A Large Annotated Reference Corpus of New High German Poetry

Thomas Haider
University of Passau
Max Planck Institute for Empirical Aesthetics, Frankfurt am Main
thomas.haider@uni-passau.de

Abstract

This paper introduces a large annotated corpus of public domain German poetry, covering the time period from 1600 to the 1920s with 65k poems. We describe how the corpus was compiled, how it was cleaned (including duplicate detection), and how it looks now in terms of size, format, temporal distribution, and automatic annotation. Besides metadata, the corpus contains reliable annotation of tokens, syllables, part-of-speech, and meter and verse measure. Finally, we give some statistics on the annotation and an overview of other poetry corpora.

Keywords: Poetry, Corpus, German, Large, Reference

1. Introduction

This paper presents a large, comprehensive, automatically annotated, and easily searchable corpus of New High German (NHG) poetry. It was built by collecting and parsing the largest digitized corpora that contain public domain German literature. The resources were cleaned, standardized, merged, and automatically annotated large scale with the tools developed by Haider (2021). Since this corpus contains the majority of digitized public domain poetry from the New High German period, we call it *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

This corpus contains the poetry from the German Text Archive (Deutsches Textarchiv: DTA)[1](texts with label 'Lyrik') and also the Digital Library of Textgrid (TGRID) (texts with label 'verse').[2] DTA was originally mined from *wikimedia commons* and Textgrid was mined from *zeno.org*. Neither DTA (Geyken et al., 2011) nor the Textgrid library (Neuroth et al., 2015) aim to provide a comprehensive corpus of German writing. However, these are the only available large and high quality text corpora that contain public domain NHG poetry.[3]

Table 1 lists size statistics of these respective corpora and the final merged DLK corpus. In total, the corpus provides over 65k unique poems and over 1.6M lines, each tokenized, syllabified, pos-tagged, and meter-tagged. We find 254 authors (where women are unsurprisingly underrepresented). Overall, there is more material to be found in these corpora, but e.g., any of the 10k line groups in DTA that do not fall under the genre label 'Lyrik' are problematic regarding their metadata (e.g., authorship), since they were published out of context (i.a. for criticism).

	TGRID	DTA	DLK v6
#syllables	12,012,846	4,421,923	15,196,215
#words	8,024,763	2,986,912	10,162,011
#tokens	9,673,124	3,549,224	12,201,712
#lines	1,320,779	458,851	1,643,076
#stanzas	205,275	63,080	246,677
#poems	50,549	22,039	65,755
#authors	227	73	254

Table 1: Sub-Corpora of the German Poetry Corpus by Size ('tokens' with punctuation, 'words' without)

Given its size and temporal distribution (cf. section 2.2), DLK is intended as a reference corpus for NHG poetry that can act as a point of comparison and a collection to sample from. In the future, we hope to extend the data. At the current stage of digitization it is hard to address questions regarding representativeness and the canon. Since we do not (yet) know what we are missing (which poetry has not been digitized and standardized) and how important the missing poetry is to the canon, we have to assume that what we do have covers the (most important work in the) canon, and further research can sample from the reference corpus according to specific research questions.

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).

[1] http://deutschestextarchiv.de

[2] http://textgrid.de

[3] We also crawled the German version of Project Gutenberg (GUT-DE; https://www.projekt-gutenberg.org/). We omit this corpus from our collection, because licensing is unclear and the corpus is fairly inconsistent regarding markup and document structure, offering metadata (publication date, etc.) for only around 1/3 of its poems. In total, GUT-DE contains 36k poems.

LREC-COLING 2024, pages 677–683
20-25 May, 2024.

Such a corpus aims to be representative of specific variables, while a reference corpus aims to be balanced and comprehensive. Both are important in distant reading research (Jockers, 2013; Underwood, 2019) because they help ensure that 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. Corpus representativeness is considered here by aiming at a considerable size of the corpus and ensuring that every time period is adequately represented.

The paper is structured as follows: First, we provide information on how we built the corpus, including cleaning, standardizing, and structuring it in different formats. We then give an overview of the temporal distribution of the corpus and how duplicates were detected when merging the two resources. We then we give an overview of the annotation layers and some statistics to get an overview of the different annotated features. Finally, we offer an overview of other poetry corpora.

2. Building the Corpus

2.1. Cleaning and Formats

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.[4] We use langdetect[5] 1.0.8 on first lines to tag every poem with its language to filter out any poems that are not German (such as Latin or French).

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, often with adequate information, sometimes without it. 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 (manual reading and heuristics), 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. Identifying and fixing these kinds of issues is an ongoing effort.

Our resources are designed in a standardized format to sustainably and interoperably archive poetry in both .json and TEI P5 XML. 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. See Figure 1 for an example from the .json corpus. We can see the header information of the poem and one annotated line. The annotation layers are discussed below.

```
"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"
                },
    [...]
```

Figure 1: A Poem Snippet with Meter Annotation from DLK in .json

Our framework for TEI is grounded in the so-called DTA-Basisformat[6] (Haaf et al., 2014), that provides a "Base Format", which constrains the data to TEI P5 guidelines, and a relaxNG schema.

2.2. Temporal Distribution

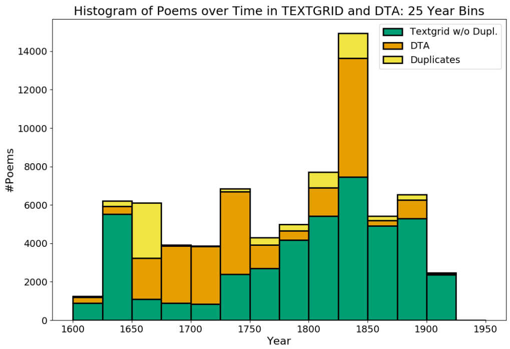

Figure 2: DTA and Textgrid Poems in 25 Year Bins. Identified duplicates are subtracted from Textgrid.

[4] 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.

[5] https://pypi.org/project/langdetect/

[6] http://www.deutschestextarchiv.de/doku/basisformat/

An important factor to consider when compiling a diachronic corpus is the temporal distribution of texts by their (publication) date. In the Textgrid source, if a distinct date of origin was identified,[7] it was tagged as publication date, if not, through *notBefore* and *notAfter* (author birth/death), of which we then took the average. Figure 2 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.

2.3. Duplicates

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. 1) between the unigrams (word forms) of two poems A and B to measure their overlap.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

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. 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 identified more than 7600 poems as duplicates (as is reflected in Table 1).

3. Annotation and Statistics

A .json poem is organized as python dictionary. Every poem has a unique index as key, some metadata (author, title, publication year, etc., see Figure 1), but also 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 respective tools were developed and are documented in previous work (Haider, 2021).

Tokenization is performed with SoMaJo (Proisl and Uhrig, 2016), with a more conservative handling of apostrophes (to leave words with elided vowels intact). The STTS tagset is used for part-of-speech tagging. The tagger was trained and evaluated for poetry and is high performing. Syllabification is carried out with an ensemble trained on CELEX with Hyphe-NN and a biLSTM-CRF.

Predicting the meter is carried out with a biLSTM-CRF on top of syllable embeddings with a multi-task-objective developed by Haider (2021), who reported a line accuracy of 88% and a syllable accuracy of 97%. To determine the measure of a line from its raw metrical annotation, we developed a set of regular expressions. We orient ourselves with the handbook of Knörrich (1971). The 'verse measure' 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. 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`,[8] the `dactylic hexameter`,[9] conventionally known as 'hexameter', and some ode forms like the `asklepiadic` verse (+−+−++−+−+).[10]

[7] https://textgrid.de/en/digitale-bibliothek

[8] Alexandrine: −+−+−+−+−+−+−?
The symbol before ? is optional

[9] Hexameter: +−?+−?+−?+−?+−+−

[10] Also see Haider (2021)

3.1. Statistics

3.1.1. Length of Poems and Lines

It is noteworthy that DTA poems are considerably shorter than Textgrid poems. As seen in Table 1, DTA contains about half the amount of poems compared to Textgrid (22k vs. 50k), but these amount to less than 40% in terms of total number of words (3M vs. 8M).

Corpus	Median Length	Mean Length
DTA	54	136
TGRID	87	164

Table 2: Median and Mean Word Length of Poems.

Table 2 shows the median and mean length of poems in the respective corpora. As expected, we find that DTA texts are overall shorter. For a point of comparison, 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 (cf. also Table 1), 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.

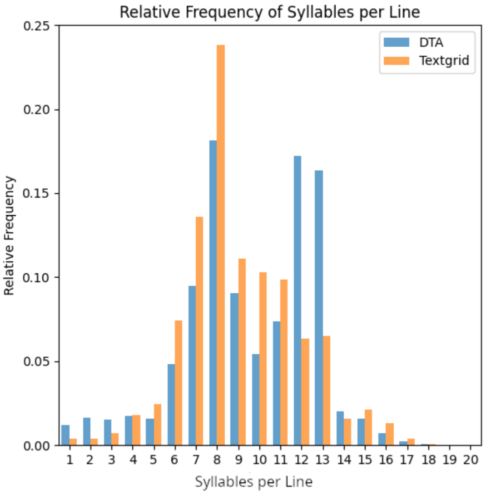

Figure 3: DTA and Textgrid Relative Frequency of Syllables in Line.

In contrast, the stylistic differences in verse length across both corpora can be seen in Figure 3, which shows the relative frequency of line length over corpora. 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.

3.1.2. Meter and Measure

Meter refers to the raw syllable annotation (+−+−+−) and measure refers to the derived label (e.g., trochaic.trimeter) as elaborated in section 3. Table 3 lists the most frequent measure labels in DLK (simplified to only show the foot type and the length). We find that the majority of lines is iambic, followed by the alexandrine and trochaic.tetrameter. The combination iambic.tetrameters and iambic.trimeters indicates the German stanza form ‘Volksliedstrophe’, which, according to Frank (1980) is among the most frequent stanza forms in German. Overall, around only 1% of lines are written in hexameter (cf. Klopstock), and for about the same proportion no meaningful measure assignment (unknown.measure) was possible (e.g., through tagging error).

Measure	Abs. Freq.	Rel. Freq.
iambic.tetra	371209	0.2259
alexandrine.iambic	236911	0.1442
iambic.penta	228454	0.1390
iambic.tri	207086	0.1260
trochaic.tetra	202300	0.1231
trochaic.penta	45216	0.0275
iambic.di	38807	0.0236
iambic.hexa	33891	0.0206
trochaic.tri	27970	0.0170
amphibrach.tri	19234	0.0117
unknown.measure	18796	0.0114
hexameter	15825	0.0096
anapaest.di	15807	0.0096
trochaic.di	15132	0.0092
amphibrach.di	14888	0.0091
trochaic.hexa	14729	0.0090
iambic.septa	11627	0.0071
trochaic.septa	10768	0.0066
amphibrach.tetra	10033	0.0061
dactylic.di	9787	0.0060
trochaic.octa	9712	0.0059
dactylic.tri	7958	0.0048
Rest	84026	0.0479

Table 3: Frequency of Measures

The temporal distribution of these verse measures is shown in Figure 4, here constrained to only show the measure of the first line of a poem to eliminate length effects. We can see that the (strict) alexandrine is the dominant verse form in pre-romantic times (before 1750), but it loses importance in later times. We see it present also in the time slot 1800–1850. Inspection of the corpus revealed that the 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. We also see that iambic.tetrameter, trochaic.tetrameter and iambic.pentameter have enjoyed continuous popularity over the whole time span.

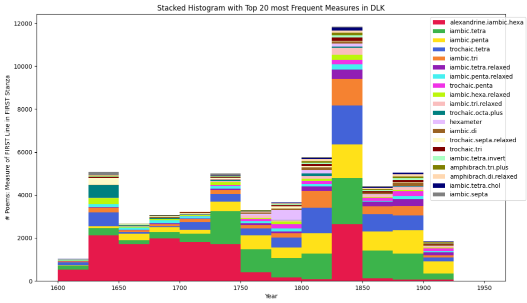

Figure 4: DLK Verse Measures over Time 1st Lines.

Tagging verse measure large scale also allows us to characterize authors by their preferred verse forms. In Figure 5 we can see that Klopstock wrote in the epic verse form of hexameter. As seen for example in Figure 6 (Heine) the most popular measures are the iambic and trochaic.

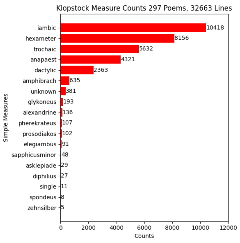

Figure 5: Measures of Klopstock: Troch., Hexam.

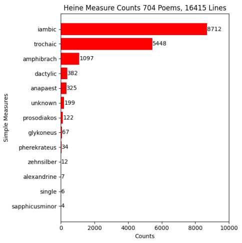

Figure 6: Measures of Heine, Iamb., Troch.

4. Related Work

Poetry corpora of varying sizes exist for a a number of languages. Larger collections of poetry are useful for large scale analysis, while small 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).

Regarding larger collections (10k–70k), we are aware of the following: The Corpus of Czech Verse (Plecháč and Kolár, 2015), the Corpus of Spanish Golden Age Sonnets (Navarro-Colorado, 2018) and the Diachronic Spanish Sonnet Corpus (Ruiz Fabo et al., 2020), the French Corpus Malherbe (Delente and Renault, 2021), the Hungarian ELTE Poetry Corpus (Horváth et al., 2022), the Italian Biblioteca italiana (bib, 2023), Portuguese Poemas (Mittmann et al., 2019), and the Corpus of Russian Poetry (Šeļa et al., 2020).

For English, most work is based on the collection of the Project Gutenberg. Parrish (2018) 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. Haider (2021) compiled a corpus from Project Gutenberg, but was quite restrictive in deleting duplicates. Underwood and Sellers (2012) released a corpus that also includes 18th and 19th century poetry.[11]

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), Old Greek (Tsagalis, 2009; Lamar and Chambers, 2019), or Chinese Tang poetry Zhang and Lapata (2014). For English, German and French rhyming corpora are available (Reddy and Knight, 2011; Sonderegger, 2011; Haider and Kuhn, 2018), alongside proposed methods to detect rhymes automatically.

Regarding rhythmic patterns, Agirrezabal et al. (2016a,b, 2019) used an English corpus, originally compiled by (Tucker, 2011). Within a project of similar scope, Anttila and Heuser (2016) manually annotated for meter and feet, according to Hanson and Kiparsky (1996). The Spanish corpora (Ruiz Fabo et al., 2020; Navarro et al., 2016; Navarro-Colorado, 2018) are also annotated for rhythm/meter 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).

Lastly, a few poetry corpora are also annotated for emotions, as discussed in Haider et al. (2020).

5. Conclusion

This paper has introduced a large annotated corpus of New High German poetry. We have shown how we built it, including cleaning, detection of duplicates, and the layers of annotation with illustratory statistics. We hope this corpus will be useful to the community. The corpus is freely available and aims to follow the FAIR principles (which will be adressed in the final version). Some work was already done on top of the corpus, e.g., (Belouadi and Eger, 2022), and it was picked up for a website to visualize the annotation layers and more.[12] Long term, we plan an API for it, according to the principles of 'programmable corpora'.

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

[12] lyrikkompass.de

6. Bibliographical References

2023. Biblioteca italiana. http://www.bibliotecaitaliana.it/.

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.

Arto Anttila and Ryan Heuser. 2016. Phonological and metrical variation across genres. In *Proceedings of the Annual Meetings on Phonology*, volume 3.

Jonas Belouadi and Steffen Eger. 2022. Bygpt5: End-to-end style-conditioned poetry generation with token-free language models. *arXiv preprint arXiv:2212.10474*.

Eliane Delente and Richard Renault. 2021. Projet anamètre: présentation, limites et avancées.

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.

Horst Joachim Frank. 1980. *Handbuch der deutschen Strophenformen*. Hanser.

Bethany Gray, Jesse Egbert, and Douglas Biber. 2017. Exploring methods for evaluating corpus representativeness. In *Corpus Linguistics International Conference*.

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, Steffen Eger, Evgeny Kim, Roman Klinger, and Winfried Menninghaus. 2020. Po-emo: Conceptualization, annotation, and modeling 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.

Kristin Hanson and Paul Kiparsky. 1996. A parametric theory of poetic meter. *Language*, pages 287–335.

Péter Horváth, Péter Kundráth, Balázs Indig, Zsófia Fellegi, Eszter Szlávich, Tímea Borbála Bajzát, Zsófia Sárközi-Lindner, Bence Vida, Aslihan Karabulut, Mária Timári, et al. 2022. Elte poetry corpus: A machine annotated database of canonical hungarian poetry. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3471–3478.

Arthur M Jacobs. 2018. The gutenberg english poetry corpus: exemplary quantitative narrative analyses. *Frontiers in Digital Humanities*, 5:5.

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

Otto Knörrich. 1971. *Die deutsche Lyrik der Gegenwart*, volume 401. Kröner.

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.

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.

Adiel Mittmann, Paulo Henrique Pergher, and Alckmar Luiz dos Santos. 2019. What rhythmic signature says about poetic corpora. *Quantitative approaches to versification*, page 153.

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. 2018. A metrical scansion system for fixed-metre spanish poetry. *Digital Scholarship in the Humanities*, 33(1):112–127.

Allison Parrish. 2018. A Gutenberg Poetry Corpus.

Petr Plecháč and Robert Kolár. 2015. The corpus of czech verse. *Studia metrica et poetica*, 2(1):107–118.

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).

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.

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*.

Artjoms Šeļa, Boris Orekhov, and Roman Leibov. 2020. Weak genres: Modeling association between poetic meter and meaning in russian poetry. *Proceedings http://ceur-ws. org ISSN*, 1613:0073.

Morgan Sonderegger. 2011. Applications of graph theory to an english rhyming corpus. *Computer Speech & Language*, 25(3):655–678.

Christos Tsagalis. 2009. Poetry and poetics in the hesiodic corpus. *Brill's Companion to Hesiod*, pages 131–78.

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 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/.

Christin Michelle Laroche Wilson. 2012. *Variation and Text Type in Old Occitan Texts*. The Ohio State University.

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*.

7. Language Resource References

Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas, and Frank Wiegand. 2011. Das deutsche textarchiv: Vom historischen korpus zum aktiven archiv. *Digitale Wissenschaft*, 157.

Heike Neuroth, Andrea Rapp, and Sibylle Söring. 2015. Textgrid: Von der community–für die community.