph_von_Eichendorff
1837_Neue_Liebe_Joseph_von_Eichendorff
1840_In_seinem_Garten_wandelt_er_allein_Theodor_Storm
1841_Aufruf_Georg_Herwegh
1841_Das_Lied_vom_Hasse_Georg_Herwegh
1841_Frische_Fahrt_Joseph_von_Eichendorff
1843_Wiegenlied_Georg_Herwegh
1844_Das_Hungerlied_Georg_Weerth
1844_Deutschland._Ein_Wintermaerchen,_Caput_18_XVIII_Heinrich_Heine
1844_Deutschland._Ein_Wintermaerchen,_Caput_7_VII_Heinrich_Heine
1844_Die_schlesischen_Weber_Weberlied_Heinrich_Heine
1844_Zur_Beruhigung_Heinrich_Heine
1851_Die_Stadt_Theodor_Storm
1852_Wiegenlied_Clemens_Brentano
1853_Nachtzauber_Joseph_von_Eichendorff
1854_Erinnerung_aus_Kraehwinkels_Schreckenstagen_Heinrich_Heine
1860_An_eine,_die_vorueberging_Charles_Baudelaire
1870_Kriegslied_Emanuel_Geibel
1879_Abendlied_Gottfried_Keller
1882_Zwei_Segel_Conrad_Ferdinand_Meyer
1883_In_einer_grossen_Stadt_Detlev_von_Liliencron
1886_John_Maynard_Theodor_Fontane
1887_Vereinsamt_Friedrich_Nietzsche
1890_Siehst_du_die_Stadt_Hugo_von_Hofmannsthal
1892_Regen_in_der_Daemmerung_Hugo_von_Hofmannsthal
1896_Die_Beiden_Hugo_von_Hofmannsthal
1902_Der_Panther_Rainer_Maria_Rilke
1902_Herbsttag_Rainer_Maria_Rilke
1903_Weltende_Else_Lasker-Schueler
1906_Das_Karussell_Rainer_Maria_Rilke
1906_Ein_alter_Tibetteppich_Else_Lasker-Schueler
1907_Liebeslied_Rainer_Maria_Rilke
1907_Mensch_im_Eisen_Heinrich_Lersch
1907_Todeserfahrung_Rainer_Maria_Rilke
1908_Papageien-Park_Rainer_Maria_Rilke
1909_Verfall_Georg_Trakl
1910_Berlin_VIII_Georg_Heym
1910_Der_Gott_der_Stadt_Georg_Heym
1910_Die_Irren_Georg_Heym
1910_Die_Tote_im_Wasser_Georg_Heym
1910_Nach_der_Schlacht_Georg_Heym
1911_Berlin_I_Georg_Heym
1911_Blauer_Abend_in_Berlin_Oskar_Loerke
1911_Der_Krieg_Georg_Heym
1911_Die_Daemonen_der_Stadt_Georg_Heym

1911__Die__Stadt__Georg__Heym
1911__Fabrikstrasse__Tags__Paul__Zech
1911__Lover's__Seat__Ernst__Stadler
1911__Weltende__Jakob__van__Hoddis
1912__Abendlied__Georg__Trakl
1912__Auf__der__Terrasse__des__Cafe__Josty__Paul__Boldt
1912__Die__Daemmerung__Georg__Trakl
1912__Halber__Schlaf__Georg__Heym
1912__In__den__Nachmittag__gefluestert__Georg__Trakl
1912__Psalm__Georg__Trakl
1912__Sonntagnachmittag__Alfred__Lichtenstein
1912__Stadt__Gerrit__Engelke
1912__Umbra__Vitae__Die__Menschen__stehen__vorwaerts__in__den__Strassen__Georg__Heym
1913__An__die__Verstummten__Georg__Trakl
1913__Die__Daemmerung__Alfred__Lichtenstein
1913__Die__Raben__Georg__Trakl
1913__Die__Stadt__Alfred__Lichtenstein
1913__Im__Daemmer__Paul__Zech
1913__Im__Winter__Georg__Trakl
1913__In__der__Welt__Paul__Boldt
1913__Punkt__Alfred__Lichtenstein
1913__Vorstadt__im__Foehn__Georg__Trakl
1914__Der__Aufbruch__Ernst__Stadler
1914__Der__Spruch__Ernst__Stadler
1914__Grodek__Georg__Trakl
1914__Morgens__Jakob__van__Hoddis
1914__Staedter__Alfred__Wolfenstein
1914__Sturmangriff__August__Stramm
1914__Traum__August__Stramm
1915__Patrouille__August__Stramm
1915__Untreu__August__Stramm
1915__Vorfruehling__August__Stramm
1917__Kriegslied__Erich__Muehsam
1924__Nachthimmel__und__Sternenfall__Rainer__Maria__Rilke
1932__Augen__in__der__Grossstadt__Kurt__Tucholsky
1936__Deutschland__im__Marschschritt__Herybert__Menzel

8.4 Regular Expressions to Determine Verse Measures

```
def get_versification(meter_line, input_type='list', measure_type='f', greek_forms=True):
        # full = f
        # short = s
        # intermediate = i
        # length = l
        meter = meter_line
        if input_type == 'list':
                meter = ''.join(meter_line)
        meter = re.sub('\+', 'I', meter)
        meter = re.sub('\-', 'o', meter)
        #print(meter)
        hexameter =       re.compile('^Ioo?Ioo?Ioo?Ioo?IooIo$')
        alxiambichexa =   re.compile("^oIoIoIoIoIoIo?$")
        asklepiade =      re.compile("^IoIooIIooIoI$") # 12 Ode
        glykoneus =       re.compile("^IoIooIoI$")     # 8  Ode
        pherekrateus =    re.compile("^IoIooIo$")      # 7  Ode
        iambelegus =      re.compile('^oIoIoIoIIooIooI$')
        elegiambus =      re.compile('^IooIooIo?oIoIoIoo?$')
        diphilius =       re.compile('^IooIooI..IooI..$')
        prosodiakos =     re.compile('^.IooIooII?$')
        sapphicusmaior =  re.compile('^IoIIIooIIooIoI.$')
        sapphicusminor =  re.compile('^IoI.IooIoI.$')
        iambicoctaplus =  re.compile("^oIoIoIoIoIoIoIoIo?")
        iambicsepta =     re.compile("^oIoIoIoIoIoIoIo?$")
        iambicpenta =     re.compile("^oIoIoIoIoIo?$")
        iambicpentaspond= re.compile("^IIoIoIoIoIo?$")
        iambictetra =     re.compile("^.IoIoIoIo?$")
        iambictri =       re.compile("^.IoIoIo?$")
        iambicdi =        re.compile("^.IoIo?$")
        iambic =          re.compile("^.IoIo?")
        iambicsingle =    re.compile("^oI$")
        trochaicoctaplus =   re.compile('^IoIoIoIoIoIoIoIo?')
        trochaicsepta =      re.compile('^IoIoIoIoIoIoIo?$')
        trochaichexa =       re.compile('^IoIoIoIoIoIo?$')
        trochaicpenta =      re.compile('^IoIoIoIoIo?$')
        trochaictetra =      re.compile('^IoIoIoIo?$')
        trochaictri =        re.compile('^IoIoIo?$')
        trochaicdi =         re.compile('^IoIo?$')
        trochaicsingle =     re.compile("^Io$")
        trochaic =           re.compile('^IoIo?')
        amphibrachdi =         re.compile('^o?IooIo$')
```

```
amphibrachdimix =      re.compile('^oIooIo')
amphibrachtri =        re.compile('^oIooIooIo?$')
amphibrachtriplus =    re.compile('^oIooIooIo')
amphibrachtetra =      re.compile('^oIooIooIooIo?$')
amphibrachtetraplus =  re.compile('^oIooIooIooIo')
amphibrachpentaplus =  re.compile('^oIooIooIooIooIo?')
amphibrachsingle =     re.compile('^oIo$')
adoneus =              re.compile('^IooI.$')
adoneusspond =         re.compile('^IooII$')
dactylicpenta =     re.compile('^IooIooIooIooIo?o?$')
dactylicpentaplus = re.compile('^IooIooIooIooIooIoo')
dactylictetra =     re.compile('^IooIooIooIo?o?$')
dactylictetraplus = re.compile('^IooIooIooIoo')
dactylictri =       re.compile('^IooIooIo?o?$')
dactylictriplus =   re.compile('^IooIooIoo')
dactylicdi =        re.compile('^IooIoo$')
dactylicdiplus =    re.compile('^IooIoo')
amphibrachiambicmix =     re.compile('^oI.*oIooIoo?I')
amphibrachtrochaicmix =   re.compile('^Io.*oIooIoo?I')
artemajor =         re.compile('^oIooIooIooIo$')
artemajorhalf =     re.compile('^oIooIo$')
iambicseptainvert=  re.compile("^IooIoIoIoIoIoIo?$")
iambichexainvert =  re.compile("^IooIoIoIoIoIo?$")
iambicpentainvert=  re.compile("^IooIoIoIoIo?$")
iambictetrainvert=  re.compile("^IooIoIoIo?$")
iambictriinvert =   re.compile("^IooIoIo?$")
iambicinvert =      re.compile('^IooIoI')
trochaicextrasyll=  re.compile('^I.*IooI.+')
iambicextrasyll=    re.compile('^o.*IooI.+')
#iambiccholstrict = re.compile('.IoI.IoIoII.$')
iambiccholstrict =  re.compile("^oIoIoIoIoIooI$")
iambicchol       =  re.compile('^o?.*IooI$')
zehnsilber =        re.compile('^...I.....I$')
anapaestdiplus =    re.compile('^ooIooI')
anapaesttriplus =   re.compile('^ooIooIooI')
anapaesttetraplus=  re.compile('^ooIooIooIooI')
anapaestinit =      re.compile('^ooI')
dactylicinit =      re.compile('^o?Ioo')
spondeus =          re.compile('^II$')
singleup =          re.compile('^I$')
singledown =        re.compile('^o$')
#alexandriner =     re.compile('oIoIoIoIoIoIo?$')
#adoneus =          re.compile('IooIo$')
#iambicamphibrachcentermix = re.compile('oIoIooIoI$')
```

```
greek = { 'asklepiade':asklepiade,\
          'glykoneus':glykoneus,\
          'pherekrateus':pherekrateus,\
          'iambelegus':iambelegus,\
          'elegiambus':elegiambus,\
          'diphilius':diphilius,\
          'prosodiakos':prosodiakos,\
          'sapphicusmaior':sapphicusmaior,\
          'sapphicusminor':sapphicusminor
         }

adoneus = {
          'adoneus':adoneus,\
          'adoneus.spond':adoneusspond
          }
verses1 = {
          'iambic.octa.plus':iambicoctaplus,\
          'iambic.septa':iambicsepta,\
          'hexameter':hexameter,\
          'alexandrine.iambic.hexa':alxiambichexa,\
          'iambic.penta':iambicpenta,\
          'iambic.penta.spondeus':iambicpentaspond,\
          'iambic.tetra':iambictetra,\
          'iambic.tri':iambictri,\
          'iambic.di':iambicdi,\
          'trochaic.octa.plus':trochaicoctaplus,\
          'trochaic.septa':trochaicsepta,\
          'trochaic.hexa':trochaichexa,\
          'trochaic.penta':trochaicpenta,\
          'trochaic.tetra':trochaictetra,\
          'trochaic.tri':trochaictri,\
          'trochaic.di':trochaicdi
          }

verses2 = {
          'dactylic.penta':dactylicpenta,\
          'dactylic.tetra':dactylictetra,\
          'dactylic.tri':dactylictri,\
          'amphibrach.penta.plus':amphibrachpentaplus,\
          'amphibrach.tetra':amphibrachtetra,\
          'amphibrach.tetra.plus':amphibrachtetraplus,\
          'amphibrach.tri':amphibrachtri,\
          'amphibrach.tri.plus':amphibrachtriplus,\
          'amphibrach.relaxed':amphibrachdi,\
          'dactylic.penta.plus':dactylicpentaplus,\
```

```
            'dactylic.tetra.plus':dactylictetraplus,\
            'dactylic.tri.plus':dactylictriplus,\
            'dactylic.di.plus':dactylicdiplus,\
            'dactylic.di':dactylicdi,\
            'anapaest.tetra.plus':anapaesttetraplus,\
            'anapaest.tri.plus':anapaesttriplus,\
            'anapaest.di.plus':anapaestdiplus,\
            'arte_major':artemajor,\
            'arte_major.half':artemajorhalf
            }
verses3 = {
            'iambic.septa.invert':iambicseptainvert,\
            'iambic.hexa.invert':iambichexainvert,\
            'iambic.penta.invert':iambicpentainvert,\
            'iambic.tetra.invert':iambictetrainvert,\
            'iambic.tri.invert':iambictriinvert,\
            'iambic.invert':iambicinvert,\
            'trochaic.relaxed':trochaicextrasyll,\
            'iambic.relaxed':iambicextrasyll,\
            'iambic.chol.strict':iambiccholstrict,\
            'iambic.relaxed.chol':iambicchol,\
            'amphibrach.single':amphibrachsingle,\
            'amphibrach.iambic.mix':amphibrachiambicmix,\
            'amphibrach.trochaic.mix':amphibrachtrochaicmix,\
            'anapaest.init':anapaestinit,\
            'dactylic.init':dactylicinit,\
            'amphibrach.di.mix':amphibrachdimix,\
            'zehnsilber':zehnsilber,\
            'spondeus':spondeus,\
            'iambic.single':iambicsingle,\
            'trochaic.single':trochaicsingle,\
            'single.down':singledown,\
            'single.up':singleup}
 verses = {}
 if greek_forms == False:
         #verses = verses1 + verses2 + verses3
         verses.update(verses1)
         verses.update(verses2)
         verses.update(verses3)
 if greek_forms == True:
         verses.update(verses1)
         verses.update(greek)
         verses.update(verses2)
         verses.update(adoneus)
         verses.update(verses3)
```

```
        #verses = verses1 + greek + verses2 + adoneus + verses3

label = None
for label, pattern in verses.items():
        result = pattern.match(meter)
        #if label == 'chol.iamb':
        #       result = pattern.search(meter)
        hebungen = meter.count('I')
        counters = {0:'zero', 1:'single', 2:'di', 3:'tri', 4:'tetra', 5:'penta', 6:'hexa', 7:'septa'}
        if hebungen > 7:
                hebungen_label = 'octa.plus'
        else:
                hebungen_label = counters[hebungen]
        if 'relaxed' in label:
                label = re.sub('.relaxed', '.' + hebungen_label + '.relaxed', label)
        if 'relaxed.chol' in label:
                label = re.sub('relaxed.chol', 'chol', label)
        #if 'chol.strict' in label:
        #        label = re.sub('relaxed.chol', 'chol', label)
        if 'chol' in label:
                label = re.sub('.chol', '.' + hebungen_label + '.chol', label)
        if 'iambic.invert' in label:
                label = re.sub('.invert', '.' + hebungen_label + '.invert', label)

        label = re.sub('di.di', 'di', label)
        label = re.sub('tri.tri', 'tri', label)
        label = re.sub('tetra.tetra', 'tetra', label)
        label = re.sub('penta.penta', 'penta', label)
        label = re.sub('hexa.hexa', 'hexa', label)
        label = re.sub('septa.septa', 'septa', label)
        label = re.sub('octa.octa', 'octa', label)
        if result != None:
                split = label.split('.')
                if measure_type == 's':
                        return split[0]
                if measure_type == 'i':
                        return '.'.join(split[:2])
                if measure_type == 'l':
                        return hebungen_label
                else:
                        return label
else: return 'unknown.measure.' + hebungen_label
```

Table 8.3: Verse measures by frequency in DTA determined with automatic annotation.

Measure	Rel. Freq.	Abs. Freq.
all lines	1	493216 (494520)
alexandrine.iambic.hexa	0.270	133297
iambic.tetra	0.172	84956
trochaic.tetra	0.097	48060
iambic.tri	0.081	39854
iambic.penta	0.079	39265
iambic.penta.relaxed	0.020	9865
iambic.tetra.relaxed	0.016	7794
iambic.di	0.015	7514
iambic.tri.relaxed	0.015	7365
iambic.hexa.relaxed	0.014	7141
trochaic.di	0.013	6261
trochaic.penta	0.010	5114
trochaic.tri	0.010	4968
trochaic.single	0.009	4696
amphibrach.single	0.008	4143
anapaest.di.plus	0.008	3907
single.up	0.008	3901
amphibrach.di.relaxed	0.007	3453
amphibrach.tri.plus	0.007	3217
trochaic.hexa	0.006	3047
iambic.single	0.006	3006
trochaic.hexa.relaxed	0.006	2784
amphibrach.tetra	0.006	2781
trochaic.septa	0.006	2750
iambic.hexa.invert	0.005	2579
trochaic.penta.relaxed	0.005	2544
hexameter	0.005	2488
dactylic.di.plus	0.005	2400
iambic.septa	0.005	2254
iambic.tri.chol	0.004	2057
iambic.tetra.invert	0.004	2043
iambic.di.chol	0.004	1994
iambic.tetra.chol	0.004	1876
trochaic.septa.relaxed	0.004	1790
iambic.penta.chol	0.004	1781
amphibrach.tri	0.004	1775
trochaic.octa.plus	0.004	1766
iambic.penta.invert	0.004	1743
single.down	0.003	1554
iambic.septa.relaxed	0.003	1504
dactylic.tetra	0.003	1462
iambic.tri.invert	0.003	1363
trochaic.tetra.relaxed	0.003	1342
prosodiakos	0.003	1338
dactylic.tri	0.003	1338
anapaest.tri.plus	0.002	1138
anapaest.init	0.002	1125
dactylic.init	0.002	866
anapaest.tetra.plus	0.002	860
glykoneus	0.002	854
dactylic.tri.plus	0.002	822
pherekrateus	0.002	801
unknown.measure.hexa	0.001	730
unknown.measure.septa	0.001	718
iambic.hexa.chol.strict	0.001	691
unknown.measure.penta	0.001	667
unknown.measure.zero	0.001	506
iambic.octa.plus	0.001	495
unknown.measure.tri	0.001	461
amphibrach.tetra.plus	0.001	381

Table 8.4: Verse measures by frequency in DTA continued.

Measure	Rel. Freq.	Abs. Freq.
unknown.measure.tetra	0.001	373
iambic.septa.invert	0.001	362
iambic.octa.plus.relaxed	0.001	353
dactylic.penta	0.001	312
iambic.septa.chol	0.001	286
iambic.hexa.chol	0.001	281
unknown.measure.octa.plus	0.001	281
unknown.measure.di	0.001	275
sapphicusminor	0.0	214
amphibrach.penta.plus	0.0	183
trochaic.octa.plus.relaxed	0.0	178
iambic.di.relaxed	0.0	150
elegiambus	0.0	145
trochaic.tri.relaxed	0.0	118
dactylic.tetra.plus	0.0	105
asklepiade	0.0	97
diphilius	0.0	87
iambic.octa.plus.invert	0.0	63
zehnsilber	0.0	49
iambic.penta.spondeus	0.0	25
unknown.measure.single	0.0	16
iambic.octa.plus.octa.plus.chol	0.0	14
iambelegus	0.0	1
adoneus	0.0	1
spondeus	0.0	1
dactylic.penta.plus	0.0	1

8.5 Verse Measure Author Characterizations

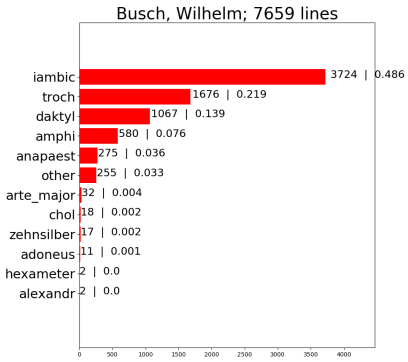

FIGURE 8.1: Measures of Wilhelm Busch

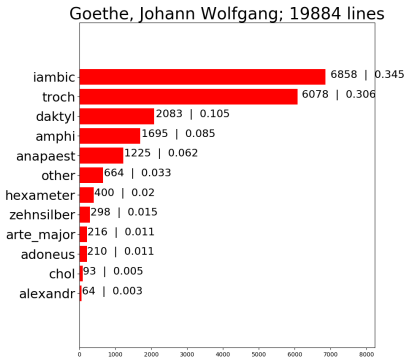

FIGURE 8.2: Measures of Goethe

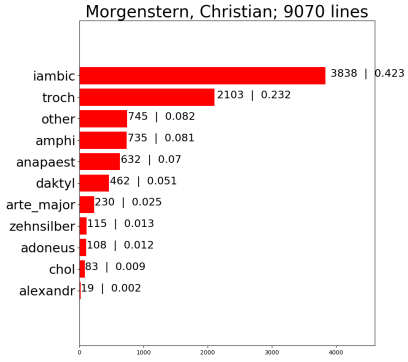

FIGURE 8.3: Measures of Christian Morgenstern

Bibliography

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Apoorv Agarwal, Augusto Corvalan, Jacob Jensen, and Owen Rambow. 2012. Social network analysis of alice in wonderland. In *Proceedings of the NAACL-HLT 2012 Workshop on computational linguistics for literature*, pages 88–96.

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2016a. Machine learning for metrical analysis of English poetry. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 772–781, Osaka, Japan. The COLING 2016 Organizing Committee.

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2019. A comparison of feature-based and neural scansion of poetry. *RANLP 2019, arXiv preprint arXiv:1711.00938*.

Manex Agirrezabal, Aitzol Astigarraga, Bertol Arrieta, and Mans Hulden. 2016b. Zeuscansion: a tool for scansion of english poetry. *Journal of Language Modelling*, 4.

Mark Algee-Hewitt, Ryan Heuser, Maria Kraxenberger, JD Porter, Jonny Sensenbaugh, and Justin Tackett. 2014. The stanford literary lab transhistorical poetry project phase ii: Metrical form. In *DH*.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005a. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 579–586.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005b. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Ouais Alsharif, Deema Alshamaa, and Nada Ghneim. 2013. Emotion classification in arabic poetry using machine learning. *International Journal of Computer Applications*, 65(16).

Susanne Althoff. 2016. Algorithms could save book publishing—but ruin novels, wired magazine.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Text, Speech and Dialogue*, pages 196–205, Berlin, Heidelberg. Springer Berlin Heidelberg.

Arto Anttila, Timothy Dozat, Daniel Galbraith, and Naomi Shapiro. 2018. Sentence stress in presidential speeches. *Lingbuzz Preprints*.

Arto Anttila and Ryan Heuser. 2016. Phonological and metrical variation across genres. In *Proceedings of the Annual Meetings on Phonology*, volume 3.

Shlomo Argamon. 2019. Computational register analysis and synthesis. *Register Studies, Forthcoming*.

Aristoteles. 1877. *The Rhetoric of Aristotle. Vol. 3.* Cambridge University Press. Ed.: Edward Meredith Cope and John Edwin Sandys.

Aristoteles. 2012. *Poetics.* [orig. 335 BCE], Hackett Indianapolis.

Jan Assmann. 2006. Form as a mnemonic device: Cultural texts and cultural memory.

Zhenisbek Assylbekov, Rustem Takhanov, Bagdat Myrzakhmetov, and Jonathan N. Washington. 2017. Syllable-aware neural language models: A failure to beat character-aware ones. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1866–1872, Copenhagen, Denmark. Association for Computational Linguistics.

Derek Attridge. 2014. *The rhythms of English poetry.* Routledge.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Hannes Bajohr. 2021. Künstliche intelligenz und digitale literatur. theorie und praxis konnektionistischen schreibens. *Digitale Literatur II. Sonderband Text+Kritik*, page 174–185.

Rebecca Balacarcel. 2013. Understanding "i felt a funeral in my brain". *YouTube, https://www.youtube.com/watch?v=Q9ISE11zqic.*

David Bamman, Chris Dyer, and Noah A Smith. 2014a. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 828–834.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014b. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

David Bamman, Olivia Lewke, and Anya Mansoor. 2019. An annotated dataset of coreference in english literature. *arXiv preprint arXiv:1912.01140*.

David Bamman, Ted Underwood, and Noah A Smith. 2014c. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379.

Linda Barros, Pilar Rodriguez, and Alvaro Ortigosa. 2013. Automatic classification of literature pieces by emotion detection: A study on quevedo's poetry. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 141–146. IEEE.

Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2009. On the syllabification of phonemes. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 308–316.

Timo Baumann, Hussein Hussein, and Burkhard Meyer-Sickendiek. 2018. Style detection for free verse poetry from text and speech. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1929–1940.

MH Beissinger. 2012. Oral poetry. *The Princeton encyclopedia of poetry and poetics*, pages 978–81.

Jonas Belouadi and Steffen Eger. 2022. Bygpt5: End-to-end style-conditioned poetry generation with token-free language models. *arXiv preprint arXiv:2212.10474*.

Walter Benjamin. 2016. *Goethes Wahlverwandtschaften*. BoD–Books on Demand.

Max Bense. 1969. *Einführung in die informationstheoretische Ästhetik*. Rowohlt (-Taschenbuch-Verlag).

Thomas Berg. 1990. Unreine reime als evidenz für die organisation phonologischer merkmale. *Zeitschrift für Sprachwissenschaft*, 9(1-2):3–27.

Wolfgang Beutin, Klaus Ehlert, Wolfgang Emmerich, Helmut Hoffacker, Bernd Lutz, and Volker Meid. 1994. *Deutsche Literaturgeschichte: von den Anfängen bis zur Gegenwart*. Springer.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2020. Cross-lingual contextualized topic models with zero-shot learning. *arXiv preprint arXiv:2004.07737*.

Douglas Biber. 1989. A typology of english texts. *Linguistics*, 27(1):3–44.

Douglas Biber, Ulla Connor, and Thomas Upton. 2007. Discourse on the move. *Using corpus analysis to describe discourse structure*.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press.

George David Birkhoff. 2013. *Aesthetic measure*. Harvard University Press.

Christoph Bläsi. 2020. Kl im verlagswesen. In *Maschinen der Kommunikation*, pages 167–187. Springer.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Andre Blessing, Peggy Bockwinkel, Nils Reiter, and Marcus Willand. 2016. Dramenwerkbank. automatische sprachverarbeitung zur analyse von figurenrede. *DHd2016. Konferenzabstracts, Universität Leipzig*, 7(12.3):2016.

Stefan Blohm, Valentin Wagner, Matthias Schlesewsky, and Winfried Menninghaus. 2018. Sentence judgments and the grammar of poetry: Linking linguistic structure and poetic effect. *Poetics*, 69:41–56.

Klemens Bobenhausen. 2011. The metricalizer2–automated metrical markup of german poetry. *Current Trends in Metrical Analysis, Bern: Peter Lang*, pages 119–131.

Marcel Bollmann. 2013. Pos tagging for historical texts with sparse training data. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 11–18.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ryan L Boyd. 2017. Psychological text analysis in the digital humanities. In *Data analytics in digital humanities*, pages 161–189. Springer.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015a. Distinguishing voices in the waste land using computational stylistics. In *Linguistic Issues in Language Technology, Volume 12, 2015-Literature Lifts up Computational Linguistics*.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015b. Gutentag: an nlp-driven tool for digital humanities research in the project gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47.

Barry Brummett. 2018. *Techniques of close reading*. Sage Publications.

Sven Buechel and Udo Hahn. 2017a. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion

analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017b. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain. Association for Computational Linguistics.

Michael Burke. 2017. Rhetoric and poetics: The classical heritage of stylistics. In *The Routledge handbook of stylistics*, pages 29–48. Routledge.

John Burrows. 2002. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17(3):267–287.

John F Burrows. 1992. Not unles you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7(2):91–109.

Thomas Carper and Derek Attridge. 2020. *Meter and meaning: an introduction to rhythm in poetry*. Routledge.

Deniz Cevher, Sebastian Zepf, and Roman Klinger. 2019. Towards multimodal emotion recognition in german speech events in cars using transfer learning. In *Conference on Natural Language Processing (KONVENS)*.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2067–2075. JMLR.org.

Marcus Tullius Cicero. 2013. *De oratore*. BG Teubner.

Isobelle Clarke and Jack Grieve. 2017. Dimensions of abusive language on twitter. In *Proceedings of the first workshop on abusive language online*, pages 1–10.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *LREC*.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229, Hong Kong, China. Association for Computational Linguistics.

Arpita Das, Harish Yenala, Manoj Chinnakotla, and Manish Shrivastava. 2016. Together we stand: Siamese networks for similar question retrieval. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 378–387.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Margaret Dickie. 1990. Reperiodization: The example of emily dickinson. *College English*, 52(4):397–409.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation

models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1147–1156, Copenhagen, Denmark.

Steffen Eger and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 52–58, Berlin, Germany. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. 2017. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 138–147.

Chris Emmery, Ákos Kádár, and Grzegorz Chrupała. 2021. Adversarial stylometry in the wild: Transferable lexical substitution attacks on author profiling. *arXiv preprint arXiv:2101.11310*.

Chris Emmery, Enrique Manjavacas, and Grzegorz Chrupała. 2018. Style obfuscation by invariance. *arXiv preprint arXiv:1805.07143*.

Alex Estes and Christopher Hench. 2016. Supervised machine learning for hybrid meter. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 1–8.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, pages 417–422.

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. Understanding and explaining delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32(suppl_2):ii4–ii16.

Nigel Fabb. 1997. *Linguistics and literature: Language in the verbal arts of the world*. Blackwell.

Nigel Fabb and Morris Halle. 2008. *Meter in poetry: A new theory*. Cambridge University Press.

Nigel Fabb and Morris Halle. 2010. *Meter in Poetry: A New Theory*. Cambridge University Press, Cambridge, UK.

Dean Falk. 2004. Prelinguistic evolution in early hominins: Whence motherese? *Behavioral and brain sciences*, 27(4):491.

Kirill Fayn, Steven Willemsen, R Muralikrishnan, Bilquis Castaño Manias, Winfried Menninghaus, and Wolff Schlotz. 2021. Full throttle: Demonstrating the speed, accuracy, and validity of a new method for continuous two-dimensional self-report and annotation. *Behavior research methods*, pages 1–15.

Irving Finkel. 2019. Cracking ancient codes: Cuneiform writing - with irving finkel. *THE ROYAL INSTITUTION, YouTube Channel, https://www.youtube.com/watch?v=PfYYraMgiBA*.

David Finkelstein. 2008. History of the book, authorship, book design, and publishing. *Handbook of Research on Writing: History, Society, School, Individual, Text*, pages 65–79.

Ruth Finnegan. 2012. *Oral Literature in Africa*, volume 1 of *World Oral Literature Series*. Open Book Publishers, Cambridge, UK.

Frank Fischer and Daniil Skorinkin. 2021. Social network analysis in russian literary studies. In *The Palgrave Handbook of Digital Russia Studies*, pages 517–536. Palgrave Macmillan, Cham.

Stanley Fish. 1980. What is stylistics and why are they saying such terrible things about it? In *Is there a text in this class?: The authority of interpretive communities*, pages 0–0. Columbia University Press.

Horst Joachim Frank. 1980. *Handbuch der deutschen Strophenformen*. Hanser.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Liye Fu, Susan Fussell, and Cristian Danescu-Niculescu-Mizil. 2020. Facilitating the communication of politeness through fine-grained paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5127–5140.

Hans Georg Gadamer. 1960. Wahrheit und methode grundzüge einer philosophischen hermeneutik.

Uta Gaidys, Evelyn Gius, Margarete Jarchow, Gertraud Koch, Wolfgang Menzel, Dominik Orth, and Heike Zinsmeister. 2017. Project description–herma: Automated modelling of hermeneutic processes. *Hamburger Journal für Kulturanthropologie (HJK)*, (7):119–123.

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1764.

Antonio García-Berrio. 2016. *A theory of the literary text*, volume 17. Walter de Gruyter GmbH & Co KG.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498.

Gérard Genette. 1992. Fiktion und diktion. übers. v. heinz jatho. *München: Fink [frz. 1991]*.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191.

Claus J Gigl. 2008. Abitur-wissen deutsch. *Deutsche Literaturgeschichte. Freising.*

Ken J Gilhooly and Robert H Logie. 1980. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior research methods & instrumentation*, 12(4):395–427.

Evelyn Gius and Janina Jacke. 2017. The hermeneutic profit of annotation: On preventing and fostering disagreement in literary analysis. *International Journal of Humanities and Arts Computing*, 11(2):233–254.

Evelyn Gius and Janina Jacke. 2022. Are computational literary studies structuralist? *Journal of Cultural Analytics*, 7(4).

Andrew Goldstone and Ted Underwood. 2014. The quiet transformations of literary studies: What thirteen thousand scholars could tell us. *New Literary History*, 45(3):359–384.

Hugo Gonçalo Oliveira. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages

11–20, Santiago de Compostela, Spain. Association for Computational Linguistics.

Jack Goody. 1987. *The Interface between the Written and the Oral*. Cambridge University Press.

Rob Van der Goot, Barbara Plank, and Malvina Nissim. 2017. To normalize, or not to normalize: The impact of normalization on part-of-speech tagging. *arXiv preprint arXiv:1707.05116*.

Amitha Gopidi and Aniket Alam. 2019. Computational analysis of the historical changes in poetry and prose. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 14–22.

Bethany Gray, Jesse Egbert, and Douglas Biber. 2017. Exploring methods for evaluating corpus representativeness. In *Corpus Linguistics International Conference*.

Erica Greene, Tugba Bodrumlu, and Kevin Knight. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 524–533.

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing*, 22(3):251–270.

Jack Grieve, Isobelle Clarke, Emily Chiang, Hannah Gideon, Annina Heini, Andrea Nini, and Emily Waibel. 2019a. Attributing the bixby letter using n-gram tracing. *Digital Scholarship in the Humanities*, 34(3):493–512.

Jack Grieve, Chris Montgomery, Andrea Nini, Akira Murakami, and Diansheng Guo. 2019b. Mapping lexical dialect variation in british english using twitter. *Frontiers in Artificial Intelligence*, 2:11.

H. S. Gross. 2017. Prosody.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pages 67–71.

E Dario Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, 4:47–60.

Loitongbam Gyanendro Singh, Lenin Laitonjam, and Sanasam Ranbir Singh. 2016. Automatic syllabification for Manipuri language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 349–357, Osaka, Japan. The COLING 2016 Organizing Committee.

Susanne Haaf, Alexander Geyken, and Frank Wiegand. 2014. The dta "base format": A tei subset for the compilation of a large reference corpus of printed text from multiple sources. *Journal of the Text Encoding Initiative*, (8).

Thomas Haider. 2021. Metrical tagging in the wild: Building and annotating poetry corpora with rhythmic features. *Proceedings of the European Association for Computational Linguistics, arXiv:2102.08858*.

Thomas Haider and Steffen Eger. 2019. Semantic change and emerging tropes in a large corpus of new high german poetry. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 216–222.

Thomas Haider, Steffen Eger, Evgeny Kim, Roman Klinger, and Winfried Menninghaus. 2020. Po-emo: Conceptualization, annotation, and model-

ing of aesthetic emotions in german and english poetry. *arXiv preprint arXiv:2003.07723*, pages 1652–1663.

Thomas Haider and Jonas Kuhn. 2018. Supervised rhyme detection with siamese recurrent networks. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 81–86.

Thomas Haider and Alexis Palmer. 2017. Modeling communicative purpose with functional style: Corpus and features for german genre and register analysis. In *Proceedings of the Workshop on Stylistic Variation*, pages 74–84.

Thomas Nikolaus Haider. 2019. Diachronic topics in new high german poetry. In *In Proceedings of the International Digital Humanities Conference DH2019, Utrecht*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Kristin Hanson and Paul Kiparsky. 1996. A parametric theory of poetic meter. *Language*, pages 287–335.

Stefan Hartmann. 2014. The diachronic change of german nominalization patterns: An increase in prototypicality. In *Selected Papers from the 4th UK Cognitive Linguistics Conference. Lancaster: Cognitive Linguistics Association*, pages 52–171. Citeseer.

Raymond C Hawkins II and Ryan L Boyd. 2017. Such stuff as dreams are made on: Dream language, liwc norms, and personality correlates. *Dreaming*, 27(2):102.

David Helbig, Enrica Troiano, and Roman Klinger. 2020. Challenges in emotion style transfer: An exploration with a lexical substitution pipeline. *arXiv preprint arXiv:2005.07617*.

Klaus W Hempfer. 2008. Überlegungen zur historischen begründung einer systematischen lyriktheorie. *Sprachen der Lyrik. Von der Antike bis zur digitalen Poesie. Für Gerhard Regn anlässlich seines*, 60:33–60.

Christopher Hench. 2017. Phonological soundscapes in medieval poetry. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 46–56.

Aurélie Herbelot. 2014. The semantics of poetry: a distributional reading. *Digital Scholarship in the Humanities*, 30(4):516–531.

J Berenike Herrmann, Karina van Dalen-Oskam, and Christof Schöch. 2015. Revisiting style, a key concept in literary studies. *Journal of literary theory*, 9(1):25–52.

J Berenike Herrmann, Arthur M Jacobs, and Andrew Piper. 2021. Computational stylistics. *Handbook of Empirical Literary Studies*, page 451.

Lena Hettinger, Fotis Jannidis, Isabella Reger, and Andreas Hotho. 2016. Classification of literary subgenres. In *The 3rd conference of the Digital Humanities DHd*.

Johann Christian August Heyse. 1827. Theoretisch-praktische grammatik oder lehrbuch zum reinen und richtigen sprechen, lesen und schreiben der deutschen sprache.

Susan Höivik and Kurt Luger. 2009. Folk media for biodiversity conservation: A pilot project from the himalaya-hindu kush. *International Communication Gazette*, 71(4):321–346.

Jack Hopkins and Douwe Kiela. 2017. Automatically generating rhythmic verse with neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 168–178.

Yufang Hou and Anette Frank. 2015. Analyzing sentiment in classical Chinese poetry. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 15–24, Beijing, China. Association for Computational Linguistics.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th international conference on World Wide Web*, pages 452–461.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394.

Joshua Conrad Jackson, Joseph Watts, Teague R Henry, Johann-Mattis List, Robert Forkel, Peter J Mucha, Simon J Greenhill, Russell D Gray, and Kristen A Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Arthur M Jacobs. 2015. Neurocognitive poetics: methods and models for investigating the neuronal and cognitive-affective bases of literature reception. *Frontiers in human neuroscience*, 9:186.

Arthur M Jacobs. 2018. The gutenberg english poetry corpus: exemplary quantitative narrative analyses. *Frontiers in Digital Humanities*, 5:5.

Roman Jakobson. 1960. Linguistics and poetics. In *Style in language*, pages 350–377. MA: MIT Press.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161*.

Harsh Jhamtani, Sanket Vaibhav Mehta, Jaime Carbonell, and Taylor Berg-Kirkpatrick. 2019. Learning rhyming constraints using structured adversaries. *arXiv preprint arXiv:1909.06743*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2020. Deep learning for text style transfer: A survey. *arXiv preprint arXiv:2011.00416*.

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

Matthew L Jockers and David Mimno. 2013. Significant themes in 19th-century literature. *Poetics*, 41(6):750–769.

Philip N. Johnson-Laird and Keith Oatley. 2016. *Handbook of emotions*, chapter Emotions in Music, Literature, and Film. Guilford Publications.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the workshop on noisy user-generated text*, pages 9–18.

Werner Jung. 1997. *Kleine Geschichte der Poetik*. Junius-Verlag.

Sofoklis Kakouros, Joris Pelemans, Lyan Verwimp, Patrick Wambacq, and Okko Räsänen. 2016. Analyzing the contribution of top-down lexical and

bottom-up acoustic cues in the detection of sentence prominence. *Proceedings Interspeech 2016*, 8:1074–1078.

Immanuel Kant. 2001. *Critique of the Power of Judgment*. (P. Guyer & E. Matthews, Trans.). Cambridge, England: Cambridge University Press (Original work published 1790).

Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 8–17.

Justine T Kao and Dan Jurafsky. 2015. A computational analysis of poetic style. *LiLT (Linguistic Issues in Language Technology)*, 12.

Jonah Katz. 2015. Hip-hop rhymes reiterate phonological typology. *Lingua*, 160:54–73.

Vaibhav Kesarwani, Diana Inkpen, Stan Szpakowicz, and Chris Tanasescu. 2017. Metaphor detection in a poetry corpus. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–9.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65.

Paul Kiparsky. 1973. The role of linguistics in a theory of poetry. *Daedalus*, pages 231–244.

Paul Kiparsky. 2020. Metered verse. *Annual Review of Linguistics*, 6:25–44.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.

Roman Klinger, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur. 2018. IEST: WASSA-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium. Association for Computational Linguistics.

Christine A Knoop, Stefan Blohm, Maria Kraxenberger, and Winfried Menninghaus. 2019. How perfect are imperfect rhymes? effects of phonological similarity and verse context on rhyme perception. *Psychology of Aesthetics, Creativity, and the Arts*.

Christine A Knoop, Valentin Wagner, Thomas Jacobsen, and Winfried Menninghaus. 2016. Mapping the aesthetic space of literature "from below". *Poetics*, 56:35–49.

Otto Knörrich. 1971. *Die deutsche Lyrik der Gegenwart*, volume 401. Kröner.

Otto Knörrich. 2005. Lexikon lyrischer formen. 2., überarb. *Aufl. Stuttgart: Kröner (479)*.

Corina Koolen, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout. 2020. Literary quality in the eye of the dutch reader: The national reader survey. *Poetics*, 79:101439.

Maximilian Köper and Sabine Schulte Im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 german lemmas. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2595–2598.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2013. Authorship attribution: What's easy and what's hard? *Available at SSRN 2274891*.

Maria Kraxenberger and Winfried Menninghaus. 2016. Mimological reveries? disconfirming the hypothesis of phono-emotional iconicity in poetry. *Frontiers in Psychology*, 7:1779.

Helmut Kreuzer and Rul Gunzenhäuser. 1965. *Mathematik und Dichtung: Versuche zur Frage einer exakten Literaturwissenschaft*. Nymphenburger Verlagshandlung GmbH., München.

Amrith Krishna, Vishnu Dutt Sharma, Bishal Santra, Aishik Chakraborty, Pavankumar Satuluri, and Pawan Goyal. 2019. Poetry to prose conversion in sanskrit as a linearisation task: A case for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1160–1166.

Jonas Kuhn. 2020. *Einleitung. Reflektierte algorithmische Textanalyse: Interdisziplinäre (s) Arbeiten in der CRETA-Werkstatt*. Walter de Gruyter GmbH & Co KG.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

William Labov. 1986. The social stratification of (r) in new york city department stores. In *Dialect and language variation*, pages 304–329. Elsevier.

William Labov. 2019. *The study of language in its social context*. De Gruyter Mouton.

Veronika Laippala, Jesse Egbert, Douglas Biber, and Aki-Juhani Kyröläinen. 2021. Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Language Resources and Evaluation*, pages 1–32.

Annie Lamar and America Chambers. 2019. Generating homeric poetry with deep neural networks. In *2019 First International Conference on Transdisciplinary AI (TransAI)*, pages 68–75. IEEE.

Harm Lameris and Sara Stymne. 2021. Whit's the richt pairt o speech: Pos tagging for scots. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 39–48.

Dieter Lamping. 2016. *Handbuch Lyrik: Theorie, Analyse, Geschichte*. Springer-Verlag.

Jey Han Lau, Trevor Cohn, Timothy Baldwin, Julian Brooke, and Adam Hammond. 2018. Deep-speare: A joint neural model of poetic language, meter and rhyme. *arXiv preprint arXiv:1807.03491*.

Heinrich Lausberg, David E Orton, and R Dean Anderson. 1998. *Handbook of literary rhetoric: A foundation for literary study*. Brill.

Xuan Le, Ian Lancashire, Graeme Hirst, and Regina Jokel. 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists. *Literary and linguistic computing*, 26(4):435–461.

David YW Lee. 2001. Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the bnc jungle.

Shin Haeng Lee, Jin-Young Tak, Eun Joo Kwak, Seonghoon Kim, and Tae Yun Lim. 2021. Revisiting sylvia plath's and anne sexton's confessional poetry: Analyzing stylistic differences and evolution of poetic voice (s) through computational text analysis. *Digital Scholarship in the Humanities*.

Geoffrey Leech. 2008. Language in literature. *Style and Foregrounding. Routledge*.

Geoffrey Leech and Mick Short. 2007. Style in fiction: A linguistic introduction to english prose fiction.

Geoffrey N Leech. 2014. *A linguistic guide to English poetry*, volume 4. Routledge.

Willem JM Levelt. 1993. *Speaking: From intention to articulation*, volume 1. MIT press.

Franklin Mark Liang. 1983. *Word Hy-phen-a-tion by Com-put-er*. Citeseer.

Johann-Mattis List, Jananan Sylvestre Pathmanathan, Nathan W Hill, Eric Bapteste, and Philippe Lopez. 2017. Vowel purity and rhyme evidence in old chinese reconstruction. *Lingua Sinica*, 3(1):1–17.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019. DENS: A dataset for multi-class emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6294–6299, Hong Kong, China. Association for Computational Linguistics.

Michaela Mahlberg. 2013. *Corpus stylistics and Dickens's fiction*. Routledge.

Enrique Manjavacas, Folgert Karsdorp, Ben Burtenshaw, and Mike Kestemont. 2017. Synthetic literature: Writing science fiction in a co-creative process. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*, pages 29–37.

Joshua J. Mark. 2009. Literature. https://www.worldhistory.org/literature/.

Nina McCurdy, Julie Lein, Katharine Coles, and Miriah Meyer. 2015a. Poemage: Visualizing the sonic topology of a poem. *IEEE transactions on visualization and computer graphics*, 22(1):439–448.

Nina McCurdy, Vivek Srikumar, and Miriah Meyer. 2015b. Rhymedesign: A tool for analyzing sonic devices in poetry. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 12–22.

Dan McIntyre and Brian Walker. 2010. How can corpora be used to explore the language of poetry and drama? In *The Routledge handbook of corpus linguistics*, pages 516–530. Routledge.

Daniel McIntyre and Brian Walker. 2019. *Corpus stylistics: Theory and practice*. Edinburgh University Press.

Katja Mellmann. 2007. Versanalyse. *Handbuch Literaturwissenschaft 2: Methoden und Theorien*, pages 81–97.

Thomas Corwin Mendenhall. 1887. The characteristic curves of composition. *Science*, 9(214):237–249.

Winfried Menninghaus, Valentin Wagner, Julian Hanich, Eugen Wassiliwizky, Milena Kuehnast, and Thomas Jacobsen. 2015. Towards a psychological construct of being moved. *PloS one*, 10(6):e0128451.

Winfried Menninghaus, Valentin Wagner, Christine A Knoop, and Mathias Scharinger. 2018. Poetic speech melody: A crucial link between music and language. *PloS one*, 13(11):e0205980.

Winfried Menninghaus, Valentin Wagner, Eugen Wassiliwizky, Thomas Jacobsen, and Christine A Knoop. 2017. The emotional and aesthetic powers of parallelistic diction. *Poetics*, 63:47–59.

Winfried Menninghaus, Valentin Wagner, Eugen Wassiliwizky, Ines Schindler, Julian Hanich, Thomas Jacobsen, and Stefan Koelsch. 2019. What are aesthetic emotions? *Psychological review*, 126(2):171.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 259–263.

Smitha Milli and David Bamman. 2016. Beyond canonical texts: A computational analysis of fanfiction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2048–2053.

Yoonsook Mo, Jennifer Cole, and Eun-Kyung Lee. 2008. Naïve listeners' prominence and boundary perception. *Proc. Speech Prosody, Campinas, Brazil*, pages 735–738.

Mohammad Mobasher and Saeed Farzi. 2021. An enhanced personality detection system through user's digital footprints. *Digital Scholarship in the Humanities*.

Saif Mohammad. 2011a. Colourful language: Measuring word-colour associations. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 97–106.

Saif Mohammad. 2011b. Even the abstract have color: Consensus in word-colour associations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–373.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier.

Saif M Mohammad and Peter D Turney. 2013a. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Saif M Mohammad and Peter D Turney. 2013b. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Gaurav Mohanty, Pruthwik Mishra, and Radhika Mamidi. 2018. Kabithaa: An annotated corpus of odia poems with sentiment polarity information. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Franco Moretti. 2005. *Graphs, maps, trees: abstract models for a literary history*. Verso.

Franco Moretti. 2009. Style, inc. reflections on seven thousand titles (british novels, 1740–1850). *Critical Inquiry*, 36(1):134–158.

Franco Moretti. 2011. Network theory, plot analysis. literary lab pamphlet 2.

Franco Moretti. 2013. *Distant reading*. Verso Books.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

David Nadeau, Catherine Sabourin, Joseph De Koninck, Stan Matwin, Peter D Turney, et al. 2006. Automatic dream sentiment analysis. In *Proceedings of the Workshop on Computational Aesthetics at the Twenty-First National Conference on Artificial Intelligence*.

Borja Navarro, María Ribes Lafoz, and Noelia Sánchez. 2016. Metrical annotation of a large corpus of spanish sonnets: representation, scansion and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4360–4364.

Borja Navarro-Colorado. 2018a. A metrical scansion system for fixed-metre spanish poetry. *Digital Scholarship in the Humanities*, 33(1):112–127.

Borja Navarro-Colorado. 2018b. On poetic topic modeling: extracting themes and motifs from a corpus of spanish poetry. *Frontiers in Digital Humanities*, 5:15.

Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. 2016. Learning text similarity with siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157.

Ani Nenkova, Jason Brenier, Anubha Kothari, Sasha Calhoun, Laura Whitton, David Beaver, and Dan Jurafsky. 2007. To memorize or to predict: Prominence labeling in conversational speech. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 9–16.

Christian Obermeier, Winfried Menninghaus, Martin Von Koppenfels, Tim Raettig, Maren Schmidt-Kassow, Sascha Otterbein, and Sonja AE Kotz. 2013. Aesthetic and emotional effects of meter and rhyme in poetry. *Frontiers in psychology*, 4:10.

Martin Opitz. 2020. *Buch von der deutschen Poeterey*. Good Press.

Myle Ott, Claire Cardie, and Jeffrey T Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 497–501.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319.

Allison Parrish. 2018. A Gutenberg Poetry Corpus.

Rebecca J Passonneau, Nancy Ide, Songqiao Su, and Jesse Stuart. 2014. Biber redux: Reconsidering dimensions of variation in american english. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 565–576.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Stephen Phillips. 2016. Can big data find the next 'harry potter'?, the atlantic.

Axel Pichler and Nils Reiter. 2020. Reflektierte textanalyse. In *Reflektierte Algorithmische Textanalyse*, pages 43–60. De Gruyter.

Andrew Piper. 2017. Think small: on literary modeling. *PMLA/Publications of the Modern Language Association of America*, 132(3):651–658.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98.

Petr Plecháč. 2020. Relative contributions of shakespeare and fletcher in henry viii: An analysis based on most frequent words and most frequent rhythmic patterns. *Oxford Journal of Digital Humanities (DSH), arXiv preprint arXiv:1911.05652*.

Robert Plutchik. 1991. *The Emotions*. University Press of America.

RL Victoria Pöhls. 2020. Engaging literature. In *Politische Emotionen in den Künsten*, pages 219–240. De Gruyter.

Octavian Popescu and Carlo Strapparava. 2013. Behind the times: Detecting epoch changes using large corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 347–355.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Beatrice Primus. 2011. *Unreine Reime und phonologische Theorie*. De Gruyter Mouton.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as*

Corpus Workshop (WAC-X) and the EmpiriST Shared Task, pages 57–62, Berlin. Association for Computational Linguistics (ACL).

Vladimir Propp. 2010. *Morphology of the Folktale*, volume 9. University of Texas Press.

Christoph Purschke and Dirk Hovy. 2019. Lörres, möppes, and the swiss.(re) discovering regional patterns in anonymous social media data. *Journal of Linguistic Geography*, 7(2):113–134.

Marcus Fabius Quintilian. 1924. The institutio oratoria of quintilian, trans. he butler.

Taraka Rama. 2016. Siamese convolutional networks for cognate identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1018–1027.

Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. 2008. The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In *ICWSM*.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Sravana Reddy and Kevin Knight. 2011. Unsupervised discovery of rhyme schemes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 77–82.

Radim Rehurek and Petr Sojka. 2011. Gensim—statistical semantics in python. *statistical semantics; gensim; Python; LDA; SVD*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ines Reinig and Ines Rehbein. 2019. Metaphor detection for german poetry. Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019).

Nils Reiter, Benjamin Krautter, Janis Pagel, and Marcus Willand. 2018. Detecting protagonists in german plays around 1800 as a classification task.

Phil Roberts. 2000. *How Poetry Works*. Penguin UK.

Thomas Rommel. 2004. Literary studies. *A companion to digital humanities*, pages 88–96.

Andrew Rosenberg. 2010. Autobi-a tool for automatic tobi annotation. In *Eleventh Annual Conference of the International Speech Communication Association*.

Ina Rösiger and Arndt Riester. 2015. Using prosodic annotations to improve coreference resolution of spoken text. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 83–88.

David Rother, Thomas Haider, and Steffen Eger. 2020. CMCE at SemEval-2020 task 1: Clustering on manifolds of contextualized embeddings to detect historical meaning shifts. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 187–193, Barcelona (online). International Committee for Computational Linguistics.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Pablo Ruiz, Clara Martínez Cantón, Thierry Poibeau, and Elena González-Blanco. 2017. 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*, pages 27–32.

Pablo Ruiz Fabo, Helena Bermúdez Sabel, Clara Martínez Cantón, and Elena González-Blanco. 2020. The diachronic spanish sonnet corpus: Tei and linked open data encoding, data distribution, and metrical findings. *Digital Scholarship in the Humanities*.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111.

Marina Santini. 2005. Linguistic facets for genre and text type identification: A description of linguistically-motivated features. *ITRI report series: ITRI-05*, 2.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Ines Schindler, Georg Hosoya, Winfried Menninghaus, Ursula Beermann, Valentin Wagner, Michael Eid, and Klaus R Scherer. 2017a. Measuring aesthetic emotions: A review of the literature and a new assessment tool. *PloS one*, 12(6):e0178899.

Ines Schindler, Georg Hosoya, Winfried Menninghaus, Ursula Beermann, Valentin Wagner, Michael Eid, and Klaus R. Scherer. 2017b. Measuring aesthetic emotions: A review of the literature and a new assessment tool. *PLOS ONE*, 12(6):1–45.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation

of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Denise Schmandt-Besserat. 2015. Evolution of writing. In James D. Wright, editor, *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, second edition edition, pages 761–766. Elsevier, Oxford.

Helmut Schmid. 1994. Treetagger-a part-of-speech tagger for many languages. *Ludwig-Maximilians-Universität Munich*.

Christof Schöch. 2016. Principal component analysis for literary genre stylistics. *The Dragonfly's Gaze*.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41, New Orleans, Louisiana. Association for Computational Linguistics.

Sarah Schulz and Jonas Kuhn. 2016. Learning from within? comparing pos tagging approaches for historical text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4316–4322.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Oscar Samuel Schwartz. 2017. *A History of Computational Poetics: Agency, Automation and the Post-Human*. Ph.D. thesis, Monash University.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. *MLMTA*, 2003:273.

Artjoms Šeļa, Petr Plecháč, and Alie Lassche. 2022. Semantics of european poetry is shaped by conservative forces: The relationship between poetic meter and meaning in accentual-syllabic verse. *Plos one*, 17(4):e0266556.

Elena Semino and Mick Short. 2004. *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*. Routledge.

Serge Sharoff. 2018. Functional text dimensions for the annotation of web corpora. *Corpora*, 13(1):65–95.

Lucius Adelno Sherman. 1892. On certain facts and principles in the development of form in literature. *Papers from the University Studies series (The University of Nebraska)*, page 8.

Victor Shlovsky. 1965. Art as technique. russian formalist criticism: Four essays. *Trans. Lee T. Lemon and Marion J. Reis. Lincoln and London: University of Nebraska Press.*

Philippa Shoemark, Debnil Sur, Luke Shrimpton, Iain Murray, and Sharon Goldwater. 2017. Aye or naw, whit dae ye hink? scottish independence and linguistic identity on social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1239–1248.

H. S. Sichel. 1975. On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351):542–547.

Paul Simpson. 2004. *Stylistics: A Resource Book for Students.* Routledge, London, UK.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634.

Carlota S Smith. 2003. *Modes of discourse: The local structure of texts*, volume 103. Cambridge University Press.

Stanisław Śnieżewski. 2014. Elocutio in quintilian's institutio oratoria, book viii. *Classica Cracoviensia*, (17):203–230.

Morgan Sonderegger. 2011. Applications of graph theory to an english rhyming corpus. *Computer Speech & Language*, 25(3):655–678.

S. P. Sreeja and G. S. Mahalakshmi. 2019. Perc-an emotion recognition corpus for cognitive poems. In *2019 International Conference on Communication and Signal Processing (ICCSP)*, pages 0200–0207.

Phil Stamper-Halpin. 2019. Chow prh uses artificial intelligence (ai) to extend your book's reach.

T Steele. 2012. Verse and prose. *The Princeton encyclopedia of poetry and poetics*, pages 1507–1513.

Shannon Wiltsey Stirman and James W Pennebaker. 2001. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine*, 63(4):517–522.

Jana Straková, Milan Straka, and Jan Hajic. 2014. Open-source tools for morphology, lemmatization, pos tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74. Association for Computational Linguistics.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA).

Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann, and Martti Vainio. 2019. Predicting prosodic prominence from text with pre-trained contextualized word representations. *arXiv preprint arXiv:1908.02262*.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Dennis Tenen. 2017. *COMPUTATIONAL POETICS. An Introduction*, pages 1–22. Stanford University Press.

Peer Trilcke, Christopher Kittel, Nils Reiter, Daria Maximova, and Frank Fischer. 2020. Opening the stage–a quantitative look at stage directions in german drama. In *DH*.

Gabriel Stephen Trop. 2010. *Aesthetic Exercises and Poetic Form in the Works of Hölderlin, Novalis, and Rococo Poets*. University of California, Berkeley.

Christos Tsagalis. 2009. Poetry and poetics in the hesiodic corpus. *Brill's Companion to Hesiod*, pages 131–78.

Reuven Tsur. 2008. *Toward a Theory of Cognitive Poetics: Second*. Liverpool University Press.

Herbert F Tucker. 2011. Poetic data and the news from poems: A" for better for verse" memoir. *Victorian Poetry*, 49(2):267–281.

Ted Underwood. 2019. *Distant horizons: digital evidence and literary change*. University of Chicago Press.

Ted Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in english-language fiction. *Journal of Cultural Analytics*, 3(2):11035.

Ted Underwood and Jordan Sellers. 2012. The emergence of literary diction. *The Journal of Digital Humanities, 1(2), Online; accessed 16-February-2021;*, pages http://journalofdigitalhumanities.org/1–2/the–emergence–of–literary–diction–by–ted–underwood–and–jordan–sellers/.

Rafael Valle, Kevin Shih, Ryan Prenger, and Bryan Catanzaro. 2020. Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis. *arXiv preprint arXiv:2005.05957*.

Bertie Vidgen, Austin Botelho, David Broniatowski, Ella Guest, Matthew Hall, Helen Margetts, Rebekah Tromble, Zeerak Waseem, and Scott Hale.

2020. Detecting east asian prejudice on social media. *arXiv preprint arXiv:2005.03909*.

Katie Wales. 2014. *A Dictionary of Stylistics*. Routledge.

Chong Wang, David Blei, and David Heckerman. 2012. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Eugen Wassiliwizky, Thomas Jacobsen, Jan Heinrich, Manuel Schneiderbauer, and Winfried Menninghaus. 2017. Tears falling on goosebumps: Co-occurrence of emotional lacrimation and emotional piloerection indicates a psychophysiological climax in emotional arousal. *Frontiers in Psychology*, 8:41.

Linda R Waugh. 1980. The poetic function in the theory of roman jakobson. *Poetics today*, 2(1a):57–82.

Swantje Westpfahl. 2014. STTS 2.0? improving the tagset for the part-of-speech-tagging of German spoken data. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 1–10, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Swantje Westpfahl and Thomas Schmidt. 2016. Folk-gold a gold standard for part-of-speech-tagging of spoken german.

Christin Michelle Laroche Wilson. 2012. *Variation and Text Type in Old Occitan Texts*. The Ohio State University.

Simone Winko. 2009. Auf der Suche nach der Weltformel. In *Grenzen der Literatur: Zu Begriff und Phänomen des Literarischen*, volume 2 of *Revisionen*, pages 374–396. De Gruyter, Berlin, Germany.

Jörg Wöckener, Thomas Haider, Tristan Miller, The-Khang Nguyen, Thanh Tung Linh Nguyen, Minh Vu Pham, Jonas Belouadi, and Steffen Eger. 2021. End-to-end style-conditioned poetry generation: What does it take to learn from examples alone? In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 57–66, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Richard Xiao. 2009. Multidimensional analysis and the study of world englishes. *World Englishes*, 28(4):421–450.

Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *CogSci*. cognitivesciencesociety.org.

C. Yang, K. H. Lin, and H. Chen. 2009. Writer meets reader: Emotion analysis of social media from both the writer's and reader's perspectives. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 287–290.

Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1128–1137.

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680.

Richard Zimmermann. 2015. The parsed corpus of middle english poetry. *Published online at http://www. pcmep. net.*

George Kingsley Zipf. 2016. *Human behavior and the principle of least effort: An introduction to human ecology.* Ravenio Books.