BRILL

RESEARCH DATA JOURNAL FOR THE HUMANITIES AND SOCIAL SCIENCES 9 (2024) 1–17

BRILL OPEN

brill.com/rdj

PoeTree: Poetry Treebanks in Czech, English, French, German, Hungarian, Italian, Portuguese, Russian, Slovenian and Spanish

Petr Plecháč | ORCID: 0000-0002-1003-4541
Corresponding author
Institute of Czech Literature, Czech Academy of Sciences, Prague, Czechia
plechac@ucl.cas.cz

Silvie Cinková | ORCID: 0000-0003-4526-3915
Institute of Czech Literature, Czech Academy of Sciences, Prague, Czechia
Charles University, Prague, Czechia
cinkova@ucl.cas.cz

Robert Kolár | ORCID: 0000-0001-8061-1917
Institute of Czech Literature, Czech Academy of Sciences, Prague, Czechia
kolar@ucl.cas.cz

Artjoms Šeļa | ORCID: 0000-0002-2272-2077
Institute of Polish Language, Polish Academy of Sciences, Warsaw, Poland
artjoms.sela@ijp.pan.pl

Mirella De Sisto | ORCID: 0000-0002-0899-5976
Tilburg University, Tilburg, the Netherlands
m.desisto@tilburguniversity.edu

Lara Nugues | ORCID: 0000-0003-1381-8090
University of Basel, Basel, Switzerland
lara.nugues@unibas.ch

Thomas Haider | ORCID: 0000-0003-1522-4026
University of Passau, Passau, Germany
thomas.haider@uni-passau.de

Published with license by Koninklijke Brill BV | DOI:10.1163/24523666-bja10044
 | ISSN: 2452-3666 (online)

Neža Kočnik | ORCID: 0009-0003-8318-2179
University of Ljubljana, Ljubljana, Slovenia
neza.kocnik@gmail.com

Received 19 December 2023 | Revised 29 April 2024 |
Accepted 21 June 2024 | Published online 12 September 2024

Abstract

This article presents a set of standardised corpora of poetry comprising over 330,000 poems in ten languages (Czech, English, French, German, Hungarian, Italian, Portuguese, Russian, Slovenian, and Spanish). Each corpus has been deduplicated, enriched with Universal Dependencies, provided with additional metadata, and converted into a unified JSON structure.

Keywords

poetry – computational poetics – corpus linguistics – digital humanities

- Related data set "PoeTree. Poetry Treebanks in Czech, English, French, German, Hungarian, Italian, Portuguese, Russian, Slovenian and Spanish" with DOI www.doi.org/10.5281/zenodo.10907309 in repository "Zenodo"
- Also access the data through the REST API (as of April 2024): https://versologie.cz/poetree/api_doc
- Also access the data through the Python library (as of April 2024): www.pypi.org/project/poetree
- Also access the data through the R library (as of April 2024): www.github.com/perechen/poetRee

1. Introduction

With advances in computational literary studies, the demand for open multilingual datasets has been increasing, be it for the purpose of comparative literary research (Storey & Mimno, 2020; Šeļa et al., 2022), as a benchmark for new stylometric methods (Du et al., 2022; Plecháč, 2021), or as training data

for multi-lingual models that aim to enhance literary text annotation and processing pipelines (Bamman, 2021; Byszuk et al., 2020; de la Rosa, 2023). Several relevant resources are already available for prose fiction, including the European Literary Text Collection or ELTeC (Odebrecht et al., 2021) and benchmark corpora built by the Computational Stylistics Group (2023). In addition to these, the expansive DraCor project (Fischer et al., 2019) contains dramatic texts across numerous languages and periods. This leaves poetry, the last of the three main literary genres, without a dedicated resource, a situation that hinders research in computational poetics and comparative poetry studies.

Several monolingual corpora of poetry have already been built (Bobenhausen & Hammerich, 2015; Delente & Renault, 2021; Grishina et al., 2009; Haider, 2021a; Horváth et al., 2022; Mittmann, 2019; Navarro-Colorado et al., 2017; Plecháč & Kolár, 2015; Ruiz Fabo et al., 2021), yet their structures and tag sets are not mutually compatible, and the depth of their annotation varies. While the recently released Python library *Averell* (Díaz Medina et al., 2021) aims to transform these resources into a unified JSON output, it is hampered by a critical problem, namely that it is not well adapted to the structural peculiarities of the original datasets. Consequently, a large part of the data is lost (out of more than 18,000 poems in the French corpus, for example, only 5,081 make it to the JSON output; similarly, almost 15,000 poems are lost from the Italian corpus).

In this article, we present a dataset entitled PoeTree (Poetry Treebanks), comprising poetry corpora in ten different languages (Czech, English, French, German, Hungarian, Italian, Portuguese, Spanish, Slovenian, and Russian), with a total of more than 330,000 poems / 89,000,000 tokens. All texts have been deduplicated, morphologically tagged, and parsed for syntactic dependencies with UDpipe. All information is encoded in a shared simple JSON structure.

2. Resources

- **Poetree deposited at Zenodo** – DOI:www.doi.org/10.5281/zenodo.10907309
- **Other access points (as of April 2024)**
  - REST API – URL:https://versologie.cz/poetree/api_doc
  - Python library – URL:www.pypi.org/project/poetree
  - R library – URL:www.github.com/perechen/poetRee
- **Temporal coverage:** 13th century-20th century; 2009–2023 (construction)

Data stems from the following resources (we refer to each corpus by its ISO 639-1 language code):

- CS: The Corpus of Czech Verse (Plecháč & Kolár, 2015)
- DE: German Poetry Corpus (Textgrid and DTA) (Bobenhausen & Hammerich, 2015; Haider, 2021a; 2021b; 2023; 2024)
- ES: Corpus of Spanish Golden Age Sonnets (Navarro-Colorado et al., 2017) + Diachronic Spanish Sonnet Corpus (Ruiz Fabo et al., 2021)
- FR: Corpus Malherbə (Delente & Renault, 2021)
- HU: ELTE Poetry Corpus (Horváth et al., 2022)
- IT: Biblioteca italiana (2023)
- PT: Poemas (Mittmann et al., 2019)
- RU: Corpus of Russian Poetry (Grishina et al., 2009)

To the best of our knowledge, there are currently two open corpora of English Poetry (Parrish, 2018, and Haider, 2021b, 2023), both based on texts available at Project Gutenberg (2023). The former is known, however, to be vastly contaminated by non-versified documents (fiction, comments, etc.; cf. Pace-Sigge, 2019), while the latter, in its efforts to clean the data of these contaminants, seems to go too far, omitting a large part of the original data. In light of this, we have decided to compile EN from scratch for the sake of the PoeTree collection. The texts were acquired from Project Gutenberg through GutenTag (Brooke et al., 2015). Although each text has been manually checked for tagging errors, some of these were beyond repair. This concerns chunks of verse that the system misclassified as prose and the lines of which were merged into a single paragraph. These parts were thus omitted from the final corpus. Although rather infrequent, it remains a known bug in EN.

In addition to these, Slovenian corpus (SL) has been compiled from texts available at wikisource platform, part of which was published within the project "Slovenska leposlovna klasika", financed by the Ministry of Culture of the Republic of Slovenia.

As is shown in Figure 1, the corpora vary largely both in size and time coverage. PoeTree is thus by no means a balanced dataset and does not really aim to be one. Should it comprise poetry in "big" languages with long-lasting traditions along with that in "small" ones, it would necessarily mean getting rid of data in the former. If certain research tasks require it, we think downsampling methods will better meet researchers' needs.

3. Cleaning Data

In the aforementioned resources were occasional texts written in foreign languages. We tried to minimise such cases by means of automatic language

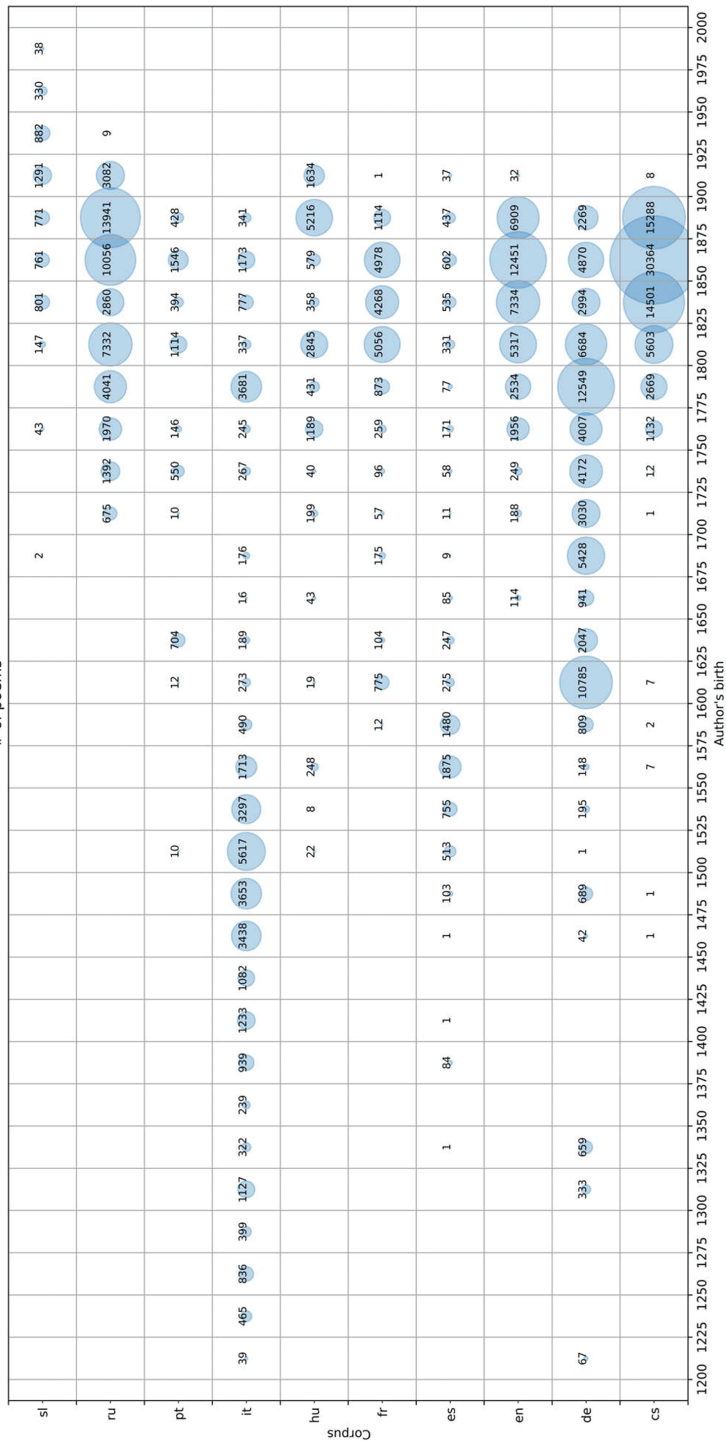


FIGURE 1 Number of poems (duplicates excluded) matched to the years of birth of their authors (25-year range)

detection. For each poem we have used the *langdetect* Python library (Danilák, 2021), using probabilities to determine the language of the text. In those cases where the probability of the respective language was lower than 0.99, the poem was subjected to a manual check and eventually removed. (Even with the threshold set this high the number of poems to inspect was in the lower hundreds.)

Another problem was posed by the existence of duplicates, which is to say multiple identical or slightly differing texts entering a single corpus from different editions. We aimed to identify these by means of approximate substring matching which – unlike the plain vanilla edit distance – is also able to capture cases such as A. Ducros' poem 'Les rubans de Marie', which occurs in FR twice: once encoded as a single poem (coming from the 1854 book *Les Capricieuses*) and once split into four parts 'Ruban blanc', 'Ruban bleu', 'Ruban vert', and 'Ruban noir' (coming from the 1896 collection *Les Caresses d'antan*). The procedure was as follows:

– Let similarity of poems A and B containing $|A| \geq |B|$ characters respectively be defined as:

$$\mathrm{sim}(A, B) = 1 - \frac{\min(\mathrm{lev}(a_1, B), \ldots, \mathrm{lev}(a_n, B))}{|B|}$$

  where $\{a_1, \ldots, a_n\}$ is the set of all possible substrings of A and $\mathrm{lev}(a_x, B)$ is the Levenshtein distance between a_x and B.
– For each author in each corpus construct an undirected graph where nodes represent their poems and an edge exists between A and B if $\mathrm{sim}(A, B) > 0.75$. (In all corpora, the distributions of poem pairs' similarities are strongly bimodal [see Figure 2] with a major peak between 0.4 and 0.5 [completely unrelated texts] and a minor peak at 1 [completely identical texts]. The threshold of 0.75 above which poems are considered duplicates roughly corresponds to their local minima.] An example of such a graph for M. Arnold is given in Figure 3a.
– For each component of each graph mark one of its nodes as a primary variant and the rest as its duplicates in the following way:
  – if the component is complete (see Figure 3b):
  – limit the primary variant candidates to the poems with the highest number of lines
  – if multiple candidates remain and if the year of creation/publication is known for all of them, limit the candidate set to the earliest ones
  – if multiple candidates remain, select the primary variant by random

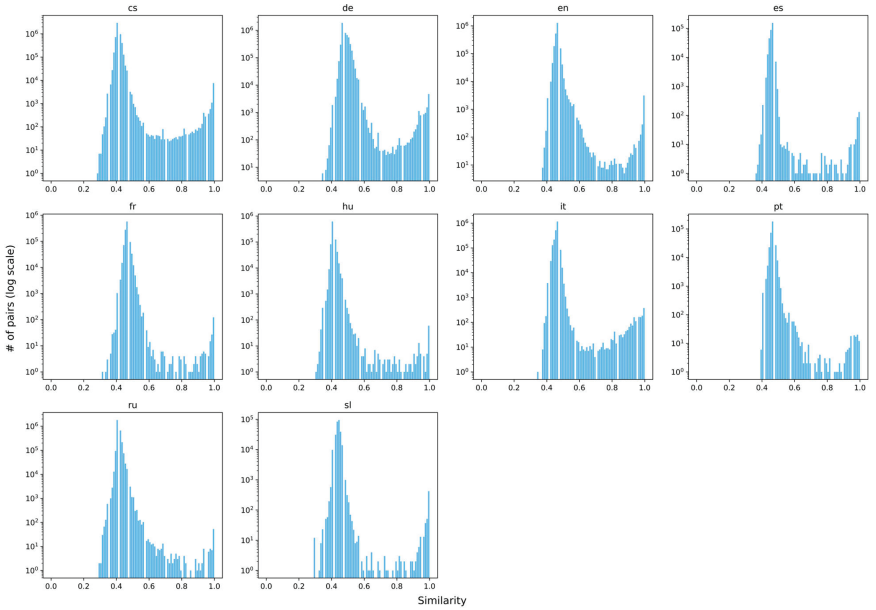
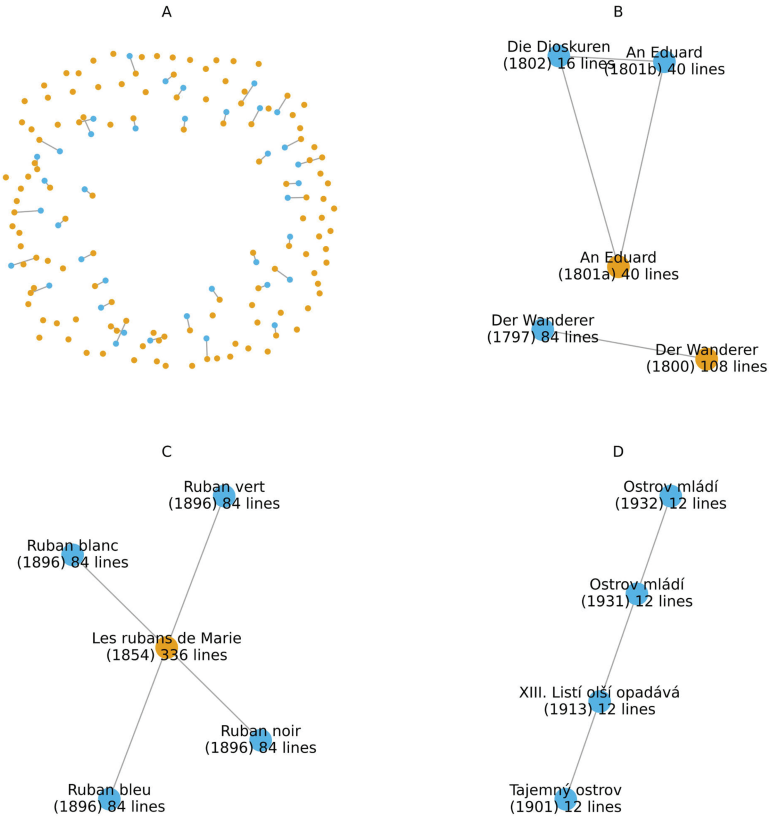


FIGURE 2 Distribution of poem pairs' similarities in each corpus

– else if the component is a star and the central node is a poem with the highest number of lines (see Figure 3c), mark the central node as the primary variant
– else: determine the primary variant manually (see Figure 3d)

In this way, 20,999 complete components (88% of which comprised just two nodes) and 585 star components were deduplicated automatically, while 75 components were processed manually (usually concerning cases when a certain poem was continuously reworked up to the point that the similarity of the initial and the final variant was below the threshold). Figure 4 gives the corpora sizes after both language detection and deduplication steps.

Given that this way of deduplication may not be suitable to all possible use cases, we have preserved information in each case not only on whether a poem was marked as duplicate according to the steps outlined above but also on the measure of similarity to its 20 nearest neighbours, with the aim of making it possible for PoeTree users to apply other deduplication criteria. Deduplication scripts are available at https://github.com/versotym/poetree_deduplication. Interactive similarity graphs may be inspected in detail at https://versologie.cz/poetree/deduplication.



Note: Orange indicates a primary variant. (A) Graph representing all poems of M. Arnolds (EN). (B) Two complete clusters from the graph of F. Hölderlin (DE). 'Der Wanderer' (1797) is marked as a duplicate since it has fewer lines than 'Der Wanderer' (1800). Within the other component 'Die Dioskuren' is ruled out on account of its length, 'An Eduard' (1801a) is then randomly selected as the primary variant since the two remaining poems have the same number of lines and come from the same year. (C) Star component from the graph of A. Ducros (FR). The central node is selected as the primary variant as it has the highest number of lines. (D) Component from the graph of E. Lešehrad (CS) to be resolved manually.

FIGURE 3 Deduplication through undirected graphs

4. Enriching Data

Where available, author records are enriched with VIAF id and wikidata entity id (the former identifier was already present in the ELTE Poetry Corpus and Diachronic Spanish Sonnet Corpus). This allows not only to unify pen names and alternate spellings under a single identity, but also to acquire additional metadata such as date of birth, date of death, and country of citizenship.

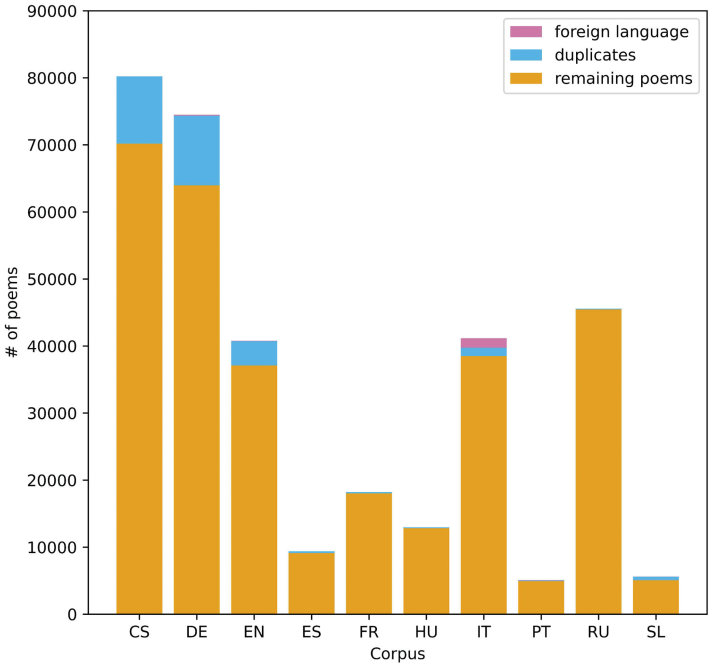


FIGURE 4 The number of poems in each corpus after language detection and deduplication

We enrich each poem with lemmatization, morphological tagging, and syntactic parsing according to the Universal Dependencies annotation scheme, using the multilingual UDPipe 2 parser (Straka, 2018). Lemmatization and morphological tags allow for retrieving words in specific contexts. Typically, a researcher might want to retrieve grammatical collocations that denote entities and their properties, or events with their participants and circumstances. This easily translates into nouns and their attributes, verbs and their arguments (subject, objects), and adjuncts (adverbials). Unlike bag-of-words approaches or ordinary linear searches, syntactic parsing allows for direct queries about syntactic elements, abstracting from auxiliary words, modifiers, and nested clauses that might obscure them.

The quality of morphosyntax-based information extraction depends on the quality of the automatic parsing. However, most language models have been trained on modern non-fiction, with the consequence that the rate at which they generate adequate results on older texts, especially in the case of poetry, may be lower than documented. The only way to assess the performance of a parser on a particular domain is by evaluating it on a manually annotated data set from that domain. We have performed such an evaluation of the largest

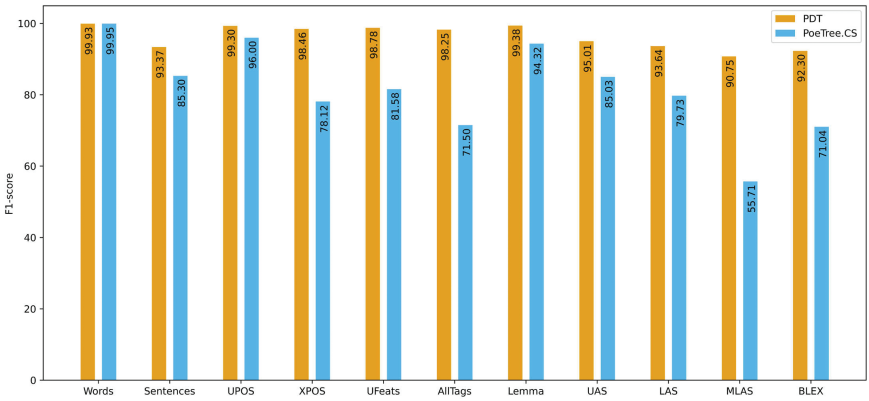


FIGURE 5 UDPipe 2 performance with Prague Dependency Treebank (PDT) as compared to the testing portion of PoeTree.CS (6591 tokens)

Czech model (based on Prague Dependency Treebank newswire texts from the 1990s) on a random sample of 29 poems from the Czech PoeTree section. The performance was indeed lower in all standard metrics (see Figure 5), and a semi-manual error analysis revealed several systematic errors that would hamper proper extraction of relevant syntactic relations, especially concerning nouns as modifiers of other nouns, which tended to be attached instead as arguments of the nearest governing predicate (cf. Cinková et al., 2024). This finding reveals the need for a domain adaptation of the Czech model to older Czech (poetry) and calls for the same procedure to be performed on the other PoeTree languages.

5. Standardisation

In PoeTree, each poem is stored as a standalone JSON file with a standardised structure. We considered using TEI-XML, a widely used format in the community but ultimately decided against it as a storage format for the corpus that focuses on encoding linguistic data over the source and editorial information. JSON also provides operational ease in our case, as it could be easily manipulated across different coding frameworks and approaches, including research communities that are not familiar with TEI. We invite the creation of converters and wrappers that adapt our corpus to TEI schemas or present them in other custom formats.

The top-level keys of the JSON structure are shown in Table 1. The last three keys hold complex data structures. The 'source' key holds an object comprising metadata on a particular book edition from which a text comes (see Table 2). The 'author' keys may hold either an object, schema (shown in Table 3), or array of such objects in the case of poems with multiple authors or multi-author books where the authorship of particular poems is unknown. The 'body' key holds the text of the poem itself. It is an array where each element corresponds to a single line (see Table 4). In each line, there is a 'words' key which holds the linguistic analysis provided by UDPipe-2. The default CoNLL-U format (Universal Dependencies, 2013) is split into an array whose elements correspond to particular tokens (see Table 5). Note that unlike CoNLL-U we do not encode multiwords as standalone tokens, but rather delegate this

TABLE 1 Structure of poem metadata

Key	Data type	Description
id	string	id of the poem; points to a source file in original resource
title	string | null	title of the poem
year_created	number | array | null	year when poem created (may precede the date of publication); time span encoded as [year_min, year_max]; only available in DE and RU
duplicate	string | false	whether poem is considered to be a duplicate; if so, id of the primary variant is given here
neighbors	array	ids of the 20 most similar poems by the same author; each one is encoded as [id, similarity measure] in descending order
source	object	metadata on poem's source book
author	object | array	metadata on poem's author
body	array	array of poem's lines

TABLE 2 Structure of metadata on source of poem

Key	Data type	Description
id	string | null	id of book
title	string | null	title of book
year_published	number | null	year when book published
publisher	string | null	name of publisher
place	string | null	place of publication (city)
corpus	string	name of resource from which data is derived

TABLE 3 Structure of metadata on author of poem

Key	Data type	Description
name	string	name of author as it appears in source; [anonymous] marks unknown authors
viaf	string | null	VIAF id of author
wiki	string | null	wikidata id of author
country	string | null	country of citizenship in iso 639-1 format
born	number | null	author's year of birth
died	number | null	author's year of death

TABLE 4 Structure of encoding line of poem

Key	Data type	Description
id	number	index of line; zero-base; increment through the entire poem (does not restart in new stanza)
stanza_id	number	index of stanza; zero-based
text	string	text of line

TABLE 4 Structure of encoding line of poem (*cont*)

Key	Data type	Description
part	string | false	if verse-line is divided into multiple text-lines, this shows whether it is an initial part ('I'), medial part ('M') or final part ('F'); false if the line is not divided
words	array	output of UDPipe 2 provided as an array of tokens

TABLE 5 Structure of linguistic annotation

Key	Data type	Description
id	number	word index; integer starting at 1 for each new sentence
form	string	word form or punctuation symbol
lemma	string	lemma or stem of word form
upos	string	universal part-of-speech tag
xpos	string	language-specific part-of-speech tag; underscore if not available
feats	string	list of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available
head	number	head of the current word, which is either a value of id or zero (0)
deprel	string	universal dependency relation to the head (root if head = 0) or a defined language-specific subtype of one
deps	string	enhanced dependency graph in the form of a list of head-deprel pairs
sentence	number	sentence index; one-based
multiword	object	this is an optional key; it appears only if a given word is part of a multiword token, if so the structure is {'form' (string): form of the multiword, 'id' (number): multiword index; integer starting at 1 for each new sentence}

information to an optional 'multiword' key of its components. For instance, while a Spanish phrase 'Esperaré del mal' gives 5 tokens in CoNLL-U:

1	Esperaré	…
2-3	del	…
2	de	…
3	el	…
4	mal	…

in PoeTree this is encoded as a 4-element list:

```
[
{id: 0, form: "Esperaré", …},
{id: 1, form: "de", …, multiword: {"form": "del", id: 1}},
{id: 2, form: "el", …, multiword: {"form": "del", id: 1}},
{id: 3, form: "mal", …},
]
```

6. Conclusion and Future Plans

PoeTree in its current state offers an extensive dataset suitable for various tasks (not only) in the field of NLP, stylometry, and computational literary studies.

In the upcoming two years, we aim to evaluate the parser performance on other languages represented in PoeTree, and, most importantly, enrich PoeTree with rhyme detection, fixed forms (sonnet, sestina, etc.) description, topic modelling, and recognised named entities that would link to a common knowledge base (wikidata). Furthermore, we plan to incorporate and standardise the annotation of poetic metres from the original resources (where available and permitted by the license) and to perform our own machine-driven metre detection in the remaining corpora. This would make PoeTree the only existing full-text dataset with comparative information on poetic forms that is aligned across languages. We hope this will enable research that was not possible before: from the evolution of poetic forms to the tracing of literary contacts across cultures, and answers to fundamental questions about the connection between form and meaning from the historical perspective.

Acknowledgements

The creation of this dataset was supported by the Czech Science Foundation (project GA23-07727S).

References

Bamman, D. (2021). *BookNLP*. GitHub. www.github.com/booknlp/booknlp.

Biblioteca italiana. (2023). *Biblioteca italiana*. www.bibliotecaitaliana.it.

Bobenhausen, K., & Hammerich, B. (2015). Métrique littéraire, métrique linguistique et métrique algorithmique de l'allemand mises en jeu dans le programme Metricalizer². *Langages*, 199, 67–87. www.cairn.info/revue-langages-2015-3-page-67.htm?contenu=article.

Brooke J., Hammond A., & Hirst, G. (2015). *GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg Corpus*. In A. Feldman, A. Kazantseva, S. Szpakowicz, & C. Koolen (Eds.), *Proceedings of the fourth workshop on computational linguistics for literature* (pp. 42–47). Association for Computational Linguistics. www.doi.org/10.3115/v1/W15-0705.

Byszuk, J., Woźniak, M., Kestemont, M., Leśniak, A., Łukasik, W., Šeļa, A., & Eder, M. (2020). Detecting direct speech in multilingual collection of 19th-century novels. In R. Sprugnoli, & M. Passarotti (Eds.), *Proceedings of LT4HALA 2020 – 1st workshop on language technologies for historical and ancient languages* (pp. 100–104). ELRA. www.lrec-conf.org/proceedings/lrec2020/workshops/LT4HALA/pdf/2020.lt4hala-1.15.pdf.

Cinková, S., Plecháč, P., & Popel, M. (2024). Rhymes and syntax. Morpho-syntactic analysis of the Czech poetry. *Primerjalna Književnost*, *47*(2), 65–88. www.doi.org/10.3986/pkn.v47.i2.04.

Computational Stylistics Group. (2023). *Resources*. https://computationalstylistics.github.io/resources.

Danilák, M. (2021). *Langdetect*. GitHub. www.github.com/Mimino666/langdetect.

de la Rosa, J., Pérez Pozo, Á., Ros, S., & González-Blanco, E. (2023). *ALBERTI, a multilingual domain specific language model for poetry analysis*. arXiv. www.doi.org/10.48550/arXiv.2307.01387.

Delente, É., & Renault, R. (2021). Projet Anamètre: présentation, limites et avancées. In A.-S. Bories, G. Purnelle, & H. Marchal (Eds.), *Plotting poetry: On mechanically-enhanced reading* (pp. 73–92). Presses universitaires de Liège.

Díaz Medina, A., Pérez Pozo, Á., & de la Rosa, J. (2021). *Averell: A corpus management tool to transform poetic corpora into a JSON format compliant with the POSTDATA ontology* (v1.2.2). Zenodo. www.doi.org/10.5281/zenodo.5702404.

Du, K., Dudar, J. & Schöch, C. (2022). Evaluation of measures of distinctiveness: classification of literary texts on the basis of distinctive words. *Journal of Computational Literary Studies*, *1*(1). www.doi.org/10.48694/jcls.102.

Fischer, F., Börner, I., Göbel, M., Hechtl, A., Kittel, C., Milling, C., & Trilcke, P. (2019). Programmable corpora: Introducing DraCor, an infrastructure for the research on European drama. In *Proceedings of DH2019*. Utrecht University. www.doi.org/10.5281/zenodo.4284002.

Grishina E., Korchagin K., Plungian V., & Sichinava, D. (2009). Poeticheskii korpus v ramkah NKRIA: obschaia struktura i perspektivy ispolzovania. In *Natsionalnii korpus russkogo iazyka: 2006–2008. Novye rezultaty i perspektivy* (pp. 71–113). Nestor-Istoria.

Haider, T. (2021a). *A German Poetry Corpus / Deutsches Lyrik Korpus* (*DLK*). GitHub. www.github.com/tnhaider/DLK.

Haider, T. (2021b). Metrical tagging in the wild: Building and annotating poetry corpora with rhythmic features. In Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 3715–3725). Association for Computational Linguistics. www.doi.org/10.18653/v1/2021.eacl-main.325.

Haider, T. (2023). *A computational stylistics of poetry: Distant reading and modeling of German and English verse*. Doctoral Thesis. In: OPUS University of Stuttgart. www.doi.org/10.18419/opus-12721.

Haider, T. (2024). A large annotated reference corpus of New High German Poetry. In *Proceedings of LREC-COLING*. Torino. www.aclanthology.org/2024.lrec-main.59/.

Horváth, P., Kundráth, P., Indig, B., Fellegi, Z., Szlávich, E., Borbála Bajzát, T., Sárközi-Lindner, Z., Vida, B., Karabulut A., Timári M., & Palkó, G. (2022). ELTE Poetry Corpus: a machine annotated database of canonical Hungarian poetry. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the 13th conference on language resources and evaluation* (*LREC 2022*) (pp. 3471–3478). ELRA. www.aclanthology.org/2022.lrec-1.372.

Mittmann, A., Esteves, E., & Luiz dos Santos, A. (2019). What rhythmic signature says about poetic corpora. In P. Plecháč, B. P. Scherr, T. Skulacheva, H. Bermúdez-Sabel, & R. Kolár (Eds.), *Quantitative approaches to versification* (pp. 153–172). ICL CAS. https://versologie.cz/conference2019/proceedings/mittmann-pergher-dossantos.pdf.

Navarro-Colorado, B., Ribez Lafoz, M., & Sánchez, N. (2017). Metrical annotation of a large corpus of Spanish sonnets: representation, scansion and evaluation. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the tenth international conference on language resources and evaluation* (*LREC 2016*) (pp. 4360–4364). ELRA. www.lrec-conf.org/proceedings/lrec2016/pdf/453_Paper.pdf.

Odebrecht, C., Burnard, L., & Schöch, C. (2021). *European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels* (v1.1.0) [Data set]. Zenodo. www.doi.org/10.5281/zenodo.4662444.

Pace-Sigge, M. (2019). Typical phraseological units in poetic texts. In G. Corpas Pastor, & R. Mitkov (Eds.), *Computational and corpus-based phraseology* (pp. 330–344). Springer. www.doi.org/10.1007/978-3-030-30135-4_24.

Parrish, A. (2018). *A Gutenberg Poetry Corpus*. GitHub. https://github.com/aparrish/gutenberg-poetry-corpus.

Plecháč, P. (2021). *Versification and authorship attribution*. Institute of Czech Literature, CAS and Karolinum Press. www.doi.org/10.14712/9788024648903.

Plecháč, P., & Kolár, R. (2015). The Corpus of Czech Verse. *Studia Metrica et Poetica*, *2*(1), 107–118. www.doi.org/10.12697/smp.2015.2.1.05.

Project Gutenberg. (2023). *Project Gutenberg*. www.gutenberg.org.

Ruiz Fabo, P., Bermúdez Sabel, H., Martínez Cantón, C., & González-Blanco, E. (2021). The diachronic Spanish sonnet corpus: TEI and linked open data encoding, data distribution, and metrical findings. *Digital Scholarship in the Humanities*, *36* (Supplement_1, June 2021), i68–i80. www.doi.org/10.1093/llc/fqaa035.

Šeļa, A., Plecháč, P., & Lassche, A. (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), Article e0266556. www.doi.org/10.1371/journal.pone.0266556.

Storey, G., & Mimno, D. (2020). Like Two pis in a pod: Author similarity across time in the Ancient Greek Corpus. *Journal of Cultural Analytics*, *5*(2). www.doi.org/10.22148/001c.13680.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In D. Zeman, & J. Hajič (Eds.), *Proceedings of CoNLL 2018: the SIGNLL conference on computational natural language learning* (pp. 197–207), Association for Computational Linguistics. www.aclanthology.org/K18-2020.

Universal Dependencies (2013). *CoNLL-U Format*. www.universaldependencies.org/format.html.