D3.3

SHOWCASES FOR THE APPLICATION OF CLS METHODS AND TOOLS

Editors: Christof Schöch, Evgeniia Fileva, Julia Dudar

Authors: Peter Andorfer, Ingo Börner, Vera Maria Charvát, Aitor Díaz Medina, Julia Dudar, Matej Ďurčo, Evgeniia Fileva, Frank Fischer, Elena González-Blanco, Carsten Milling, Martin Anton Müller, Álvaro Pérez Pozo, Salvador Ros Muñoz, Javier de la Rosa Pérez, Christof Schöch, Artjoms Šeļa, Henny Sluyter-Gäthje, Gerd-Hermann Susen, Peer Trilcke, Laura Untner, Evgeniya Ustinova

Date: 29 February, 2024

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004984

Project Acronym: CLS INFRA

Project Full Title: Computational Literary Studies Infrastructure

Grant Agreement No.: 101004984

Deliverable/Document Information

Deliverable No.: D3.3

Deliverable Title: SHOWCASES FOR THE APPLICATION OF CLS METHODS AND TOOLS

Editors: Christof Schöch, Evgeniia Fileva, Julia Dudar

Authors: Peter Andorfer, Ingo Börner, Vera Maria Charvát, Aitor Díaz Medina, Julia Dudar, Matej Ďurčo, Evgeniia Fileva, Frank Fischer, Elena González-Blanco, Carsten Milling, Martin Anton Müller, Álvaro Pérez Pozo, Salvador Ros Muñoz, Javier de la Rosa Pérez, Christof Schöch, Artjoms Šeļa, Henny Sluyter-Gäthje, Gerd-Hermann Susen, Peer Trilcke, Laura Untner, Evgeniya Ustinova

Dissemination Level: PRIVATE OR PUBLIC

Document History

Version / Date	Changes / Approval	Authors / Approved by
v0.9.0, 2024-02-10	Initial version	All authors of individual showcase descriptions, approved by Christof Schöch.
v1.0.0, 2024-02-29	First complete version	All authors of individual showcase descriptions, approved by Christof Schöch.

Preface

The purpose of the showcases for the application of methods and tools in Computational Literary Studies (CLS) presented here is to illustrate, in a concrete, visual and interactive way, how some of the key methods in CLS work when they are applied to collections of literary texts. In addition, the showcases are designed as demonstrations, with a relatively low threshold for access, of how tools and datasets as elements of an integrated infrastructure such as the one developed in CLS INFRA, can be used in research. This is done here in the hope that the showcases also help us draw additional scholars towards learning about and using such methods, datasets and infrastructure services.

With this goal, we have implemented and described four showcases. They are all available from https://showcases.clsinfra.io/ and bring together different methods and kinds of datasets:

1. The "Multilingual Stylometry with ELTeC" showcase combines sets of novels in multiple languages with the method of stylometric authorship attribution. It was designed by a team in Trier, Germany (TCDH) and Kraków, Poland (IJP PAN).
2. The "Detecting Small Worlds in a Corpus of Thousands of Theater Plays" showcase is concerned with the structural analysis of character networks in European drama and was developed by the team in Potsdam, Germany (UP).
3. The showcase on "Scansion tools for corpora annotation and visualization: the Poetrylab Suite" (part 1: Averell and part 2: Poetrylab and Rantanplan) concerns poetry scansion, that is the extraction of stress patterns of lines or verses from a number of poetry corpora that can be obtained and analyzed with considerable flexibility; it was developed by the team in Madrid, Spain (UNED);
4. The "Mapping Arthur Schnitzler in Space and Time" showcase concerns spatio-temporal mapping in correspondences and other fictional and non-fictional materials using Linked Open Data, and was developed by a team in Vienna, Austria (OEAW).

Each of the showcases consists of three elements: First, a visual and web-based presentation of functionality and/or results that allows users to interact with the data and the analyses, or to perform certain steps of manipulation of the data, complete with very short explanations of functionality and context. (This element, the showcases proper, are freely available online at the links indicated above.) Second, a somewhat longer explanation of the method and dataset used in each showcase, and of the key functionality and results. (This second element documents the showcases, is collected in the present deliverable and will also be made available online alongside the showcases proper). And third, each showcase is complemented by a more developed and more detailed scholarly paper that explains the research question, datasets used, the methods employed and the results in considerable depth. (Depending on the showcase, these scholarly papers have already been published elsewhere in open access or are in the process of being submitted or reviewed.)

Christof Schöch

Table of contents

Multilingual Stylometry with ELTeC

How well does stylometric authorship attribution work across languages and corpora? A showcase on multilingual stylometry using corpora from the European Literary Text Collection (ELTeC).

URL

https://showcases.clsinfra.io/stylometry

Repository

https://gitlab.clsinfra.io/cls-infra/d33/

Status / publication date

Showcase available online, 25.02.2024.

Creators / developers / authors

- Julia Dudar (Trier University)
- Evgeniia Fileva (Trier University)
- Artjoms Šeļa (Institute of Polish Language, Kraków)
- Christof Schöch (Trier University)

Target audience

The target audience of this showcase are scholars and students interested in a key methodological aspect of stylometry-based authorship attribution as well as those interested in cross-lingual approaches within the digital humanities. The showcase is intended both as a research output and as a pedagogical tool that could be used in workshop or classroom settings alike.

Executive summary

This showcase aims to disentangle the dual influence of language and corpus composition on the performance of stylometric methods of authorship attribution.

Authorship attribution is the task to assign texts of unknown or disputed authorship to their most likely authors (Byszuk, 2023). Fundamentally, the way stylometric methods approach authorship attribution is to use the frequencies of a large number of simple features, such as words or character sequences, for a determination of the degree of similarity between texts. These similarities, in turn, are interpreted as an indicator of the likelihood for two texts to have been written by the same author. A key methodological challenge in this area is the influence of both corpus composition and language not only on the results for a specific case, but also for the overall performance and reliability of stylometric methods of authorship attribution.

This show-cases tackles this issue by investigating four distinct but broadly comparable corpora in a classification scenario. The corpora are all derived from the European Literary Text Collection (ELTeC; Schöch et al. 2021) and each contains novels by 8-10 different well known authors. They can therefore be used as benchmarking datasets, where the true authorship is known for all novels under investigation. The corpora differ both in terms of their composition, which can affect overall difficulty of attribution and therefore attribution performance, and in terms of their language, which can likewise affect attribution performance. Specifically, corpora in English, French, Hungarian and Ukrainian are included.

In order to investigate the role of composition and language separately, all four corpora were automatically translated into the other three languages using the DeepL Pro machine translation system. This allows us to vary corpus composition and language independently, and tease out effects of both in terms of attribution performance (measured as classification accuracy).

In addition, and in order not to disadvantage any one corpus or language by an unsuitable method or selected feature, a number of further parameters have been varied, namely: the type of feature considered (word forms, lemmas, part of speech or characters); the length of the feature sequence considered as the unit of analysis (unigrams, bigrams, 3-grams, 4-grams or 5-grams); the total number of different features considered (from 50 to 2000 in several increments); and the length of the textual segments considered (from 5000 words to 10000 words as well as using entire novels).

Testing all of these various parameters, corpora and languages in all of their possible combinations results in a large number of individual results. In order to make them accessible to inspection and analysis, we have developed a visualization that displays the attribution quality (i.e. the classification accuracy) in a heatmap as a function of the parameters described above. While the heatmap displays the accuracy for the entire range of features and segment lengths at the same time, other parameters can be selected by the users in order for the heatmap to be updated with the corresponding accuracy values. Two such heatmaps are provided for easy comparison of results between any two configurations of parameters, whether corpora, languages or other settings.

The detailed results remain to be investigated more closely, but some initial observations can be made. For example, the difference in performance between corpus A and corpus B in two different languages can usually be reduced if for corpus B, rather than the original version, the version translated into the language of corpus A is selected (or vice versa). Similarly, translated corpora usually lead to a lower attribution accuracy, overall, compared to their counterpart in the original language.

Research questions and objectives

The primary research question this showcase addresses is to what extent, how, and under what conditions, the composition of both language and corpus influence the performance of stylometric methods of authorship attribution.

The primary objective of this showcase, therefore, is to provide detailed data that allows us to evaluate the effectiveness of the stylometric method for corpora of different compositions and in different languages. Specifically, the study aims to determine how the results of stylometry are influenced by the language of the texts (using translations) and the composition of the corpus, considering the degree of similarity both within an author's works and between texts of different authors.

An additional, secondary research question is the following: what makes a corpus “difficult” in terms of authorship attribution?We know that this is likely to depend both on corpus composition and on language; but predicting the difficulty based on corpus composition is particularly complex, since it's usually unrealistic to control for every potential source of variation. E.g. if we are attributing a gothic novel, it’s unlikely to find a reference corpus to consist only of gothic novels written with the same narrative techniques by authors that only come from the same cohort and social background.

Several key aspects likely need to be considered: The level of distinctness or overlap of the two similarity distributions for pairs of works known to be written by the same author, on the one hand, and those known to be written by two different authors, on the other hand, are likely to be a good indicator of attribution difficulty. However, the properties of the works that determine this similarity a priori, that is before a particular stylometric test is performed to establish this similarity based e.g. on word frequencies, is not entirely clear: time period, literary genre and subgenre, narrative perspective or metric form, author gender, author age at the time of writing, and several others may have some relevance. However, it is not clear how to quantify the various metadata configurations for a sensible prediction of overall corpus difficulty. This showcase makes a preliminary attempt some suggestions for this issue but leaves most of this to future work.

Data

The text corpus was formed based on the ELTeC corpora in four languages: French, English, Hungarian, and Ukrainian. The choice of languages is determined by their representativeness for the corresponding language group. For example, using Germanic, Romance, Finno-Ugric, and Slavic languages, one can illustrate the effectiveness of the stylometry method and see how language affects the analysis results. The original ELTeC collection includes one hundred novels in each corpus. For our research, a selection of novels was made, resulting in the following collections of texts: the English corpus includes 44 novels, the French one 30, the Hungarian one 27, and the Ukrainian one 26 novels. Each of the corpora was translated into the other three languages, thus the entire corpus includes 16 sub-corpora. The dataset also includes a metadata table collected for the novels in each language. The metadata table includes information about the author, year of publication of digital and print editions, identification number, language, number of words in the novel, as well as subgenre (social, historical, adventure, detective or sentimental novel, bildungsroman or other) and narrative perspective of the novel (heterodiegetic, homodiegetic, epistolary, dialogue or mixed).

Data preparation / preprocessing

Data preparation involved two primary stages: translation of texts into the other languages and their linguistic annotation. The texts were translated automatically using DeepL Pro. For subsequent analysis, extracting lemmas and POS (part of speech) tags was necessary, accomplished using the SpaCy library for both original and translated texts. Additionally, unigrams and n-grams (from 2 to 5) were extracted using the stylo package.

Method(s)

We use grid search over the features space to assess the general performance of authorship attribution, inspired by Rybicki and Eder (2011). The goal is not to optimize for performance, but to find reasonable approaches that work for different languages and to understand the nature of differences in performance across languages and corpora. There are three main levels of variation: features, sample size and vector length.

Features

We use frequencies of lemmas, word forms, character n-grams, and part-of-speech (POS) tags. Each feature is cut to n-grams of different n: 1-3 for words and lemmata, 2-5 for character and POS n-grams. Each n-gram length is tested independently.

Sample size

Frequency-based approaches to authorship attribution naturally depend on the available size of the text. There is a considerable variation in text sizes in ELTeC; to mitigate this, we draw a random sample of consecutive tokens (a “chunk”) for each text based on the smallest text across corpora (10k words). We use test sizes of 5000 to 10000 tokens for word-based features, and 10000 to 50000 for character-based ones. The number of available tokens per text differs dramatically between words and characters; so sample sizes mean different things for n-grams based on these features.

To account for the variability that is introduced with taking only one limited sample out of sometimes very large novels, at each step we take a random consecutive sample out of all available ‘chunks’ and record performance 100 times. As the last step, we perform classification on full-length texts.

Vector length

Vectors that represent texts are constructed based on most frequent feature cutoffs length: 50 to 2000 with incrementally increasing step sizes.

For author-based text classification, we use the Support Vector Machine classifier and perform leave-one-out cross-validation. In each step, one text is removed from the data, the model is then trained on the remaining texts, then the authorship of the left-out text is predicted and the result is recorded. This process continued until all texts were left out once. We report two performance measures: accuracy (proportion of correctly predicted authors), Cohen’s kappa (which is used for measuring inter-annotator agreement, adjusted for classification by chance and thus can be useful when comparing classifiers with different number of classes).

Visualization

The visualization consists of two heatmaps, enabling users to compare and contrast two sets of results simultaneously. The x-axis represents various settings of the most frequent features (mff) used in specific analysis, while the y-axis denotes distinct sample sizes, ranging from 5000 to 10000 tokens. Certain mff values are excluded from the x-axis due to either the absence of results. Each heatmap cell correlates mff and sample size, with the color intensity indicating accuracy levels, offering a visual metric of analysis precision. A mouseover provides further information, such as the features used as well as numerical indications of accuracy and Cohen’s Kappa.

Users can engage with the data through several selectors, enhancing the exploration and analysis process. The available choices include:

1. Corpus selector: Enables the selection from a range of available corpora, facilitating comparative studies;
2. Ngram size selector: Offers options to choose n-gram sizes from 1 to 5, allowing for detailed linguistic patterns examination;
3. Feature level selector: Permits users to toggle between 'words' and 'chars' (characters) feature types;
4. Feature type selector: Allows for the choice between plain text, lemmas, or POS tags.

To create the interactive visualizations of the showcase, Bokeh was employed, which is a Python library offering extensive capabilities in crafting interactive and dynamic data visualizations. Given the high information density of the visualization, we have integrated the Bokeh's hover tool to provide users with informative tooltips. This feature unveils detailed data points, such as accuracy and kappa values, when a user hovers over any part of the heatmap, significantly enhancing the data's interpretability.

A color scale on the right side of each graph delineates the accuracy levels, ranging from 0 to 0.9. Here, cooler tones like blues and greens signify lower values, whereas warmer hues, orange and red, indicate higher values. This color coding serves as a primary indicator of stylometric effectiveness across various parameter settings, vividly illustrating the success rate of stylometric analysis per chosen corpus and selector criteria.

While our visualization provides extensive opportunities for interpreting the results, the scope of this report allows only for a summary of key findings, therefore we focus on the outcomes of token- and lemma-based 1-gram stylometric analysis.

Outcomes

In relation to the main outcomes, we aim to present a concise overview of the optimally performed settings. First, we focus on evaluating the performance of each corpus, and then provide a brief account of the overall performance for each language.

The Hungarian corpus (Fig. 1) demonstrates the highest accuracy score, reaching approximately 90% across most sample sizes. This is achieved by using the following settings: the number (N) of the most frequent words ranging from 50 to 700, with the analysis based on 1-gram lemmas. Although slightly less effective, analyses based on 1-gram word forms still yield high scores. When assessing the accuracy of translations from Hungarian into other languages, analyses based on 1-gram lemmas slightly underperform the original language but still achieve relatively high accuracy, reaching up to 80%. The performance of analyses based on 1-gram plain word forms is even more impressive, with accuracy reaching up to 90%, especially when using a smaller number of most frequent features (up to 500). Notably, English translations (Fig. 2) from Hungarian outperform translations into other languages, achieving the best results with a feature set of 1-gram word forms and a maximum of 500 features.

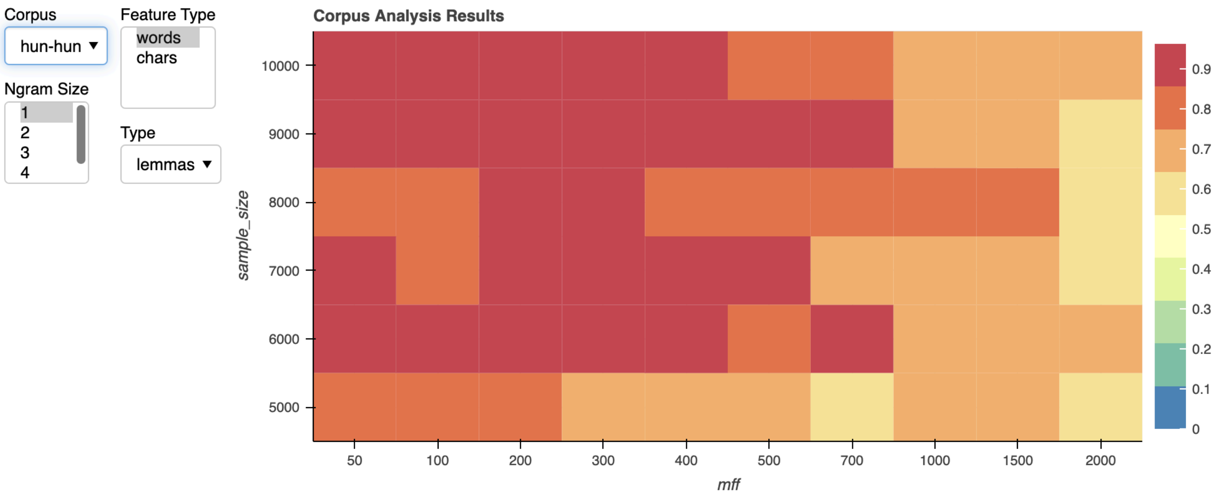
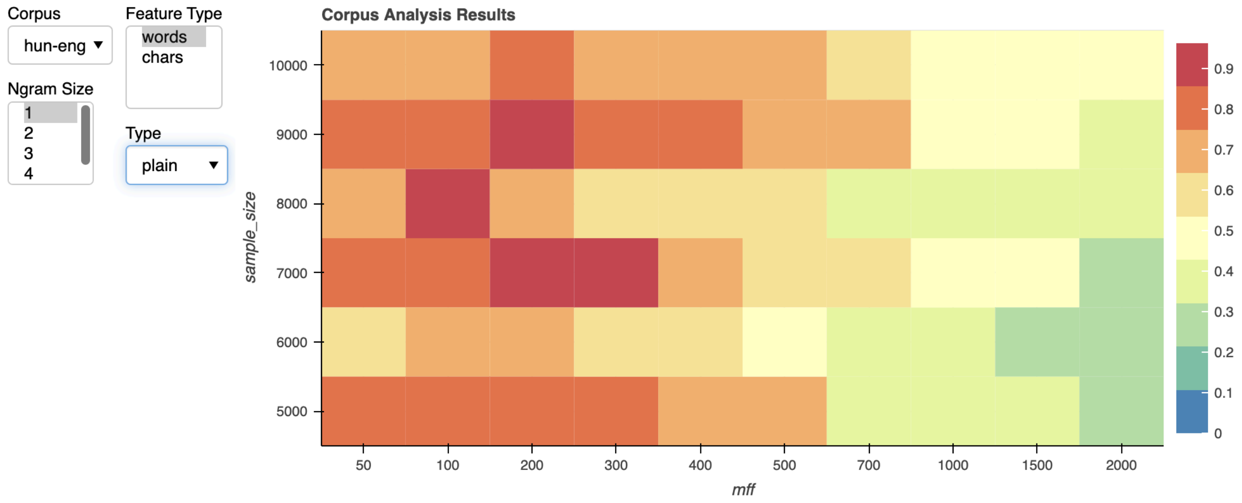

Corpus
hun-hun ▾
Feature Type
words
chars
Ngram Size
1
2
3
4
Type
lemmas ▾

Fig. 1: Results of the Hungarian corpus showing the best accuracy score across all corpora.



Corpus: hun-eng ▼

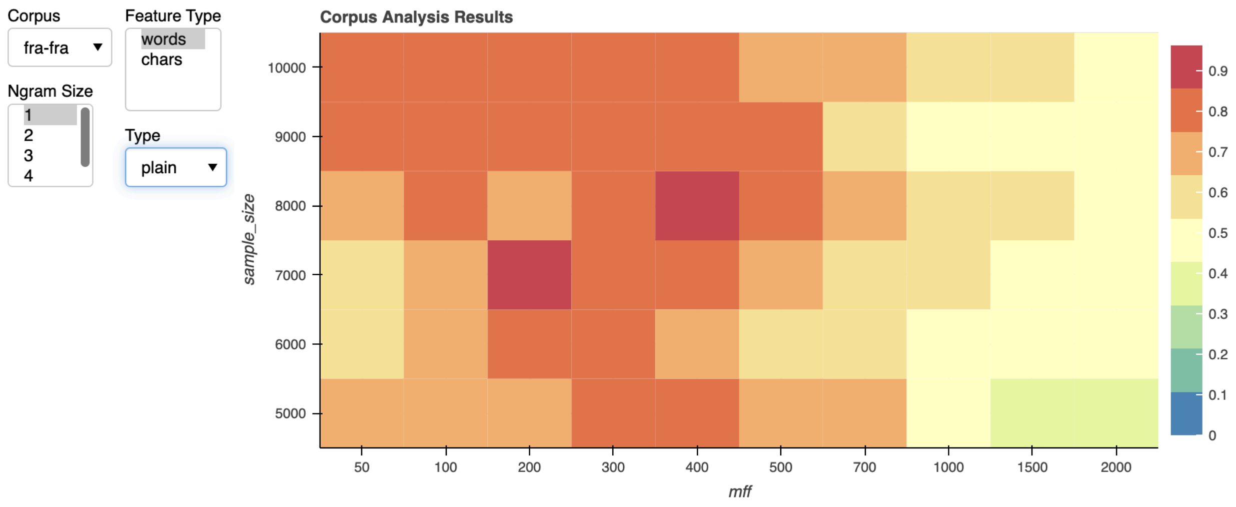
Feature Type: words, chars

Ngram Size: 1, 2, 3, 4

Type: plain ▼

Fig. 2: Performance of the Hungarian corpus translated into English.

The French corpus (Fig. 3) demonstrates robust performance, with accuracy reaching up to 80% and, in some instances, up to 90% for original texts based on 1-gram word forms with N features up to 700. Stylometric analysis based on 1-gram lemmas shows slightly lower performance. Ukrainian and Hungarian translations from French demonstrate a slight decrease in performance compared to the original texts in lemma-based analyses, with plain word form-based analyses showing even lower performance (Fig. 4, maximum accuracy around 70%). The English translation (Fig. 5), while displaying comparatively lower results in lemma-based analyses, achieves superior performance compared to Ukrainian and Hungarian translations in plain word form-based analyses.



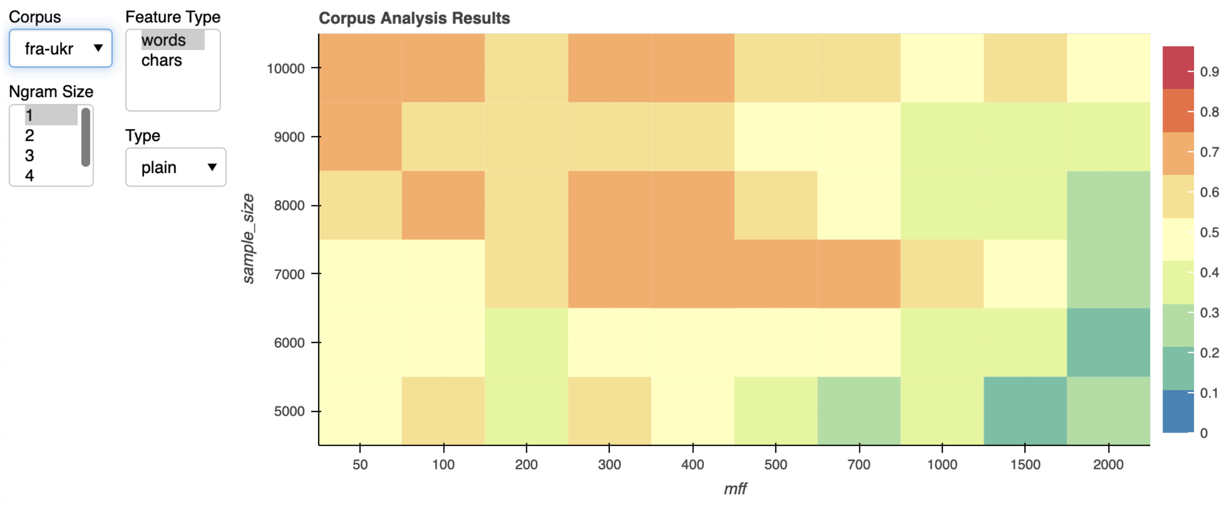
Corpus: fra-fra

Feature Type: words (selected), chars

Ngram Size: 1 (selected), 2, 3, 4

Type: plain

Fig. 3: Results of original texts in the French corpus, based on 1-gram plain word forms.



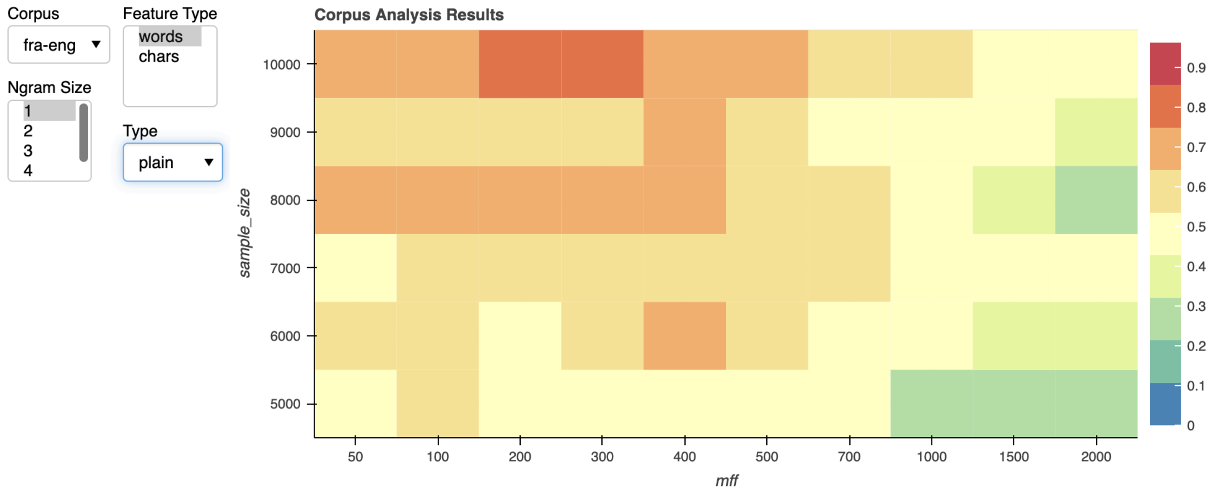
Corpus: fra-ukr

Feature Type: words (selected), chars

Ngram Size: 1 (selected), 2, 3, 4

Type: plain

Fig. 4: The performance of Ukrainian translations of French plain texts.



Corpus: fra-eng

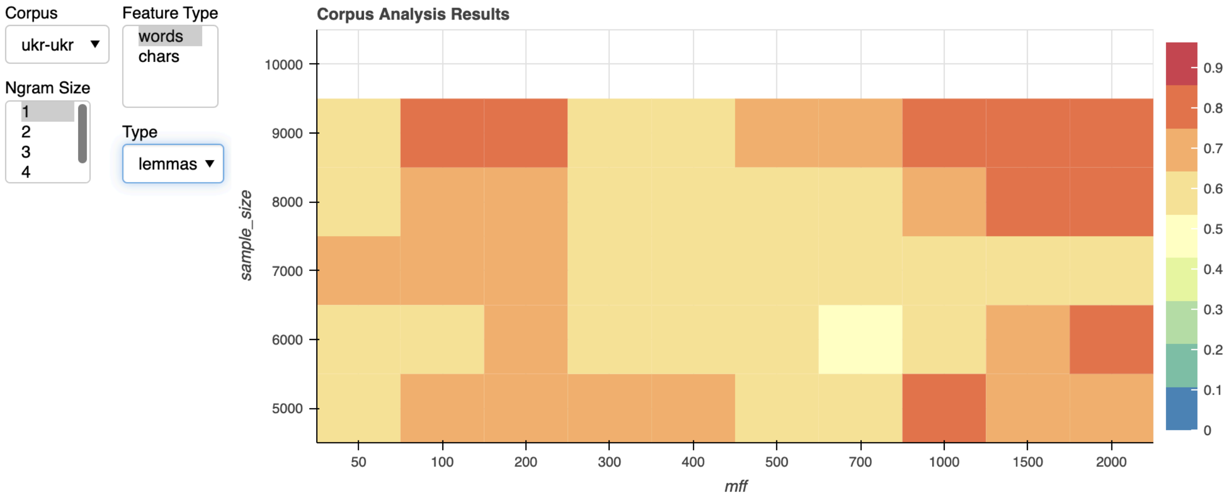
Feature Type: words

Ngram Size: 1

Type: plain

Fig. 5: Performance of the French corpus translated into English. Plain word forms results.

The Ukrainian corpus (Fig. 6) yields slightly lower performance than the French and Hungarian original corpora, achieving accuracy of up to 80% for analyses with 1-gram lemmas. Results for analyses with 1-gram plain word forms are nearly identical. Interestingly, analyses based on lemmas show higher performance when the number of features is high (ranging from 1000 to 2000), while analyses based on plain word forms yield better results with a lower number of features (up to 300). Translations of the Ukrainian corpus into other languages display significantly lower results than the original texts, with accuracy around 50% and a minimum accuracy of approximately 20%. Among the translations from Ukrainian, English translation demonstrates the highest results, reaching up to 70% (Fig. 7).



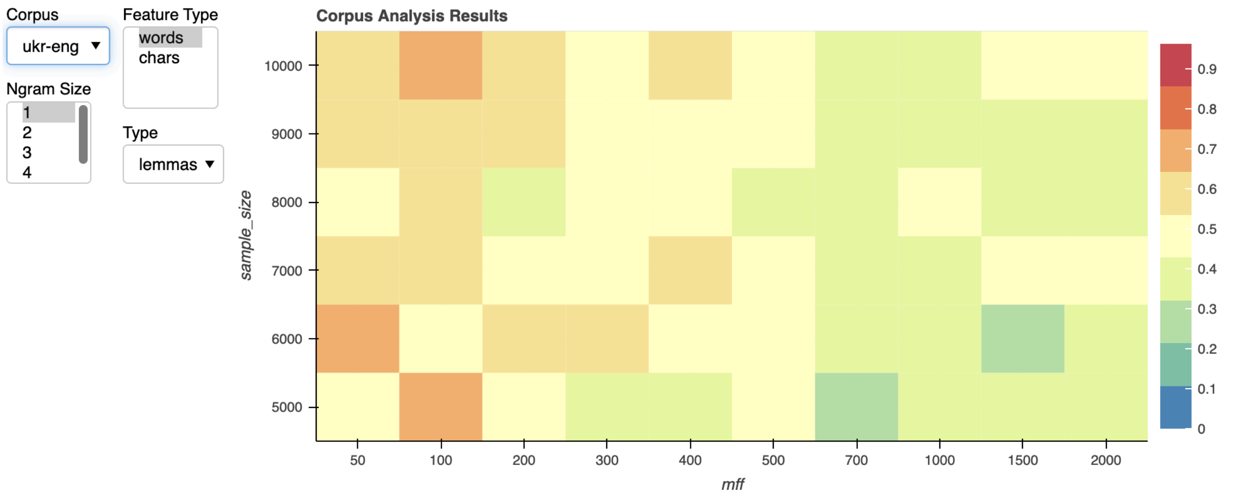
Corpus: ukr-ukr

Feature Type: words

Ngram Size: 1

Type: lemmas

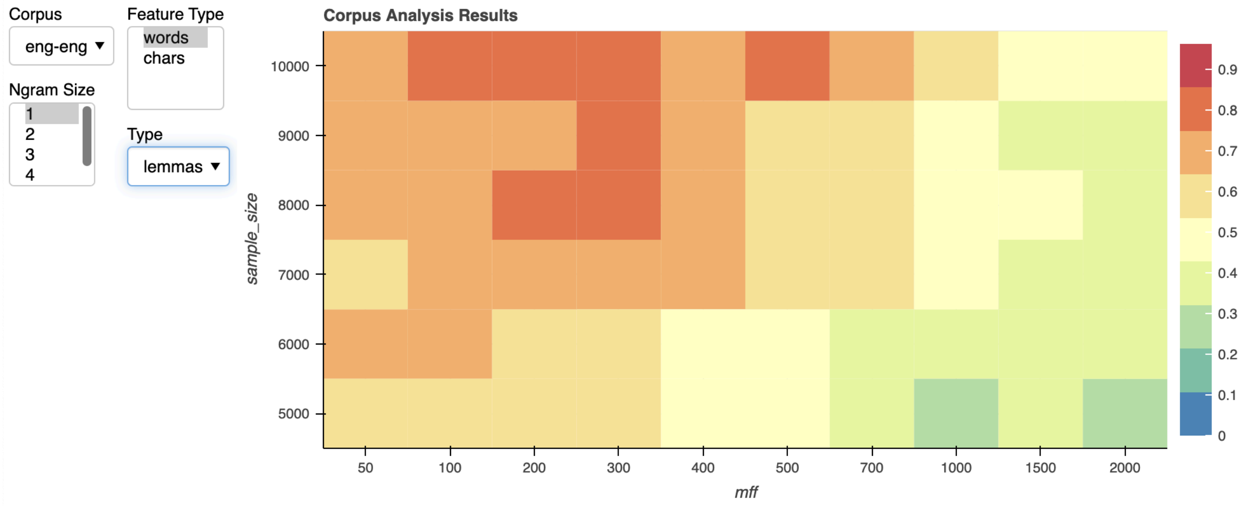
Fig. 6: Ukrainian corpus, results for 1-ngram lemmas.



Corpus: ukr-eng
Feature Type: words
Ngram Size: 1
Type: lemmas

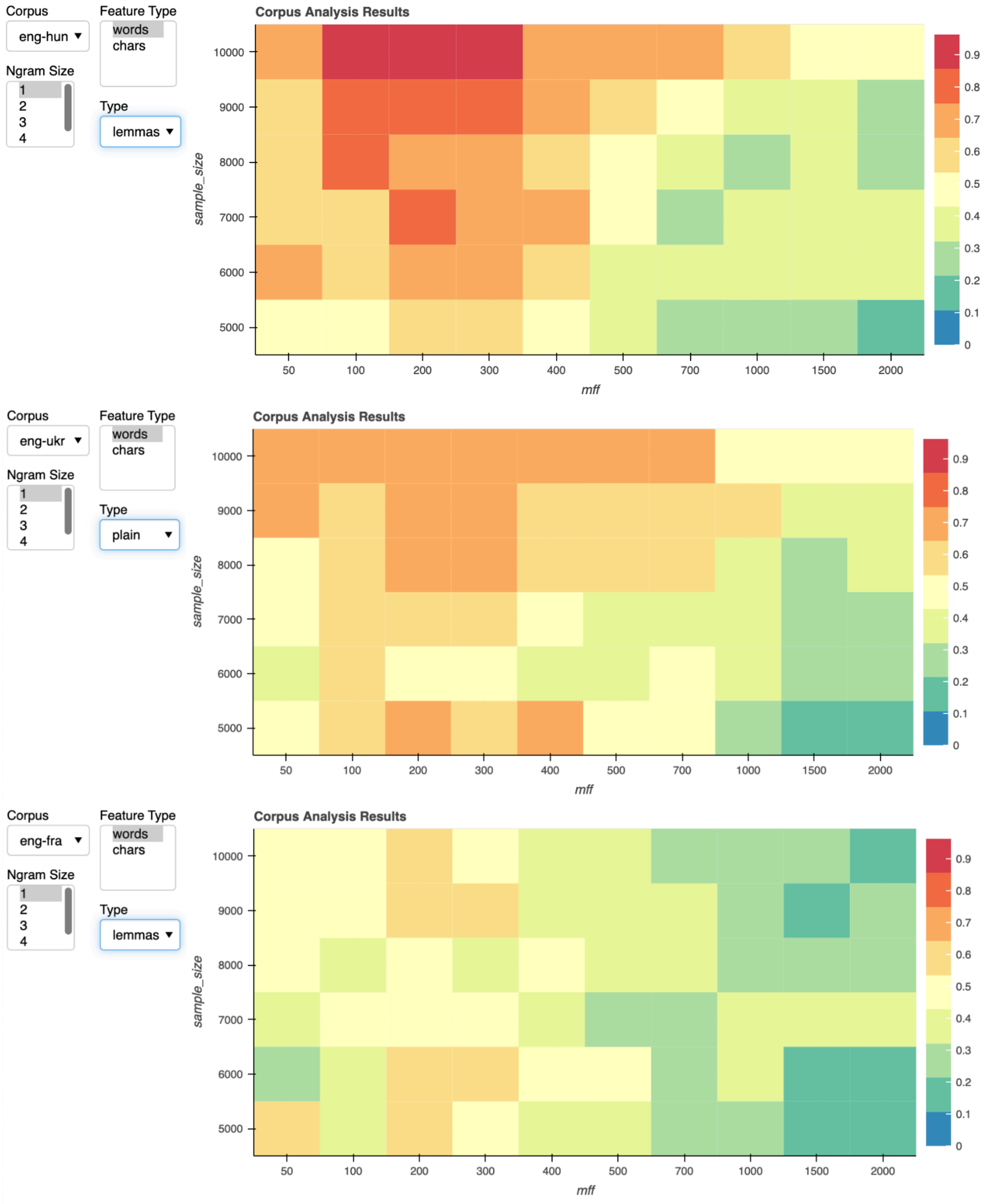
Fig. 7: English translations of Ukrainian texts, performance for 1-gram lemmas.

The English corpus (Fig. 8) achieves an average performance, with accuracy occasionally reaching 80%, but demonstrating high variability, including the lowest results around 20% for both lemma-based and plain word form-based analyses. Importantly, the Hungarian translation (Fig. 9) mirrors the performance of the original, with lemma-based analyses achieving an accuracy of up to 90% in some cases. The Ukrainian translation shows performance up to 70%, while the French translation attains maximal performance only around 60%, with most cases hovering around 20% (Fig. 9).



Corpus: eng-eng
Feature Type: words
Ngram Size: 1
Type: lemmas

Fig. 8: Results of the English original corpus. Lemma-based analysis.



Corpus: eng-hun ▾

Feature Type: words, chars

Ngram Size: 1, 2, 3, 4

Type: lemmas ▾

Corpus: eng-ukr ▾

Feature Type: words, chars

Ngram Size: 1, 2, 3, 4

Type: plain ▾

Corpus: eng-fra ▾

Feature Type: words, chars

Ngram Size: 1, 2, 3, 4

Type: lemmas ▾

Fig. 9: Performance of translations from English into Hungarian, Ukrainian and French.

Regarding the performance of each language, there are some interesting findings as well. Each language achieves its highest performance in analyzing novels presented in the original language. English attains the best results in translations from Hungarian and Ukrainian, as well as in plain word form-based analyses for translations from French. However, English demonstrates the lowest performance in the analysis of original texts compared to other languages.

The Hungarian language outperforms other languages in the analysis of original texts and nearly matches the performance of original texts in analysis based on translations from English. Alongside the Ukrainian, Hungarian demonstrates high performance in translations from French but receives low scores, similar to other translations, when analysis is based on translations from Ukrainian.

Novels in Ukrainian yield average scores in analyses based on original texts. Concerning the translations, texts translated into Ukrainian from Hungarian show slightly lower results compared to English translations. Regarding translations from French, Ukrainian texts demonstrate almost the same results as translations into Hungarian. For translations from English, Ukrainian translations yield average results and perform slightly lower than translations into Hungarian.

French performs relatively well in analyzing original texts but yields average to low results in translations from other languages.

Further information

For background on stylometric methods of authorship attribution in the context of Computational Literary Studies, see the relevant chapters on Authorship in the Survey of Methods in CLS (Schöch, Dudar, Fileva 2023).

A more extensive explanatory paper will be published in due course.

References

- Joanna Byszuk (2023): “What is Authorship Attribution?”. In: Survey of Methodological Issues in Compuational Literary Studies (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Evegniia Fileva, Julia Dudar. Trier: CLS INFRA. URL: https://methods.clsinfra.io/what-author.html.
- Jan Rybicki, Maciej Eder (2011): “Deeper Delta across genres and languages: do we really need the most frequent words?”, *Literary and Linguistic Computing*, Volume 26, Issue 3, September 2011, Pages 315–321. DOI: 10.1093/llc/fqr031.
- Christof Schöch, Roxana Patraș, Diana Santos, Tomaž Erjavec: “Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives”, *Modern Languages Open* 1/25, 2021. – DOI: 10.3828/mlo.v0i0.364.
- Christof Schöch, Julia Dudar, Evegniia Fileva, eds. (2023). *Survey of Methodological Issues in Compuational Literary Studies*. With contributions by Joanna Byszuk, Julia

Dudar, Evegniia Fileva, Andressa Gomide, Lisanne van Rossum, Christof Schöch, Artjoms Sela and Karina van Dalen-Oskam. Version 1.0.0, March 31, 2023. Trier: CLS INFRA. URL: https://methods.clsinfra.io, DOI: 10.5281/zenodo.7782363.

Detecting Small Worlds in a Corpus of Thousands of Theater Plays

URL

https://showcases.clsinfra.io/network-analysis

Repository

Code: https://github.com/dracor-org/small-world-paper

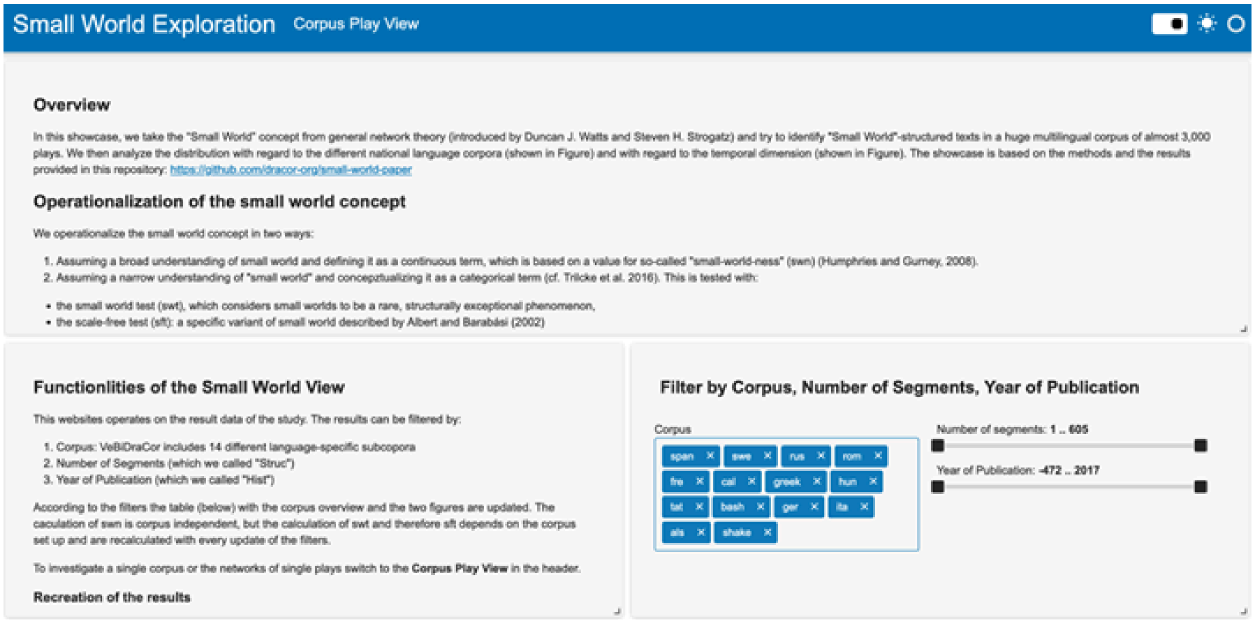
Data: https://github.com/dracor-org

Status / publication date

Showcase available online: 28.02.2024

Creators / developers / authors

- Henny Sluyter-Gäthje (Potsdam University)
- Peer Trilcke (Potsdam University)
- Evgeniya Ustinova
- Ingo Börner (Potsdam University)
- Frank Fischer (Free University Berlin)
- Carsten Milling



Overview

In this showcase, we take the "Small World" concept from general network theory (introduced by Duncan J. Watts and Steven H. Strogatz) and try to identify "Small World"-structured texts in a huge multilingual corpus of almost 3,000 plays. We then analyze the distribution with regard to the different national language corpora (shown in Figure) and with regard to the temporal dimension (shown in Figure). The showcase is based on the methods and the results provided in this repository: https://github.com/dracor-org/small-world-paper

Operationalization of the small world concept

We operationalize the small world concept in two ways:

1. Assuming a broad understanding of small world and defining it as a continuous term, which is based on a value for so-called "small-world-ness" (swn) (Humphries and Gurney, 2008).
2. Assuming a narrow understanding of "small world" and conceptualizing it as a categorical term (cf. Trilcke et al. 2016). This is tested with:

- the small world test (swt), which considers small worlds to be a rare, structurally exceptional phenomenon,
- the scale-free test (sft): a specific variant of small world described by Albert and Barabási (2002)

Functionlities of the Small World View

This websites operates on the result data of the study. The results can be filtered by:

1. Corpus: VeBiDraCor includes 14 different language-specific subcopora
2. Number of Segments (which we called "Struc")
3. Year of Publication (which we called "Hist")

According to the filters the table (below) with the corpus overview and the two figures are updated. The caculation of swn is corpus independent, but the calculation of swt and therefore sft depends on the corpus set up and are recalculated with every update of the filters.

To investigate a single corpus or the networks of single plays switch to the **Corpus Play View** in the header.

Recreation of the results

Filter by Corpus, Number of Segments, Year of Publication

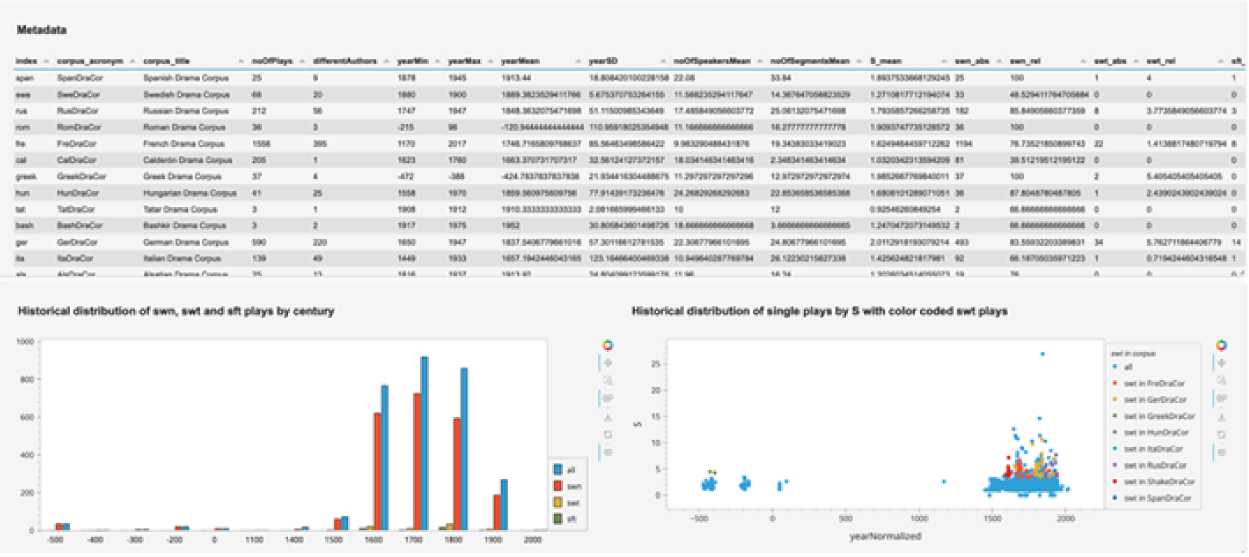
Corpus

span ✕ swe ✕ rus ✕ rom ✕ fre ✕ cal ✕ greek ✕ hun ✕ tat ✕ bash ✕ ger ✕ ita ✕ als ✕ shake ✕

Number of segments: 1 .. 605

Year of Publication: -472 .. 2017

Fig. 1: Overview over the concept of small worlds and short description how the results of the small world study can be explored in the dashboard.



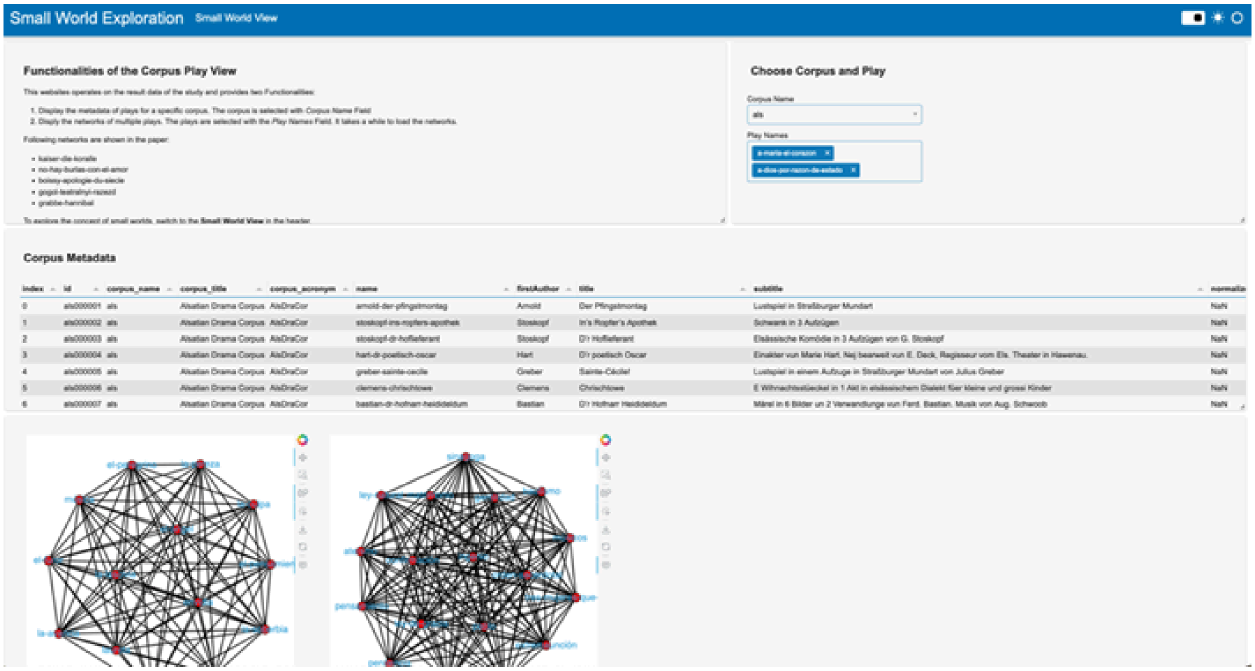
Metadata

index	corpus_acronym	corpus_title	noOfPlays	differentAuthors	yearMin	yearMax	yearMean	yearSD	noOfSpeakersMean	noOfSegmentsMean	S_mean	swn_abs	swn_rel	swt_abs	swt_rel	sft_
span	SpanDraCor	Spanish Drama Corpus	25	9	1878	1945	1913.44	18.808420100228158	22.08	33.84	1.8937533668129245	25	100	1	4	1
swe	SweDraCor	Swedish Drama Corpus	68	20	1880	1900	1889.3823529411766	5.675370753264155	11.588235294117647	14.367647058823529	1.2710817712194074	33	48.529411764705884	0	0	0
rus	RusDraCor	Russian Drama Corpus	212	56	1747	1947	1848.3632075471698	51.11500985343649	17.485849056603772	25.061320754716698	1.7935857266258735	182	85.84905660377359	8	3.7735849056603774	3
rom	RomDraCor	Roman Drama Corpus	36	3	-215	98	-120.94444444444444	110.95918025354948	11.166666666666666	16.27777777777778	1.9093747735128572	36	100	0	0	0
fre	FreDraCor	French Drama Corpus	1556	395	1170	2017	1746.7165809768637	85.56463498586422	9.983290488431876	19.34383033419023	1.6249484459712262	1194	76.73521850899743	22	1.4138817480719794	8
cal	CalDraCor	Calderón Drama Corpus	205	1	1623	1760	1663.370731707317	32.56124127372157	18.034146341463416	2.346341463414634	1.0320342313594209	81	39.51219512195122	0	0	0
greek	GreekDraCor	Greek Drama Corpus	37	4	-472	-388	-424.7837837837838	21.934416304488875	11.297297297297296	12.972972972972974	1.9852687769840011	37	100	2	5.405405405405405	0
hun	HunDraCor	Hungarian Drama Corpus	41	25	1558	1970	1859.560975609756	77.91439173236476	24.26829268292683	22.853658536585368	1.6808101289071051	36	87.8048780487805	1	2.4390243902439024	0
tat	TatDraCor	Tatar Drama Corpus	3	1	1908	1912	1910.3333333333333	2.081665999466133	10	12	0.925462603649254	2	66.66666666666666	0	0	0
bash	BashDraCor	Bashkir Drama Corpus	3	2	1917	1975	1952	30.805843601498726	18.666666666666668	3.6666666666666665	1.2470472073149532	2	66.66666666666666	0	0	0
ger	GerDraCor	German Drama Corpus	590	220	1650	1947	1837.5406779661016	57.30116612781535	22.306779661016695	24.806779661016695	2.0112918193079214	493	83.55932203389831	34	5.762711864406779	14
ita	ItaDraCor	Italian Drama Corpus	139	49	1449	1933	1657.1942446043165	123.16466400469338	10.949640287769784	26.12230215827338	1.4256248218179 81	92	66.18705035971223	1	0.7194244604316548	1
als	AlsDraCor	Alsatian Drama Corpus	25	13	1816	1937	1913.92	34.80409912159 6176	11.96	18.24	1.3026034514055073	19	76	0	0	0

Historical distribution of swn, swt and sft plays by century

Historical distribution of single plays by S with color coded swt plays

Fig. 2: Result table and two diachronic plots which are updated according to the chosen filters.



Functionalities of the Corpus Play View

This websites operates on the result data of the study and provides two Functionalities:

1. Display the metadata of plays for a specific corpus. The corpus is selected with Corpus Name Field
2. Disply the networks of multiple plays. The plays are selected with the Play Names Field. It takes a while to load the networks.

Following networks are shown in the paper:

- kaiser-die-koralle
- no-hay-burlas-con-el-amor
- boissy-apologie-du-siecle
- gogol-teatralnyi-razezd
- grabbe-hannibal

To explore the concept of small worlds, switch to the **Small World View** in the header.

Choose Corpus and Play

Corpus Name: als

Play Names: a-marte-el-corazon ×, a-dios-por-razon-de-estado ×

Corpus Metadata

index	id	corpus_name	corpus_title	corpus_acronym	name	firstAuthor	title	subtitle	normaliz
0	als000001	als	Alsatian Drama Corpus	AlsDraCor	arnold-der-pfingstmontag	Arnold	Der Pfingstmontag	Lustspiel in Straßburger Mundart	NaN
1	als000002	als	Alsatian Drama Corpus	AlsDraCor	stoskopf-ins-ropfers-apothek	Stoskopf	In's Ropfer's Apothek	Schwank in 3 Aufzügen	NaN
2	als000003	als	Alsatian Drama Corpus	AlsDraCor	stoskopf-dr-hoflieferant	Stoskopf	D'r Hoflieferant	Elsässische Komödie in 3 Aufzügen von G. Stoskopf	NaN
3	als000004	als	Alsatian Drama Corpus	AlsDraCor	hart-dr-poetisch-oscar	Hart	D'r poetisch Oscar	Einakter von Marie Hart. Nej bearweit vun E. Deck, Regisseur vom Els. Theater in Hawenau.	NaN
4	als000005	als	Alsatian Drama Corpus	AlsDraCor	greber-sainte-cecile	Greber	Sainte-Cécile!	Lustspiel in einem Aufzuge in Straßburger Mundart von Julius Greber	NaN
5	als000006	als	Alsatian Drama Corpus	AlsDraCor	clemens-chrischtowe	Clemens	Chrischtowe	E Wihnachtsstieckel in 1 Akt in elsässischem Dialekt fier kleine und grossi Kinder	NaN
6	als000007	als	Alsatian Drama Corpus	AlsDraCor	bastian-dr-hofnarr-heidideldum	Bastian	D'r Hofnarr Heidideldum	Märel in 6 Bilder un 2 Verwandlunge vun Ferd. Bastian. Musik von Aug. Schwoob	NaN

Fig. 3: Overview over the functionalities of the Corpus Play dashboard. This dashboard allows to display the subcorpus specific metadata and shows the networks of the chosen plays.

Target audience

The target audience of this showcase are scholars and students who have basic knowledge of network theory and are interested in the typology of theater plays from a network based perspective. The showcase is intended both as a research output and as a pedagogical tool that could be used in workshop or classroom settings alike.

Executive summary

With platforms like DraCor, homogenized TEI corpora of theater plays from different languages are becoming more and more available. This enables a specific approach of comparative study which is based on the method of formal network analysis and its modeling of texts as semantic structures. In this showcase, we take the “Small World” concept from general network theory and try to identify "Small World"-structured texts in a huge multilingual corpus of almost 3,000 plays.

The small world concept was introduced by Duncan J. Watts and Steven H. Strogatz. Phenomenologically described, small world networks (or more precisely: small world graphs) are, like regular networks, “highly clustered”, whereas random graphs are “poorly clustered”; at the same time, small worlds have “small characteristic path lengths, like random graphs” (Watts and Strogatz 1998, p. 440). We operationalize the small world concept in two ways. From research by Humphries and Gurney (2008), we adopt an operationalization that assumes a broad understanding of small world and defines it as a continuous term, which is based on a value for so-called "small-world-ness". This operationalization (we abbreviate it as "swn") understands small worlds as a widespread, general phenomenon. In addition, we take up an operationalization by Trilcke et al. (2016), which assumes a narrow understanding of "small world" and conceptualizes it as a categorical term. The small world test (we abbreviate it as "swt") developed by Trilcke et al. considers small worlds to be a rare, structurally exceptional phenomenon. Finally, we consider a specific variant of small world (in the sense of "swt"): the scale-free networks described by Albert and Barabási (2002), which we trace with the "sft" test.

The formal operationalization of the three types is mathematically explicated in the paper on which this showcase is based (see link below).

In this showcase, we typify the approximately 3,000 plays in our corpus with regard to the three small world concepts. We then analyze the distribution with regard to the different national language corpora and with regard to the temporal dimension.

Research questions and objectives

- Which theater plays in a huge multilingual corpus can be typified as "Small Worlds"?
- What are the descriptive benefits and limitations of different conceptualizations of dramatic small worlds?
- What is the historical distribution of the "dramatic small worlds" in our huge multilingual corpus?

Data

For our analyses we use VeBiDraCor – our "very big drama corpus", which we created by aggregating all individual corpora available through DraCor (https://dracor.org). VeBiDraCor was created on August 09, 2022 using a dedicated, fully functional Docker image of DraCor (incl. metrics services and API functions). Cf. https://github.com/dracor-org/vebidracor

Method(s)

For the study, we draw in particular on methods of social network analysis. As a first step, we model all plays as networks, where the speaking characters are set as network nodes and a relationship (edge) between two characters is set if both perform at least one speech act in a scene (or, if there are no scenes, an act). In a second step, we calculate measures from social network analysis (clustering coefficient, average path length, node degree distribution) for plays. To further analyze the values for the network analytical measures, we use simple methods of descriptive statistics.

Visualization

We use network graphs as visualizations to illustrate the structural specifics of different types of plays. In the context of descriptive statistics, we use basic diagrammatic forms such as Venn diagrams or scatter plots to illustrate different aspects of distributions.

Outcomes

- "swn" typifies 76.2% of the plays in our corpus as small worlds; "swt" typifies 2.7% of the plays in our corpus as small worlds; "sft" typifies 1.1% of the plays in our corpus as small worlds.
- Our findings provide reasons to hypothesize that the different concepts of small worlds can be located on different levels of theorizing the form of plays: On the one hand, small-world-ness "swn" could turn out to be something like a general, at least transhistorical and transnational form property of dramatic networks. On the other hand, dramatic networks of the "swt" small world type offer an approach for a network-based account to dramatic genres, with genres understood as historical forms that, accordingly, emerge under certain historical conditions – and may disappear again.
- Dramatic small worlds of the "swt" type turn out to be a primarily modern phenomenon. With the exception of two plays by Ancient Greek playwright Aristophanes, the first dramatic small worlds of the "swt" type do not appear until the end of the 16th century. These first dramatic SWT small worlds of the modern era are plays by Shakespeare. Furthermore, the dramatic scale-free networks of the SFT type prove to be a genuinely modern phenomenon. Here, too, Shakespeare is the first. So, the extraordinary impact of Shakespeare on the history of dramatic form seems to be confirmed by network-grounded analyses.

Further information

- Full study: Trilcke, P., Ustinova, E., Börner, I., Fischer, F., Milling, C. (to be published later in 2024). Detecting Small Worlds in a Corpus of Thousands of Theatre Plays. A DraCor Study in Comparative Literary Network Analysis. In Reiter, N., Andresen, M. ed. *Computational Drama Analysis. Reflecting on Methods and Interpretations*. De Gruyter. 28 p.
- Preprint: https://github.com/dracor-org/small-world-paper/blob/conference-version/Detecting_Small_World_Networks__in_a_Huge_Multilingual_Corpus_of_Theater_Plays.pdf

References

- Albert, Réka and Albert-László Barabási (2002). "Statistical Mechanics of Complex Networks". In: *Reviews of Modern Physics* 74.1, pp. 47–97. DOI: https://doi.org/10.1103/RevModPhys.74.47.
- Humphries, Mark D. and Kevin Gurney (2008). "Network 'Small-World-Ness': A Quantitative Method for Determining Canonical Network Equivalence". In: *PLoS ONE* 3.4. DOI: https://doi.org/10.1371/journal.pone.0002051.
- Trilcke, Peer, Frank Fischer, Mathias Göbel, and Dario Kampkaspar (2016). "Theatre Plays as 'Small Worlds'? Network Data on the History and Typology". In: *DH2016. Digital Identities: the Past and the Future*. Jagiellonian University, Pedagogical University of Krakow, pp. 385– 387. DOI: https://doi.org/10.5281/zenodo.6974706.
- Watts, Duncan J. and Steven H. Strogatz (1998). "Collective Dynamics of 'Small-World' Networks". In: *Nature* 393.6684, pp. 440–442. DOI: https://doi.org/10.1038/30918.

Scansion tools for corpora annotation and visualization: The Poetrylab Suite

Introducing the Poetrylab Suite: A comprehensive toolkit designed to enhance the analysis and visualization of Spanish poetry through scansion tools. At its core are three essential components: rantanplan, a Python library specialized in the scansion of Spanish poems; Averell, offering a user-friendly web interface for downloading annotated corpora; and Poetrylab, a dynamic web application enabling users to perform scansion and visualize poetic structures with precision. This suite provides researchers, scholars, and poetry enthusiasts with resources to delve into the rhythmic and metrical intricacies of Spanish poetry, fostering deeper understanding and appreciation.

Averell: A corpus management tool to transform poetic corpora into a JSON format compliant with the POSTDATA ontology (Poetrylab Suite, part 1)

Averell is an open-source tool that allows accessing many different annotated poetic corpora in a standardized way, providing them in formats ready for processing and computing.

URL

http://poetry.linhd.uned.es:1944/

Repository

https://github.com/linhd-postdata/averell

Status / publication date

Showcase available online: 01.02.2024.

Creators / developers / authors

- Aitor Díaz Medina (UNED)
- Álvaro Pérez Pozo (UNED)
- Javier de la Rosa Pérez (National Library of Norway)
- Salvador Ros Muñoz (UNED)
- Elena González-Blanco (Clibrain)

Target audience

The target audience of this showcase are scholars and students interested in poetry. It allows merging different corpora into a single new corpus based on the present metadata and granularity (e.g., verse, syllable).

Executive summary

Averell is a tool that tries to lower the barrier for researchers interested in the study of multilingual poetry corpora. It provides a unified interface to query, manage, download, and merge corpora of poetic nature in multiple languages based on features relevant for poetry scansion and meter analysis.

At its core, Averell is web application, a Python library and command line application that connects existing annotated corpora in either JSON, XML, or TEI formats, and makes them available into TEI and JSON-lines formats that can be later converted into semantic RDF according to the POSTDATA network of ontologies (González-Blanco et al., 2020).

Averell provides an easy-to-use front-end that allows users to customize the output options and it is also conceived as a command line tool for its direct use from the terminal.

The focus of this deliverable is to describe Averell which aims to be of great help for literary scholars wanting to work with poetic corpora and combine them into a unified format.

Research questions and objectives

Corpus creation has been part of the research practices of linguists and philologists for decades, and it has recently entered the computer sciences via the mixture field of natural language processing (NLP). Corpora have become a key resource in the development and evaluation of computer systems that deal with language. As these approaches from NLP are being re-discovered, applied, and enriched within the computational humanities, the making of these corpora and their transformation into structured or plain digital texts is of vital importance. Just in the literary domain, there are arguably thousands of corpora available to download or query. In a comprehensive survey, Xiao (2010) describes over a hundred well-known and highly influential corpora in English and other languages. Smaller corpora for understudied or endangered languages have also recently appeared (see Scannell 2007, Ostler 2008, Cox 2011). Notably, only five corpora in these surveys contained poetry and only one of them was annotated with relevant poetic features. As newer poetic corpora with rich annotations are becoming available, we need a proper tool to uniformly access them.

Among the characteristics that should guide the building of a corpus (McEnery and Wilson, 2001; Gries and Berez, 2015), two are commonly desired: machine readability and availability to researchers. Unfortunately, even when corpora are made fully available in electronic format, it is often the case that scholars struggle to find a proper way to address their research questions using the ready-made corpora (see e.g., Xiao, 2010).

Data

Averell has a collection of corpora whose features must meet some conditions:

1. To have metric information.
2. To be used by the community.
3. To be available online.
4. To have an open license.
5. To have a working group behind to update it.

These corpora can then be parsed into averell JSON format and be used in the corpus builder mode. If you want to build your own corpus made by one or more corpora, you have to choose the ones you want, choose the granularity and the output format.

A basic workflow of Averell is shown in Figure 1.

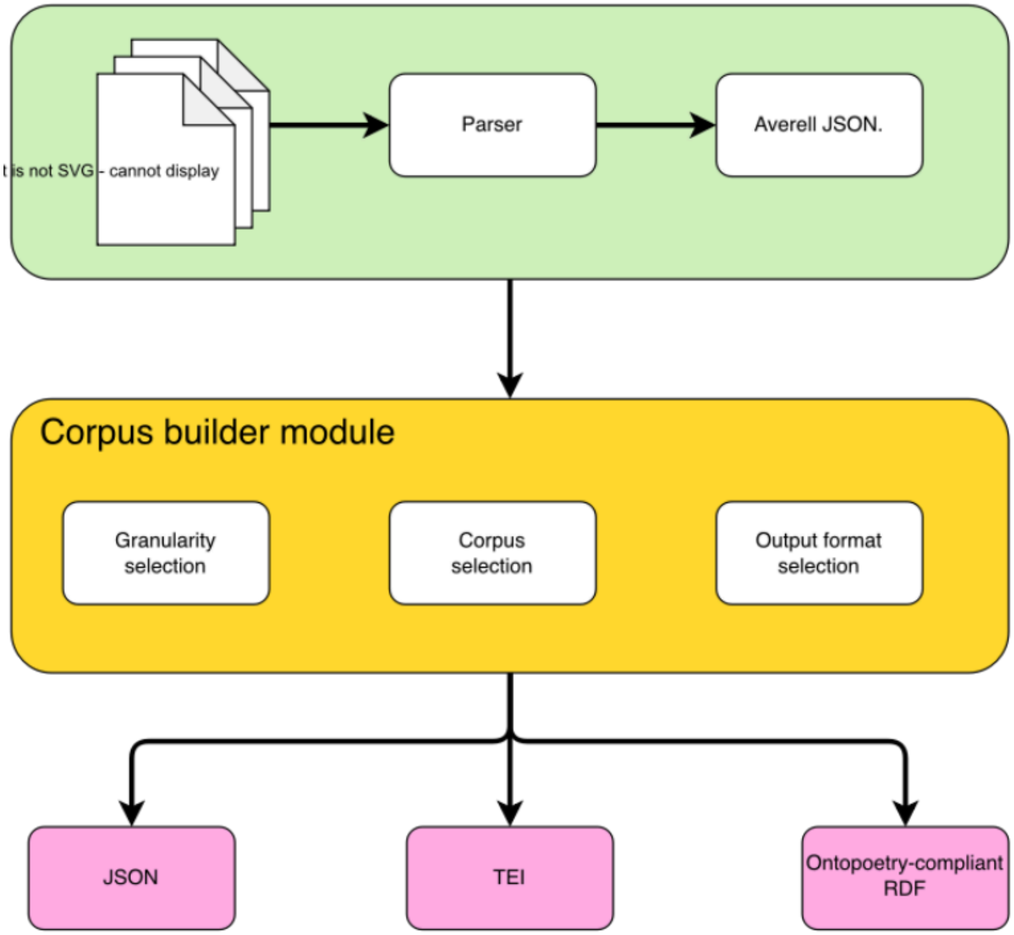


Fig. 1: Averell data flow.

Since corpora have different sizes, formats, and metrical information, we pre-processed each corpus looking for common metadata tags and structures. We then created reusable parsers to extract the relevant information exposed by Averell. The result is a JSON-lines structure capable of capturing the common details of the different corpora. From this common intermediate format, Averell can produce data in formats suitable for analysis such as JSON, XML TEI, and even POSTDATA RDF triplets.

id	name	lang	size	docs	words	granularity	license
1	Disco V2.1 (disco2_1)	es	22M	4088	381539	stanza line	CC-BY

2	Disco V3 (disco3)	es	28M	4080	377978	stanza line	CC-BY
3	Sonetos Siglo de Oro (adso)	es	6.8M	5078	466012	stanza line	CC-BY-NC 4.0
4	ADSO 100 poems corpus (adso100)	es	128K	100	9208	stanza line	CC-BY-NC 4.0
5	Poesía Lírica Castellana Siglo de Oro (plc)	es	3.8M	475	299402	stanza line word syllable	CC-BY-NC 4.0
6	Gongocorpus (gongo)	es	9.2M	481	99079	stanza line word syllable	CC-BY-NC-ND 3.0 FR
7	Eighteenth Century Poetry Archive (ecpa)	en	2400M	3084	2063668	stanza line word	CC BY-SA 4.0
8	For Better For Verse (4b4v)	en	39.5M	103	41749	stanza line	Unknown
9	Métrique en Ligne (mel)	fr	183M	5081	1850222	stanza line	Unknown
10	Biblioteca Italiana (bibit)	it	242M	25341	7121246	stanza line	Unknown

						word	
11	Corpus of Czech Verse (czverse)	cs	4100M	66428	12636867	stanza line word	CC-BY-SA
12	Stichotheque (stichopt)	pt	11.8M	1702	168411	stanza line	Unknown

Table 1. Averell available corpora.

Data preparation / preprocessing

Corpora included in averell are very heterogeneous. Some of them are in a custom format, and others use several flavours of TEI. For the custom formats we must make specific mappings, that differ from one corpus to another.

On the other hand, as TEI-corpora mapping is more standardized and precise, reusable mapping is available. The following table shows the mappings for the current TEI corpora into JSON:

Input TEI tag	Averell JSON tag
head	poem title
title	
bibl/title	

headAuthor author bibl/author	author
lg	stanzas
type	stanza_type
l	lines
met real stress	metrical_pattern
w seg	word
word Split by "|" seg	syllables

Table 2. Input TEI to JSON mapping.

Methods

After the mapping, we can re-generate the corpus in JSON format, in all the possible granularities that the original corpus has. Each level of granularity contains all the information of the granularities above them.

Granularity:

Averell is structured around two key aspects: the catalogue and its granularity. Each corpus defines a granularity level at which its documents can be split. All corpora support splitting by poems and lines (verses), but a line can also be split into words, and then syllables, for which metrical patterns might be provided. In some cases, stanzas, a set of structural and often semantical units within the poem, are also available. Extra information such as the lengths of verses, the number of lines per stanza, or the type of rhymes is also added when available. This granular annotation allows scholars to merge different corpora and extract sets of poems that meet specific criteria. For example, a corpus of hendecasyllabic sapphic verses, or poems for a specific period only at the level of the stanza. Instances of the use of Averell to carry out studies in poetry already exist.

Docker insight:

The docker creation process is straightforward. The only notable thing is that we have previously generated all TEI and JSON corpora and added it to the docker image. This is done for convenience, as some corpora are very large. Nevertheless, the ability to download, parse, and transform the corpora is still available in case you need to redo some of the previous steps.

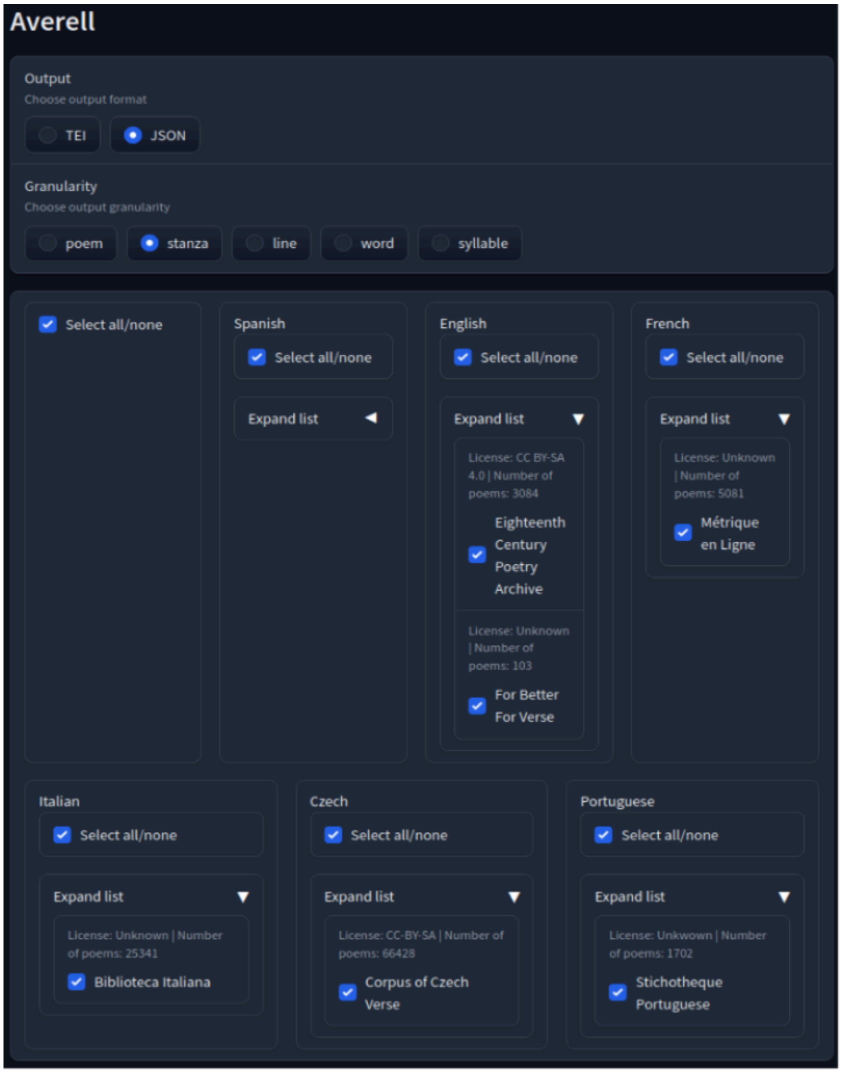
The Dockerfile is as simple as it can be:

```
FROM python:3.11


WORKDIR /usr/src/app


COPY . .
RUN pip install -e .
RUN pip install --no-cache-dir -r requirements.txt


EXPOSE 5741


CMD [ "gradio", "./app.py" ]
```

Visualization

Averell is presented as an easy-to-use interface (Figure 2) for building new corpus based on the Averell catalogue. Therefore, you can choose the input corpora, their granularity (if the corpus supports it), and finally select the output format (TEI/JSON). This application is available via the docker image. The basic usage is explained below:

1. Choose the output format.
2. Choose desired granularity.
3. Select all the corpora that you want to merge. You can choose all of them, by language, or choose specific corpora.



Averell

Output
Choose output format
○ TEI ● JSON

Granularity
Choose output granularity
○ poem ● stanza ○ line ○ word ○ syllable

☑ Select all/none

Spanish
☑ Select all/none
Expand list ◀

English
☑ Select all/none
Expand list ▼
License: CC BY-SA 4.0 | Number of poems: 3084
☑ Eighteenth Century Poetry Archive
License: Unknown | Number of poems: 103
☑ For Better For Verse

French
☑ Select all/none
Expand list ▼
License: Unknown | Number of poems: 5081
☑ Métrique en Ligne

Italian
☑ Select all/none
Expand list ▼
License: Unknown | Number of poems: 25341
☑ Biblioteca Italiana

Czech
☑ Select all/none
Expand list ▼
License: CC-BY-SA | Number of poems: 66428
☑ Corpus of Czech Verse

Portuguese
☑ Select all/none
Expand list ▼
License: Unkwown | Number of poems: 1702
☑ Stichotheque Portuguese

Fig. 2: Averell user interface.

Outcomes

Averell is built to generate two different output formats, JSON and TEI that are common formats used in the computational literary tasks.

JSON

JSON is essential for large corpus research and processing because of its ability to handle structured data, adapt to diverse data formats, and provide interoperability with a wide range of tools and technologies. Its flexibility, efficiency, and human-readable nature make it a valuable choice for managing and analyzing extensive textual datasets in research contexts.

Averell use the same set of tags explain above for the JSON output.

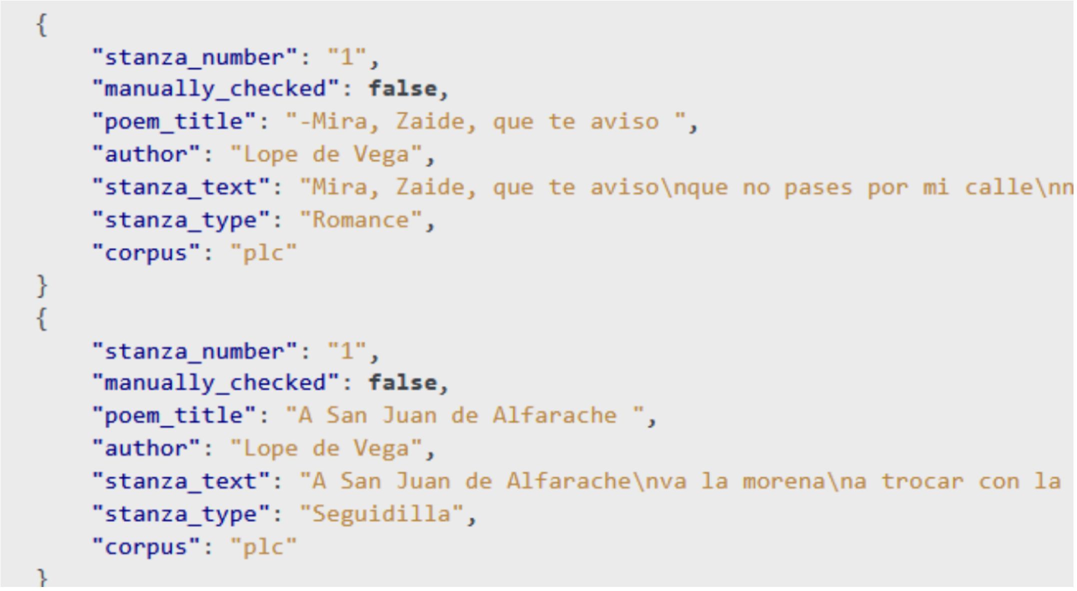
Averell Output JSON tag
poem title
author
stanzas
stanza_type
lines
metrical_pattern
word
syllables

For each of the granularities, slightly different mappings are made, adding more information the more we deepen into the structure. Common tags are **corpus**, **poem_title**, and **author**, that include the corpus name, poem title and author respectively. Then, specific granularity information is added such as the stanza information (stanza number, stanza text, and stanza number), and the equivalent information for the other granularities.

Mapping of the granularities

Stanza granularity

Figure 3 is an example of the stanza granularity. This is a list of the stanzas of the poems with simple information about the corpus it belongs to, the author, name of the poem, if it has been manually annotated or not, the stanza type (if any) and the number of the stanza within the poem.

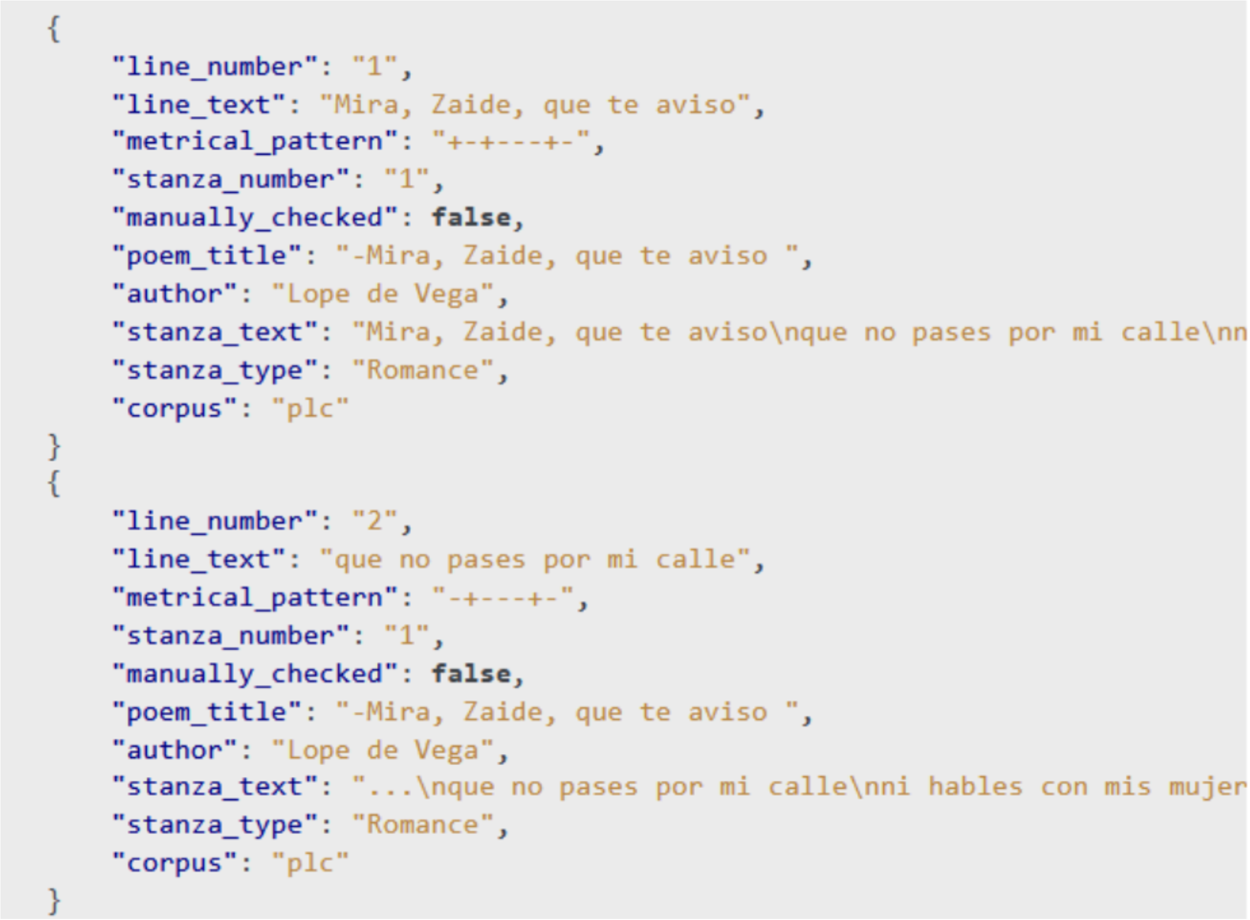


```
{
    "stanza_number": "1",
    "manually_checked": false,
    "poem_title": "-Mira, Zaide, que te aviso ",
    "author": "Lope de Vega",
    "stanza_text": "Mira, Zaide, que te aviso\nque no pases por mi calle\nn
    "stanza_type": "Romance",
    "corpus": "plc"
}
{
    "stanza_number": "1",
    "manually_checked": false,
    "poem_title": "A San Juan de Alfarache ",
    "author": "Lope de Vega",
    "stanza_text": "A San Juan de Alfarache\nva la morena\na trocar con la
    "stanza_type": "Seguidilla",
    "corpus": "plc"
}
```

Fig. 3: Stanza granularity.

Line granularity

In Figure 4 we can see another example of the JSON output, this time at the line level of granularity. These are the two first lines of a poem. On top of the previous information in Figure 3, there is now data about the metrical pattern and the number of the verse within the stanza.

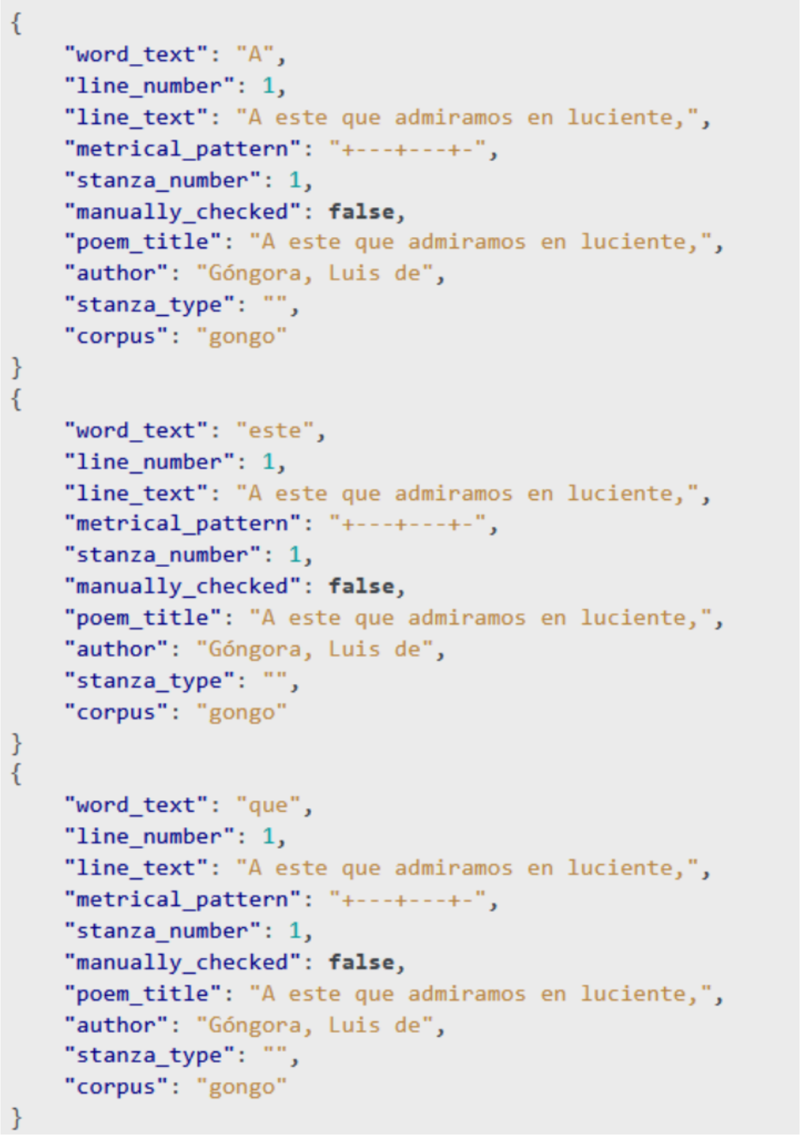


```
{
    "line_number": "1",
    "line_text": "Mira, Zaide, que te aviso",
    "metrical_pattern": "+-+---+-",
    "stanza_number": "1",
    "manually_checked": false,
    "poem_title": "-Mira, Zaide, que te aviso ",
    "author": "Lope de Vega",
    "stanza_text": "Mira, Zaide, que te aviso\nque no pases por mi calle\nn
    "stanza_type": "Romance",
    "corpus": "plc"
}
{
    "line_number": "2",
    "line_text": "que no pases por mi calle",
    "metrical_pattern": "-+---+-",
    "stanza_number": "1",
    "manually_checked": false,
    "poem_title": "-Mira, Zaide, que te aviso ",
    "author": "Lope de Vega",
    "stanza_text": "...\nque no pases por mi calle\nni hables con mis mujer
    "stanza_type": "Romance",
    "corpus": "plc"
}
```

Fig. 4: Line granularity.

Word granularity

If we go a step further in granularity as shown on Figure 5, we get the output at the word level. With all the previous information also contained here, it is possible to rebuild the whole poem by parsing the JSON.

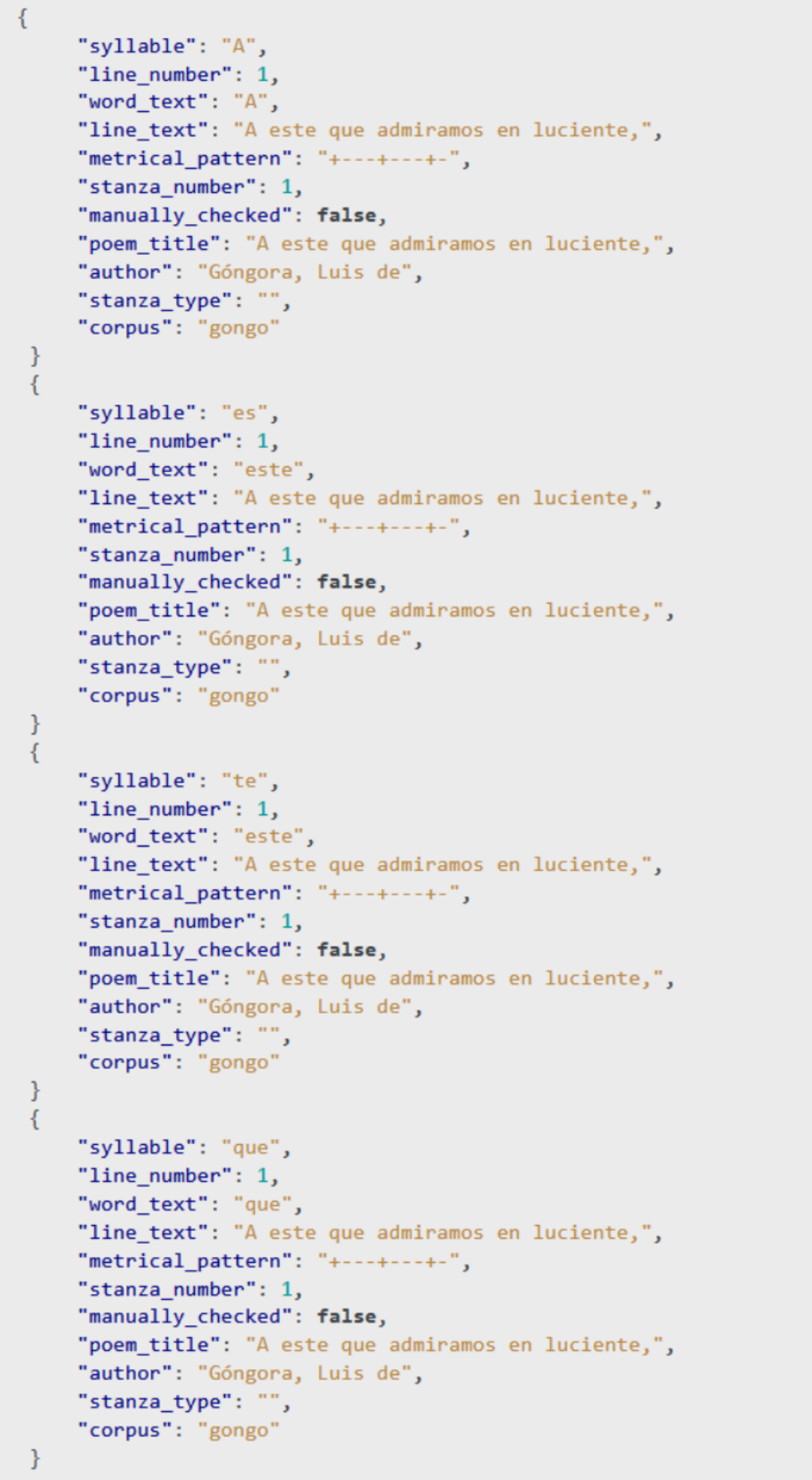


```
{
    "word_text": "A",
    "line_number": 1,
    "line_text": "A este que admiramos en luciente,",
    "metrical_pattern": "+---+---+-",
    "stanza_number": 1,
    "manually_checked": false,
    "poem_title": "A este que admiramos en luciente,",
    "author": "Góngora, Luis de",
    "stanza_type": "",
    "corpus": "gongo"
}
{
    "word_text": "este",
    "line_number": 1,
    "line_text": "A este que admiramos en luciente,",
    "metrical_pattern": "+---+---+-",
    "stanza_number": 1,
    "manually_checked": false,
    "poem_title": "A este que admiramos en luciente,",
    "author": "Góngora, Luis de",
    "stanza_type": "",
    "corpus": "gongo"
}
{
    "word_text": "que",
    "line_number": 1,
    "line_text": "A este que admiramos en luciente,",
    "metrical_pattern": "+---+---+-",
    "stanza_number": 1,
    "manually_checked": false,
    "poem_title": "A este que admiramos en luciente,",
    "author": "Góngora, Luis de",
    "stanza_type": "",
    "corpus": "gongo"
}
```

Fig. 5: Word granularity.

Syllable granularity

Figure 6 shows the poem with focus on the syllables. The output is and ordered list of dictionaries, each one with information related to the syllables of the poem. Each dictionary contains enough information to rebuild the poem as it originally was.



```
{
     "syllable": "A",
     "line_number": 1,
     "word_text": "A",
     "line_text": "A este que admiramos en luciente,",
     "metrical_pattern": "+---+---+-",
     "stanza_number": 1,
     "manually_checked": false,
     "poem_title": "A este que admiramos en luciente,",
     "author": "Góngora, Luis de",
     "stanza_type": "",
     "corpus": "gongo"
 }
 {
     "syllable": "es",
     "line_number": 1,
     "word_text": "este",
     "line_text": "A este que admiramos en luciente,",
     "metrical_pattern": "+---+---+-",
     "stanza_number": 1,
     "manually_checked": false,
     "poem_title": "A este que admiramos en luciente,",
     "author": "Góngora, Luis de",
     "stanza_type": "",
     "corpus": "gongo"
 }
 {
     "syllable": "te",
     "line_number": 1,
     "word_text": "este",
     "line_text": "A este que admiramos en luciente,",
     "metrical_pattern": "+---+---+-",
     "stanza_number": 1,
     "manually_checked": false,
     "poem_title": "A este que admiramos en luciente,",
     "author": "Góngora, Luis de",
     "stanza_type": "",
     "corpus": "gongo"
 }
 {
     "syllable": "que",
     "line_number": 1,
     "word_text": "que",
     "line_text": "A este que admiramos en luciente,",
     "metrical_pattern": "+---+---+-",
     "stanza_number": 1,
     "manually_checked": false,
     "poem_title": "A este que admiramos en luciente,",
     "author": "Góngora, Luis de",
     "stanza_type": "",
     "corpus": "gongo"
 }
```

Fig. 6: Syllable granularity.

TEI

TEI (Text Encoding Initiative) is used by a diverse range of individuals and organizations across various fields. It provides a standardized and efficient way to encode, preserve, analyze, and share historical and cultural texts. Its use contributes to the preservation of cultural heritage and the conduct of advanced research in the digital humanities. For this reason, we wanted to give Averell the possibility of exporting any corpus into TEI format.

This conversion is made by re-parsing the common JSON format and mapping each element to its corresponding TEI tag, and generating all appropriate TEI headers with the available corpus information such as author, poem name, corpus name, etc.

Input TEI tag	Averell JSON tag
head title bibl/title	poem title
headAuthor author bibl/author	author
lg	stanzas
type	stanza_type
l	lines
met real stress	metrical_pattern

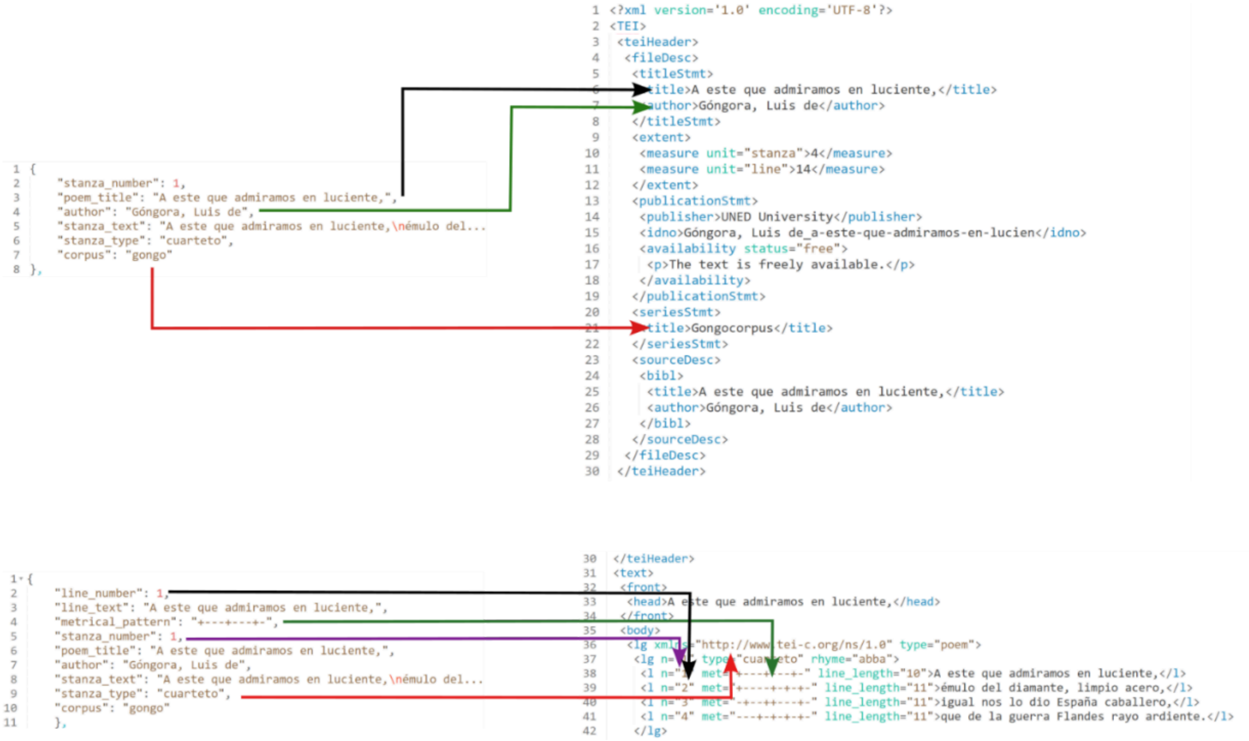
w seg	word
word Split by "|" seg	syllables

Table 3: JSON to TEI mapping.

Table 3 shows specific mappings from JSON to TEI. **author** and **title** are mapped within the **titleStmt** of the TEI header, while the corpus name uses the **title** tag of the **seriesStmt**.

For the poems structural information, we make use of the **lg** and **l** tags. The parent **lg** contains the whole poem structure while the inner **lg** tags contain the list of **l** tags representing the verses of the poem.

The scansion information is mapped as attributes of either the **lg** or **l** groups. The number of the verse and stanza use the **n** attribute in both **lg** and **l**. For **lg**, stanza type uses the **type** attribute, and the rhyme scheme uses **rhyme**. On the other hand, the **l** tag contains the metrical information and the number of syllables in the **met** and **line_length** attributes respectively.



```
{
    "stanza_number": 1,
    "poem_title": "A este que admiramos en luciente,",
    "author": "Góngora, Luis de",
    "stanza_text": "A este que admiramos en luciente,\némulo del...
    "stanza_type": "cuarteto",
    "corpus": "gongo"
},
```

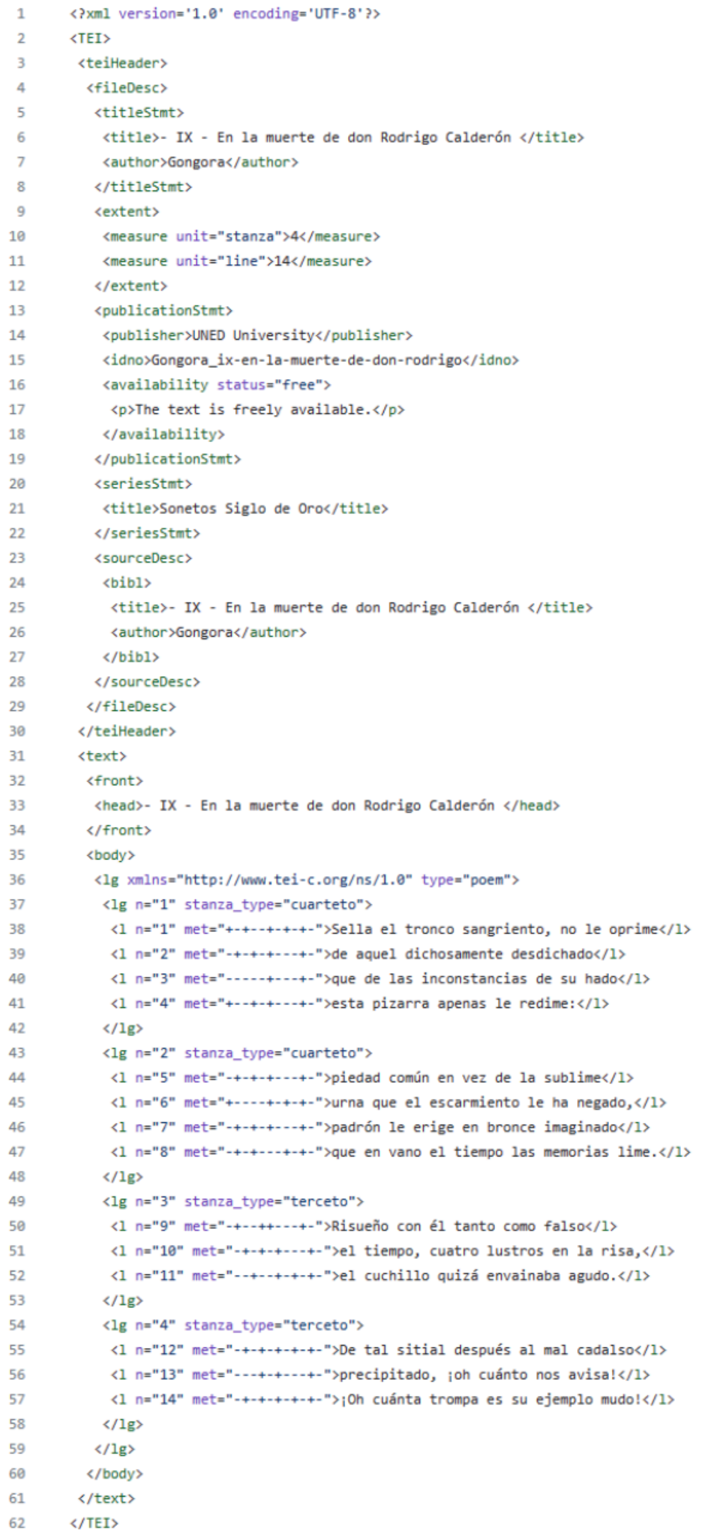
```
<?xml version='1.0' encoding='UTF-8'?>
<TEI>
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title>A este que admiramos en luciente,</title>
    <author>Góngora, Luis de</author>
   </titleStmt>
   <extent>
    <measure unit="stanza">4</measure>
    <measure unit="line">14</measure>
   </extent>
   <publicationStmt>
    <publisher>UNED University</publisher>
    <idno>Góngora, Luis de_a-este-que-admiramos-en-lucien</idno>
    <availability status="free">
     <p>The text is freely available.</p>
    </availability>
   </publicationStmt>
   <seriesStmt>
    <title>Gongocorpus</title>
   </seriesStmt>
   <sourceDesc>
    <bibl>
     <title>A este que admiramos en luciente,</title>
     <author>Góngora, Luis de</author>
    </bibl>
   </sourceDesc>
  </fileDesc>
 </teiHeader>
```

```
{
    "line_number": 1,
    "line_text": "A este que admiramos en luciente,",
    "metrical_pattern": "+---+---+-",
    "stanza_number": 1,
    "poem_title": "A este que admiramos en luciente,",
    "author": "Góngora, Luis de",
    "stanza_text": "A este que admiramos en luciente,\némulo del...
    "stanza_type": "cuarteto",
    "corpus": "gongo"
    },
```

```
 </teiHeader>
 <text>
  <front>
   <head>A este que admiramos en luciente,</head>
  </front>
  <body>
   <lg xmlns="http://www.tei-c.org/ns/1.0" type="poem">
    <lg n="1" type="cuarteto" rhyme="abba">
     <l n="1" met="+---+---+-" line_length="10">A este que admiramos en luciente,</l>
     <l n="2" met="+----+-+-+-" line_length="11">émulo del diamante, limpio acero,</l>
     <l n="3" met="-+--++---+-" line_length="11">igual nos lo dio España caballero,</l>
     <l n="4" met="---+-+-+-+-" line_length="11">que de la guerra Flandes rayo ardiente.</l>
    </lg>
```

Fig. 7: JSON to TEI mapping example.

After this conversion, we can generate a valid TEI document (Figure 8). This allows Averell to mix and unify several corpora, that were previously not in TEI, into a large, homogeneous corpus. This is a valuable tool for literary studies and the creation of critical editions because it enables precise, mass encoding of multiple large corpora in multiple languages.



```
<?xml version='1.0' encoding='UTF-8'?>
<TEI>
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title>- IX - En la muerte de don Rodrigo Calderón </title>
    <author>Gongora</author>
   </titleStmt>
   <extent>
    <measure unit="stanza">4</measure>
    <measure unit="line">14</measure>
   </extent>
   <publicationStmt>
    <publisher>UNED University</publisher>
    <idno>Gongora_ix-en-la-muerte-de-don-rodrigo</idno>
    <availability status="free">
     <p>The text is freely available.</p>
    </availability>
   </publicationStmt>
   <seriesStmt>
    <title>Sonetos Siglo de Oro</title>
   </seriesStmt>
   <sourceDesc>
    <bibl>
     <title>- IX - En la muerte de don Rodrigo Calderón </title>
     <author>Gongora</author>
    </bibl>
   </sourceDesc>
  </fileDesc>
 </teiHeader>
 <text>
  <front>
   <head>- IX - En la muerte de don Rodrigo Calderón </head>
  </front>
  <body>
   <lg xmlns="http://www.tei-c.org/ns/1.0" type="poem">
    <lg n="1" stanza_type="cuarteto">
     <l n="1" met="+-+--+-+-+-">Sella el tronco sangriento, no le oprime</l>
     <l n="2" met="-+-+-+---+-">de aquel dichosamente desdichado</l>
     <l n="3" met="-----+---+-">que de las inconstancias de su hado</l>
     <l n="4" met="+--+-+---+-">esta pizarra apenas le redime:</l>
    </lg>
    <lg n="2" stanza_type="cuarteto">
     <l n="5" met="-+-+-+---+-">piedad común en vez de la sublime</l>
     <l n="6" met="+----+-+-+-">urna que el escarmiento le ha negado,</l>
     <l n="7" met="-+-+-+---+-">padrón le erige en bronce imaginado</l>
     <l n="8" met="-+-+---+-+-">que en vano el tiempo las memorias lime.</l>
    </lg>
    <lg n="3" stanza_type="terceto">
     <l n="9" met="-+--++---+-">Risueño con él tanto como falso</l>
     <l n="10" met="-+-+-+---+-">el tiempo, cuatro lustros en la risa,</l>
     <l n="11" met="--+--+-+-+-">el cuchillo quizá envainaba agudo.</l>
    </lg>
    <lg n="4" stanza_type="terceto">
     <l n="12" met="-+-+-+-+-+-">De tal sitial después al mal cadalso</l>
     <l n="13" met="---+-+---+-">precipitado, ¡oh cuánto nos avisa!</l>
     <l n="14" met="-+-+-+-+-+-">¡Oh cuánta trompa es su ejemplo mudo!</l>
    </lg>
   </lg>
  </body>
 </text>
</TEI>
```

Fig. 8: Averell TEI output.

Further information

Deploying Averell

Averell is offered as docker image. To install and run it, follow these steps:

- Download Docker desktop from: https://www.docker.com/products/docker-desktop/
- Open the Docker desktop app.
- From the Docker Dashboard you can use Quick Search, which is in the Dashboard header, to search for:
    - Any container or Compose app on your local system. You can see an overview of associated environment variables or perform quick actions, such as start, stop, or delete.
    - Public Docker Hub images, local images, and images from remote repositories. Depending on the type of image you select, you can either pull the image by tag, view documentation, go to Docker Hub for more details, or run a new container using the image.
    - Extensions. From here, you can learn more about the extension and install it with a single click. Or, if you already have an extension installed, you can open it straight from the search results.
    - Any volume. From here you can view the associated container.
- Search for **linhdpostdata/averell-ui** and download it.
- Run the container and enter URL 127.0.0.1:5741 in your browser to access the UI.

Interoperability with Horace

RDF-POSTDATA

RDF is a foundational technology for the Semantic Web. It enables the representation of data in a way that is both structured and highly interoperable, making it an essential framework for various data-intensive applications. RDF is important for large corpus research and processing because it provides a powerful framework for representing, linking, and querying data, offering researchers the means to navigate the complexity of extensive textual datasets while facilitating interoperability and integration with other datasets and domains.

In this context, OntoPoetry is an ontology designed for the domain of poetry and literary studies. The core module of OntoPoetry has been designed to align with the FRBRoo ontology, ensuring that it can interoperate with other systems and ontologies. In this paper, the primary focus is on describing the core module, including its classes and relationships, as well as the design decisions made during its development. Additionally, the paper discusses the controlled vocabularies proposed for the core module and their connections with the other modules within the Ontopoetry ontology.

The process of exporting a corpus to RDF is made by horace. We develop this tool to map the JSON output of averell to RDF.

Table 4 shows the complete mapping, when possible, from JSON to the POSTDATA ontology OntoPoetry:

properties	Ontology classes and properties	Description
corpus	In "transmission" ontology	Corpus name
poem_title	Title	
author	agent role structure	Poem author
poem_alt_title	allternativetitle	Alternative poem title
manually_checked	Not in ontology	If the metrical information was manually annotated or not
stanzas	Poetic_metrical	List of stanzas
stanza_number	Poetic_metrical	
stanza_type	Poetic_metrical	
stanza_text	Poetic_metrical	
lines	Poetic_metrical	List of lines (of each stanza)
line_number	Poetic_metrical	
line_text	Poetic_metrical	

metrical_pattern	Poetic_metrical	
words	Poetic_metrical	List of words (of each line)
word_text	Poetic_metrical	
has_synalepha	Poetic_metrical	
syllables	Poetic_metrical	List of strings with the syllables (of each word)
corpus_license	In "transmision" ontology	
corpus_language	transmission /poetic work	
corpus_url	In "transmision" ontology	
corpus_doc_quantity	Not in ontology	
corpus_word_quantity	Not in ontology	
corpus_granularity	Not in ontology	
List of stanzas: [		
List of lines: [		
tokens	WordList	List of words (of each line)
word	SyllableList	List of syllables (of each word)

syllable	pdm:content	The text of the syllable
is_stressed	pdm:isStressed (Domain: MetricalSyllable)	Whether the syllable is stressed or not
is_word_end	pdm:isLastSyllableOf (Domain:SyllableList)	Whether the syllable is the end of a word or not
has_synalepha	Property needed from pdm:Metaplasm (i.e. pdm:affects to relate Metaplasm with two MetricalSyllables) It is also needed to specify pdm:typeOfMetaplasm	Whether or not the syllable can be conjoined with the next one
has_sinaeresis	Like the previous but including pdm:typeOfMetaplasm to Sinaeresis	Whether or not the syllable can be conjoined with the next one
stress_position	Not in ontology (covered in is_stressed at syllable level)	Index, starting from 0, for the stressed syllable of the word. If the index is negative, the syllable position is counted from the end of the word
symbol	a word with pdm:partOfSpeech equal to Symbol	If the token is a not a word, it is shown as symbol.
phonological_groups	pdm:MetricalSyllable (there must be a correspondence between metrical and grammatical using pdm:analyses)	List of phonological groups
syllable	pdm:content	The text of the syllable

is_stressed	Not in ontology	Whether the syllable is stressed or not.
is_word_end	pdm:isLastMetricalSyllableOf (domain MetricalSyllable and range the MetricalSyllableList) Inferred by structure	Whether the syllable is the end of a word or not.
synalepha_index	Not in ontology	The index of the character where the syllable is conjoined with the next one:
rhythm		JSON meta-key where we find information about the verse itself on the rhythm key
stress	pdm:patterningMetricalScheme.	Pattern of the unstressed (-) and stressed (+) syllable. This output can be changed with the parameter rhythm_format
type	We have pdm:MetricalEncoding	Metrical Encoding for the patterns
length	We only have pdm:syllabicMetricalScheme for a pdm:LinePattern	Proposed length for the verse
length_range	We also have pdm:altSyllabicMetricalScheme for a pdm:LinePattern	Minimum and maximum verse length possible. This is calculated considering all possible sinaeresis and synalephas.

in_length	Not in ontology	
max_length	Not in ontology	
structure	pdm:typeOfStanza	The name of the stanza that has been detected
rhyme	pdm:rhymeLabel	A letter code to match rhyming verses.
rhymescheme	pdm:rhymeScheme	Letters showing rhymes in a stanza
ending	pdm:rhymeGrapheme	What part of the last word is rhyming
ending_stress	Not in ontology	Negative index (-1 for last, -2 for penultimate, etc.) for the vowel that carries the stress of the rhyming part
rhyme_type	pdm:typeOfRhymeMatching	Whether the rhyme is consonant or assonant
rhyme_relaxation	Not in ontology	Whether rules for rhyme relaxation are applied. For example, removing weak vowels on diphthongs or making letters match when they are pronounced the same, for example c and z

Table 4: JSON to OntoPoetry Mapping.

Command line

Averell can be installed with **pip install averell.** This is the preferred method to install averell, as it will always install the most recent stable release. The complete and official documentation can be found at https://averell.readthedocs.io/en/latest/.

Poetrylab + rantanplan: Using rantanplan for accurate scansion analysis and visualization (Poetrylab Suite, part 2)

Poetrylab, in conjunction with rantanplan, offers a comprehensive platform for precise scansion analysis of Spanish poetry. Built upon the foundation of advanced computational linguistics, Poetrylab harnesses the capabilities of rantanplan, a specialized Python library tailored specifically for the scansion of Spanish poems. Through meticulous algorithms and linguistic analysis, rantanplan facilitates the identification and categorization of syllabic patterns, meter, and rhythmic structures within Spanish poetic compositions. Integrated seamlessly within Poetrylab's user-friendly interface, rantanplan empowers users to conduct detailed and accurate scansion analysis, providing valuable insights into the metrical intricacies and poetic nuances of Spanish verse.

URL

Poetrylab app: http://poetry.linhd.uned.es:3000/

Rantanplan API: http://poetry.linhd.uned.es:5000/ui/

Repository:

https://github.com/linhd-postdata/rantanplan

Status / publication date:

Showcase available online: 28.02.2024.

Creators / developers / authors

- Javier de la Rosa Pérez (National Library of Norway)
- Álvaro Pérez Pozo (UNED)
- Aitor Díaz Medina (UNED)
- Salvador Ros Muñoz (UNED)
- Elena González-Blanco (Clibrain)

Target audience

The target audience for Poetrylab, with its incorporation of rantanplan, encompasses a diverse spectrum of individuals ranging from scholars and researchers to students and poetry enthusiasts. Researchers in linguistics, literature, and cultural studies benefit from its sophisticated tools for in-depth analysis of Spanish poetic forms, enabling them to explore the nuances of meter, rhyme, and structure within various literary traditions. Students of Spanish literature and language find Poetrylab invaluable for enhancing their understanding and appreciation of poetry, offering a hands-on approach to studying poetic techniques and forms. Furthermore, poets and writers seeking inspiration or validation in their own creative endeavors utilize Poetrylab as a resource for studying and analyzing established poetic works. Its intuitive interface and robust functionality make it accessible to users with varying levels of expertise, fostering a vibrant community of individuals passionate about Spanish poetry and literary analysis.

Executive summary

Poetrylab presents a suite of tools for the analysis and visualization of Spanish poetry. This comprehensive platform is designed to cater to a diverse audience, including scholars, researchers, students, and poetry enthusiasts. At the heart of Poetrylab lies rantanplan, a specialized Python library meticulously crafted for the scansion of Spanish poems. Leveraging advanced computational linguistics algorithms, rantanplan facilitates the identification and categorization of syllabic patterns, meter, and rhythmic structures within Spanish poetic compositions. Users can engage with Poetrylab's user-friendly interface to conduct detailed scansion analyses, gaining valuable insights into the metrical intricacies and poetic nuances of Spanish verse. Whether exploring established literary traditions, studying poetic techniques, or seeking inspiration for creative endeavors, Poetrylab serves as an invaluable resource, fostering deeper understanding and appreciation of Spanish poetry among its users.

Research questions and objectives

Automatic analysis of poetry is a field in expansion. Implementing the investigation of verse by using automatic tools is becoming more and more common and is allowing for new and more efficient approaches to unanswered theoretical questions.

The development of new technologies of natural language processing and of text analysis has greatly contributed to the growth of the field of automatic poetry analysis; also, it has led to the emergence of multiple and varied tools. Each of them addresses one or more aspects of verse and is built with a different purpose, for instance, metrical annotation or poetry visualization. Most tools analyze English; however, other languages and language varieties are the focus of several programs.

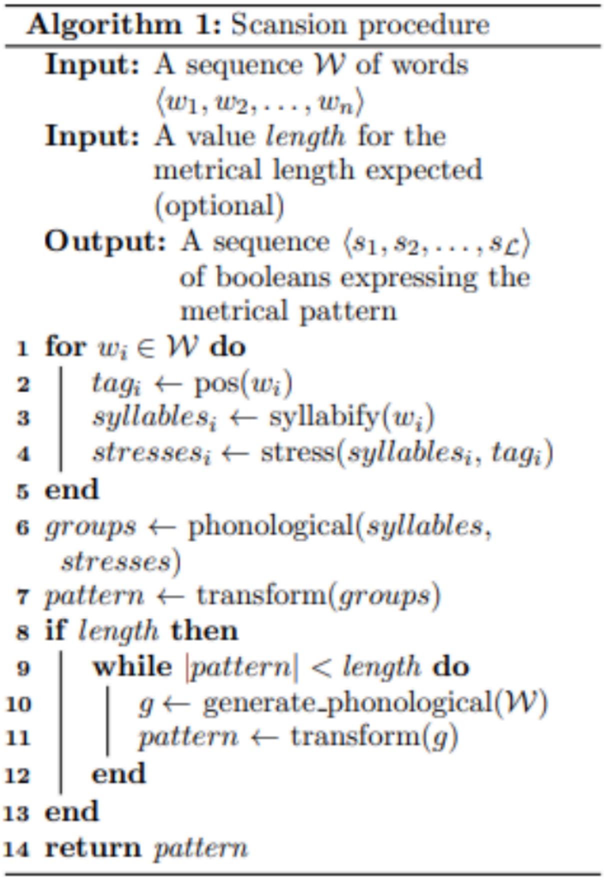
What is often missing is a connection between these tools and theoretical scholars. These tools can be a great mean of addressing theoretical questions by supporting them with large data analysis; nevertheless, too often the field of automatic tool developments and of theoretical research are not well connected. In this context we introduce rantanplan, aiming to bridge this gap and provide a comprehensive solution for scansion analysis in Spanish poetry. We also present the visual interface Poetrylab, designed to complement rantanplan by providing an intuitive platform for users to engage with the results of scansion analysis.

Methods

Rantanplan, which is comprised of four modules that work together to perform scansion of both fixed-meter as well as mixed-meter poetry: POS tagger, syllabification, stress assignment, and metrical adjustment. The general algorithm, described in Figure 9 Algorithm for *Rantanplan*, operates at the line level with a sequence of words. First, for each word in a line of verse the POS information is extracted, and the word split into syllables (lines 2-3 in algorithm 1). Combining the POS information and the syllabified word, the stress for each syllable is assigned according to the rules for oxytone, paroxytone, and proparoxytone words, plus a few exceptions detailed below (line 4).

In the process, all possible synalephas and synereses are marked at the syllable level. With the enriched syllabic data, a new sequence of phonological groups is created by applying all possible synalephas and Rantanplan, fast and accurate syllabification and scansion of Spanish poetry synereses and keeping the information about the stress positions (line 6). This sequence of phonological groups is translated directly into a metrical pattern (line 7), since each phonological group represents a prosodic unit of pronunciation. The only consideration to factor in is the stress of the ending word, so an extra symbol could be added or subtracted accordingly when necessary. From here, two situations can occur:

1. The expected metrical length is not known, in which case the calculated pattern is returned (line 14).
2. The expected metrical length is known and its value greater than the length of the calculated pattern (lines 8-13). This means some of the applied synalephas and synereses must be undone until both lengths match. The metrical adjustment module will try every option iteratively giving priority based on a heuristic. For each attempt, a new metrical pattern and its corresponding length is calculated and checked against the expected metrical length. If no match is found, the last pattern calculated is returned.



```
Algorithm 1: Scansion procedure
  Input: A sequence $\mathcal{W}$ of words $\langle w_1, w_2, \ldots, w_n \rangle$
  Input: A value $length$ for the metrical length expected (optional)
  Output: A sequence $\langle s_1, s_2, \ldots, s_{\mathcal{L}} \rangle$ of booleans expressing the metrical pattern
  for $w_i \in \mathcal{W}$ do
      $tag_i \leftarrow \mathrm{pos}(w_i)$
      $syllables_i \leftarrow \mathrm{syllabify}(w_i)$
      $stresses_i \leftarrow \mathrm{stress}(syllables_i, tag_i)$
  end
  $groups \leftarrow \mathrm{phonological}(syllables, stresses)$
  $pattern \leftarrow \mathrm{transform}(groups)$
  if $length$ then
      while $|pattern| < length$ do
          $g \leftarrow \mathrm{generate\_phonological}(\mathcal{W})$
          $pattern \leftarrow \mathrm{transform}(g)$
      end
  end
  return $pattern$
```

Fig. 9: Algorithm for Rantanplan.

POS Tagging

We built Rantanplan on top of the industrial strength NLP framework spaCy for speed (Honnibal & Montani, 2017). As mentioned previously, in Spanish some words are stressed depending on their function in the sentence, hence the need for a proper part of speech tagger. AnCora (Delor et al., 2008), the gold standard corpus many modern statistical language models are trained on for POS tagging of Spanish texts, splits most affixes thus causing some failures in the tags it assigns on prediction. To circumvent this limitation and to ensure clitics3 were handled properly, we integrated Freeling's affixes rules via a custom-built pipeline for spaCy. The resulting package, spacy_affixes, splits words with affixes before assigning POS, and can be plugged in to a regular spaCy pipeline loading one of the statistical models for Spanish. In our approach, only suffixes on verbs are enabled in spacy_affixes to guarantee clitics are handled adequately by spaCy and POS tags are assigned correctly.

Syllabification

Our method then follows a rule-based algorithm inspired by Ríos Mestre (Mestre, 1998), Caparrós (Domínguez Caparrós, 2014) and Navarro Tomás (Navarro Tomás, 1918) to split words into syllables. The procedure relies heavily on regular expressions to extract the letter groups that form the syllables. It is comprised of three steps:

1. Pre-syllabification rules are applied, which include the detection of consonant groups other than double ‘l’, as in ‘aislar’, and the handling of the prefixes ‘sin- ’ and ‘des-’ when followed by consonants, as in ‘deshielo’.
2. Regular letter clusters are identified and separated from the rest.
3. Post-syllabification exceptions for consonant clusters and diphthongs are applied.

Apart from the official rules for syllabification, there are cases with more than one correct way to proceed. The first of these cases was the ‘tl’ group. Let’s take the word ‘atlántico’ for example, its syllabification changes according to the territory. We decided not to split the group ‘tl’ since most of the Spanish speakers around the world do not separate it. In the case of words of Nahuatl origin this separation should not be made either. Compound words and words with an ‘h’ in between were also challenging.

As an example of the former let’s take the word ‘reutilizar’. Although intuitively it may seem that the prefix ‘re-’ should be separated from the rest of the word, the Fundéu recommends not to do it this way, splitting instead as ‘reu-ti-li-zar’. Similarly, the ‘h’ in a middle position does not split diphthongs, so ‘desahijar’ would be syllabified as ‘de-sahi-jar’, which might feel odd at a first pass, but it agrees with the pronunciation of the word. Moreover, we also included possible diereses as part of our alternative syllabification exceptions. One such word is ‘hiato’7 which can be split either as ‘hia-to’ or ‘hi-ato’.

As noted by Navarro-Colorado (Navarro-Colorado, 2017), another common case is the word ‘suave’, which poets tend to apply dieresis to thus resulting in ’sua-ve’ instead of the default split as ‘su-a-ve’. Therefore, our method relies on a list of words with alternative syllabifications compiled from Ríos Mestre. These alternatives are only considered by the metrical adjustment module.

Stress assignment and phonological groups

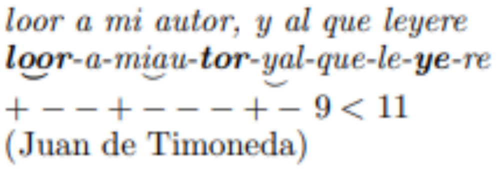
Once syllables and part of speech of a word are extracted, stress can be assigned. The assignment of stress follows very closely the rules defined by the Real Academia Española (RAE), adding exceptions for certain parts of speech, consonant groups, and words that are usually stressed but are not for metrical reasons.

The sequence of phonological groups that will be used to extract the metrical pattern is calculated by applying all possible synereses and synalephas to the list of syllables of words per line and propagating the stress information when needed. For example, the words ‘me ama’ are split into the syllables ‘me-ama’, and after applying synalepha the resulting phonological groups, ‘mea-ma’, keep the stress in place.

Word ends are also marked since they are needed to adjust the length of the metrical pattern according to the position of the stress of the last word. Phonological groups are then transformed into a metrical pattern representation and returned if the expected metrical length of the verse is not known beforehand.

Metrical adjustment

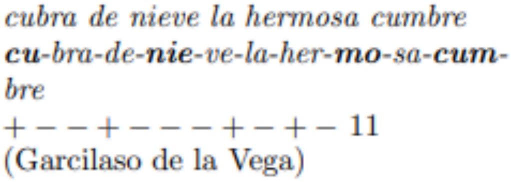
However, there are situations where the expected metrical length is known, such as processing a corpus of sonnets which tend to be hendecasyllables. In cases like this, verses with applied synalephas or synereses but a metrical length lower than the expected would trigger the adjustment module. In Figure 10, the expected metrical length is 11 but our system returns 9, thus triggering the metrical adjustment module.



loor a mi autor, y al que leyere
loor-*a-miau*-***tor***-*yal-que-le*-***ye***-*re*
$+ - - + - - - + -$ 9 $<$ 11
(Juan de Timoneda)

Fig. 10: Example 2.

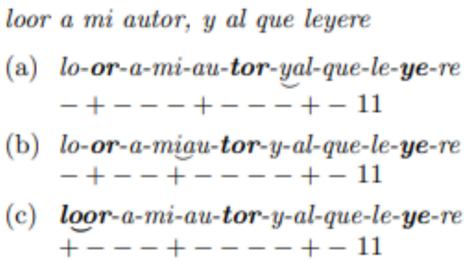
This means that 11 − 9 = 2 of the applied synalephas and synereses must be undone until both lengths match. The metrical adjustment module tries every possible metrical pattern combining synereses, synalephas, and alternative syllabifications. Priority is given to keep the synalephas since they are rarely broken, and syneresis are usually undone. The same happens for the alternative syllabifications, which deals with dieresis and adds more combinations to check for. A special case adding possibilities to the search space is the handling of synalephas between words with an initial ‘h’ and vowel ending words. Up to the 16th century, the initial ‘h’ in words was aspired instead of silent. This depends on the etymology of some words. For example, in the verse ‘cubra de nieve la hermosa cumbre’ (see example 3) there should not be synalepha between ‘la’ and ‘hermosa‘ since ‘hermosa’ evolved from the Latin ‘fermosa’ and as such a synalepha was not possible at all. To this day, this remains an option to the author, who can decide whether to apply a synalepha in such cases.



cubra de nieve la hermosa cumbre
cu-*bra-de*-***nie***-*ve-la-her*-***mo***-*sa*-***cum***-
bre
$+ - - + - - - + - + -$ 11
(Garcilaso de la Vega)

Fig. 11: Example 3.

For each attempt, a new metrical pattern and length is calculated and checked against the expected metrical length. If no match is found, the last pattern calculated is returned. For the verse in example 2, the generated possible metrical patterns are shown in example 4. Pattern 4 a, with no synereses and one synalepha between ‘y’ and ‘al’ would be generated first and returned afterwards. Since the metrical pattern has the correct length, it is returned as such and the generation stops, saving the time it takes to generate any other possible pattern. This is also a limitation of our approach since more than one correct metrical pattern can be generated that matches the desired length.



loor a mi autor, y al que leyere

(a) *lo-**or**-a-mi-au-**tor**-yal-que-le-**ye**-re*
− + − − − + − − − + − 11

(b) *lo-**or**-a-miau-**tor**-y-al-que-le-**ye**-re*
− + − − + − − − − + − 11

(c) ***loor**-a-mi-au-**tor**-y-al-que-le-**ye**-re*
+ − − − + − − − − + − 11

Fig. 12: Example 4.

Visualization

PoetryLab interface is presented in a minimalist design. The central feature on the page shown in Figure 13 is a large dashed-bordered box with an icon indicating the functionality to drag and drop a file for the supported formats .doc or .txt, or to upload one. You also have the option to write or paste your own poem.

The interface's simplicity is intended to be user-friendly, allowing for easy navigation and interaction without overwhelming the user with too many options or complex instructions.

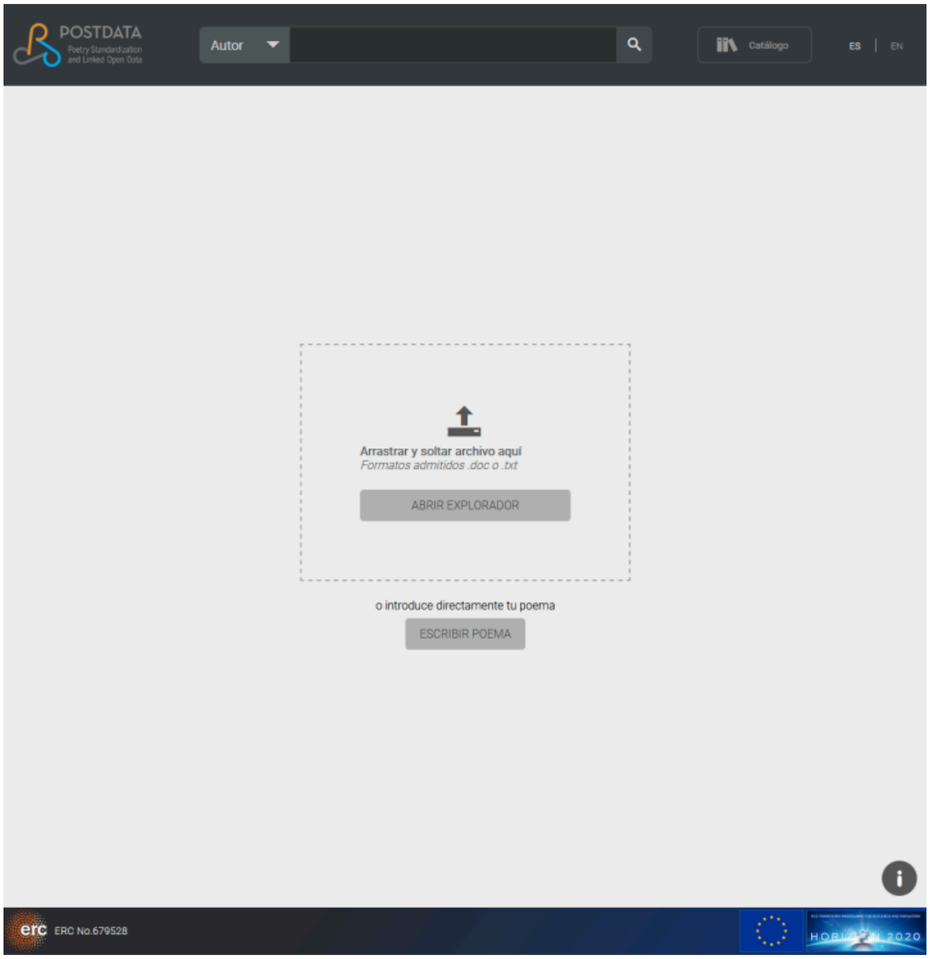
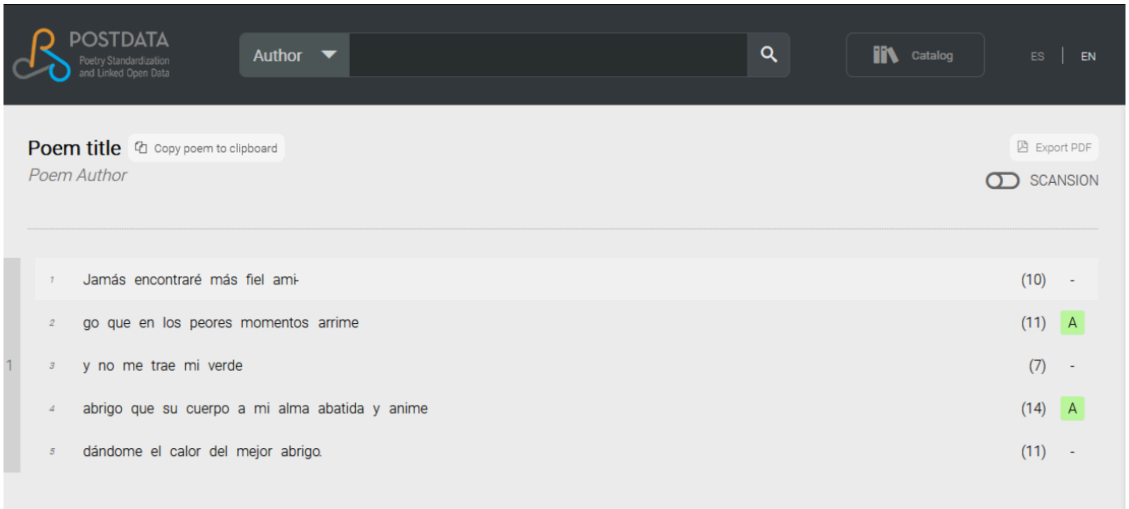

POSTDATA
Poetry Standardization and Linked Open Data
Autor
Catálogo
ES | EN
Arrastrar y soltar archivo aquí
Formatos admitidos .doc o .txt
ABRIR EXPLORADOR
o introduce directamente tu poema
ESCRIBIR POEMA
erc ERC No.679528
HORIZON 2020

Fig. 13: PoetryLab home screen.

Once the poem is processed, we can see an advanced scansion and enjambment visualization as shown in Figure 14.

This image displays a basic poetry scansion visualization interface from the PoetryLab web application. The verses are numbered for reference. Each line is followed by a number in parentheses that indicates the total syllable count for that line. Lines are marked with a letter next to the syllable count, which indicates rhyming verses.

This visualization is designed to provide a clear and straightforward presentation of the scansion of a poem, giving users a quick understanding of its metric structure.



Poem title
Poem Author

Jamás encontraré más fiel ami-	(10)	-
go que en los peores momentos arrime	(11)	A
y no me trae mi verde	(7)	-
abrigo que su cuerpo a mi alma abatida y anime	(14)	A
dándome el calor del mejor abrigo.	(11)	-

Fig. 14: Simple scansion and enjambment view.

If we switch the "SCANSION" toggle at the top right corner, we can switch to the advanced, and more detailed, scansion and enjambment visualization, as shown in Figure 15.

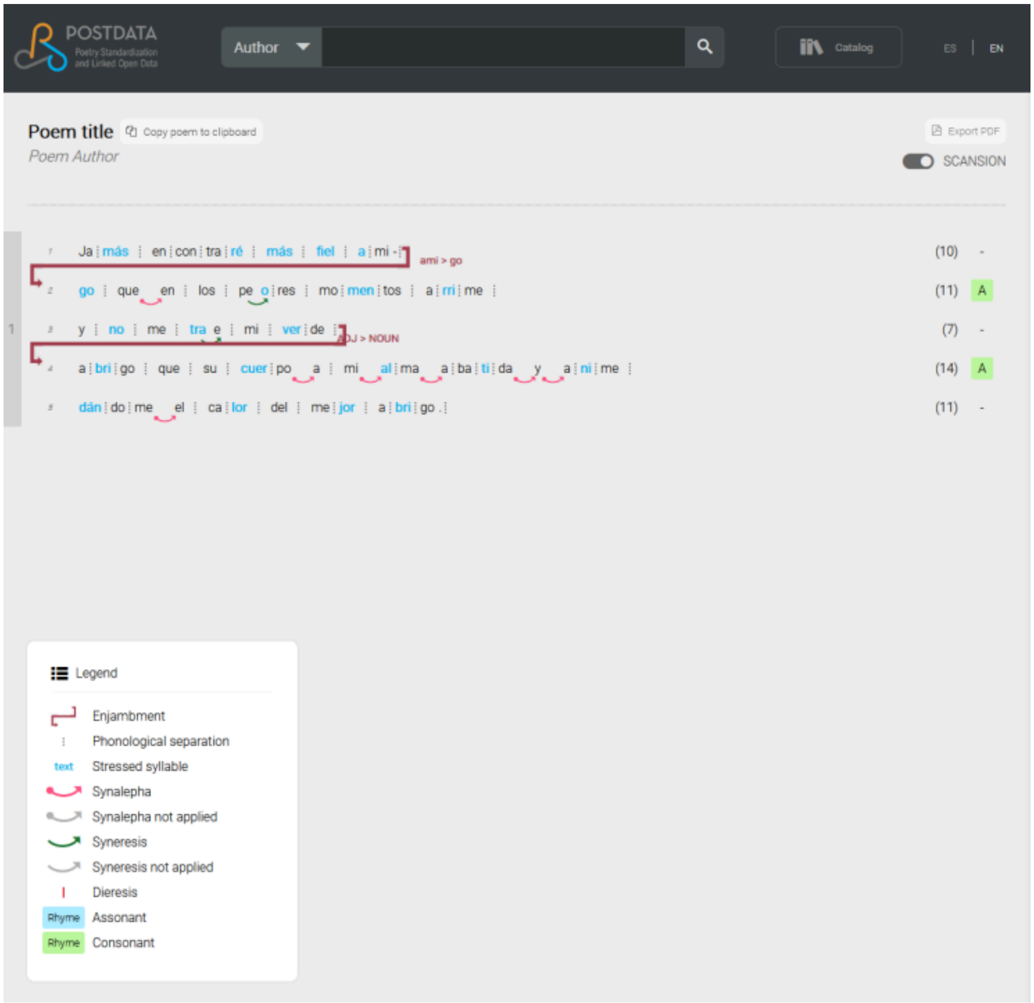
In this view, the poem's text is broken down into syllables and annotated with various symbols to represent different phonological and metric properties. The poem text is displayed with vertical bars (|) indicating phonological separation between syllables, and stressed syllables are denoted with an accent mark above the vowel. The visualization includes specific symbols to indicate poetic devices and metric characteristics: A red bracket indicates enjambment, showing the run-on line structure between verses. Small arcs above syllables represent synalepha, where the vowel at the end of one word merges with the vowel at the beginning of the next. A squiggly line represents sinaeresis, the merging of two syllables into one. A straight line denotes dieresis, the separation of a diphthong into two syllables. The color coding at the bottom of the words signifies rhyme schemes, with different colors for assonant and consonant rhymes. As with the simple view, each line of the poem is followed by a number in parentheses indicating the syllable count, and a letter which marks rhyming verses.

On the left side of the screen, there is a legend explaining the symbols and annotations used in the scansion:

- Red brackets for enjambment.
- Double bars for phonological separation.
- Green check marks for synalepha applied, and red crosses for synalepha not applied.
- Green squiggly line for syneresis and a red squiggly line for syneresis not applied.
- A straight line for dieresis.
- Colors under the text for assonant and consonant rhymes.

This visualization method provides a detailed analysis of the poem, which can be beneficial for literary studies, allowing users to understand the intricate patterns and structures of Spanish poetry. With this level of detail, researchers and students can study the metrical composition and phonetic features of the poem, such as how syllables combine or separate across lines and how rhymes are structured.

Finally, the interface offers an option to "Export PDF" button to download the analysis.



1 Ja ¦ más ¦ en ¦ con ¦ tra ¦ ré ¦ más ¦ fiel ¦ a ¦ mi - ¦ (10) -
2 go ¦ que en ¦ los ¦ pe o ¦ res ¦ mo ¦ men ¦ tos ¦ a ¦ rri ¦ me ¦ (11) A
3 y ¦ no ¦ me ¦ tra e ¦ mi ¦ ver ¦ de ¦ (7) -
4 a ¦ bri ¦ go ¦ que ¦ su ¦ cuer ¦ po a ¦ mi al ¦ ma a ¦ ba ¦ ti ¦ da y a ¦ ni ¦ me ¦ (14) A
5 dán ¦ do ¦ me el ¦ ca ¦ lor ¦ del ¦ me ¦ jor ¦ a ¦ bri ¦ go . ¦ (11) -

Fig. 15: Advanced view for scansion and enjambment.

Outcomes

The output of Rantanplan is a complex structure that will be broken down for clarity.

First, Rantanplan will show a list of stanzas. Each stanza is then shown as two separate lists. A list of tokens, and a list of “phonological groups” i.e., the phonological units that form a verse after synalephas and sinaereris are considered.

Tokens

If the token is a word, it shows a list of the syllables it is made of, with the following information:

- *syllable*: The text of the syllable.
- *is_stressed*: Whether the syllable is stressed or not.
- *is_word_end*: Whether the syllable is the end of a word or not.
- *has_synalepha* or *has_sinaeresis*: Whether the syllable can be conjoined with the next one.
- *stress_position*: Index, starting from 0, for the stressed syllable of the word. If the index is negative, the syllable position is counted from the end of the word:
    - 0: First syllable
    - -1: Last syllable
    - -2: Penultimate syllable
    - *etc*

If the token is not a word, it is shown as a symbol.

List of tokens example:

```
{'tokens': [{'word': [{'syllable': 'co', 'is_stressed': True},
  {'syllable': 'mo',
   'is_stressed': False,
   'has_synalepha': True,
   'is_word_end': True}],
 'stress_position': 0},
 {'word': [{'syllable': 'au', 'is_stressed': False}
...
 {'symbol': ','}],
...
```

Phonological groups

The next element of the output is a list of phonological groups. We use this term to refer to the phonological unit that makes up a poem when it is read, after synalephas and sinaereris are considered.

Phonological groups are quite like the token list but have no word boundaries because this is lost when applying synalephas. Each syllable within *phonological_groups* can carry the following information:

- *syllable*: The text of the syllable.
- *is_stressed*: Whether the syllable is stressed or not.
- *is_word_end*: Whether the syllable is the end of a word or not.
- *synalepha_index* or *sinaeresis_index*: The index of the character where the syllable is conjoined with the next one:
    - ○ 0: No synalepha or sinaeresis has been realized.
    - ○ Any other number: List of indexes on the syllable, starting from 0, where the original syllable or syllables have been conjoined with the next one:

§ Example: The syllable *moau* was originally split at position 1:

§ {'syllable': 'moau', 'is_stressed': **False**, 'synalepha_index': [1]}

§ Indexes of the syllable:

m o a u

0 1 2 3

We split at position 1: o, so then, we know that the original syllables are *mo* and *au.*

Phonological groups example

```
{'phonological_groups': [{'syllable': 'Me',

 'is_stressed': False,

 'is_word_end': True},

{'syllable': 'gus', 'is_stressed': True},

{'syllable': 'tas', 'is_stressed': False, 'is_word_end': True},

{'syllable': 'cuan', 'is_stressed': False},

{'syllable': 'do', 'is_stressed': False, 'is_word_end': True},

{'syllable': 'ca', 'is_stressed': True},

{'syllable': 'llas', 'is_stressed': False, 'is_word_end': True},

{'syllable': 'por', 'is_stressed': False},

{'syllable': 'quees', 'is_stressed': False, 'synalepha_index': [2]},

{'syllable': 'tás', 'is_stressed': True, 'is_word_end': True},

{'syllable': 'co', 'is_stressed': False},

{'syllable': 'moau', 'is_stressed': False, 'synalepha_index': [1]},
{'syllable': 'sen', 'is_stressed': True},

{'syllable': 'te', 'is_stressed': False, 'is_word_end': True}],
```

Metrical information

Finally, at the verse level we find information about the verse itself on the *rhythm* key:

- *rhythm*: Pattern of the unstressed (-) and stressed (+) syllable. This output can be changed with the parameter *rhythm_format*. You can find more information about how this parameter works on the documentation.
- *length*: Proposed length for the verse.
- *length_range*: Minimum and maximum verse length possible. This is calculated considering all possible sinaeresis and synalephas.

Metrical information example

```
'rhythm': {'stress': '---+----+----+-',

 'length': 14,

 'length_range': {'min_length': 13, 'max_length': 16}},

 ...
```

Stanza detection

Rantanplan is also able to detect the stanza type from a list of popular Spanish stanzas.

When this option is enabled with the *rhyme_analysis*, additional information about the stanza is shown on the output.

If we take this “cuarteto” for example:

Yo persigo una forma que no encuentra mi estilo,

botón de pensamiento que busca ser la rosa;
se anuncia con un beso que en mis labios se posa
al abrazo imposible de la Venus de Milo

If we call *get_scansion* with the *rhyme_analysis* parameter set to *True*, the following information is added to the analysis of each line:

- *structure*: The name of the stanza that has been detected.
- *rhyme*: A letter code to match rhyming verses. In this example, verse 1 rhymes with verse 4, and verse 2 rhymes with verse 3, and a letter is assigned to verses that rhyme together as shown below:

○	Yo persigo una forma que no encuentra mi estilo,	a
○	botón de pensamiento que busca ser la rosa;	b
○	se anuncia con un beso que en mis labios se posa	b
○	al abrazo imposible de la Venus de Milo	a

- *ending*: What part of the last word is rhyming.
- *ending_stress*: Negative index (-1 for last, -2 for penultimate, etc.) for the vowel that carries the stress of the rhyming part.
- *rhyme_type*: Whether the rhyme is consonant or assonant:
- Consonant: All characters from the last stressed vowel to the end the word coincides on verses that rhyme. For example:

○	est**ILO**
○	m**ILO**

- Assonant: Same as consonant rhyme but only if all vowels match:

○	am**A**d**O**
○	cach**A**rr**O**

- *rhyme_relaxation*: Whether ot not rules for rhyme relaxation are applied. For example, removing weak vowels on diphthongs or making letters match when they are pronounced the same, for example c and z.

Stanza detection example

```
'structure': 'cuarteto',

'rhyme': 'a',

'ending': 'ilo',

'ending_stress': -3,

'rhyme_type': 'consonant',

'rhyme_relaxation': True},

 ...
```

Full output example

A complete example of Rantanplan output is shown here:

```
 [{'tokens': [{'word': [{'syllable': 'Me',

   'is_stressed': False,

   'is_word_end': True}],

 'stress_position': 0},
{'word': [{'syllable': 'gus', 'is_stressed': True},
  {'syllable': 'tas', 'is_stressed': False, 'is_word_end': True}],
 'stress_position': -2},
{'word': [{'syllable': 'cuan', 'is_stressed': False},
  {'syllable': 'do', 'is_stressed': False, 'is_word_end': True}],
 'stress_position': 0},
{'word': [{'syllable': 'ca', 'is_stressed': True},
  {'syllable': 'llas', 'is_stressed': False, 'is_word_end': True}],
 'stress_position': -2},
{'word': [{'syllable': 'por', 'is_stressed': False},
  {'syllable': 'que',
   'is_stressed': False,
   'has_synalepha': True,
   'is_word_end': True}],
 'stress_position': 0},
{'word': [{'syllable': 'es', 'is_stressed': False},
   {'syllable': 'tás', 'is_stressed': True, 'is_word_end': True}],
  'stress_position': -1},
 {'word': [{'syllable': 'co', 'is_stressed': False},
   {'syllable': 'mo',
    'is_stressed': False,
    'has_synalepha': True,
    'is_word_end': True}],
  'stress_position': 0},
 {'word': [{'syllable': 'au', 'is_stressed': False},
   {'syllable': 'sen', 'is_stressed': True},
   {'syllable': 'te', 'is_stressed': False, 'is_word_end': True}],
  'stress_position': -2},
 {'symbol': ','}],
'phonological_groups': [{'syllable': 'Me',
  'is_stressed': False,
  'is_word_end': True},
 {'syllable': 'gus', 'is_stressed': True},
 {'syllable': 'tas', 'is_stressed': False, 'is_word_end': True},
 {'syllable': 'cuan', 'is_stressed': False},
 {'syllable': 'do', 'is_stressed': False, 'is_word_end': True},
 {'syllable': 'ca', 'is_stressed': True},
 {'syllable': 'llas', 'is_stressed': False, 'is_word_end': True},
 {'syllable': 'por', 'is_stressed': False},
 {'syllable': 'quees', 'is_stressed': False, 'synalepha_index': [2]},
 {'syllable': 'tás', 'is_stressed': True, 'is_word_end': True},
 {'syllable': 'co', 'is_stressed': False},
 {'syllable': 'moau', 'is_stressed': False, 'synalepha_index': [1]},
 {'syllable': 'sen', 'is_stressed': True},
 {'syllable': 'te', 'is_stressed': False}],
'rhythm': {'stress': '-+---+---+--+-', 'type': 'pattern', 'length': 14}},
 ...
```

Further information

For a guide on how to install and use rantanplan, please refer to our GitHub (https://github.com/linhd-postdata/rantanplan), or to our official documentation for an in-depth explanations of the functions (https://rantanplan.readthedocs.io/).

Publications

- De la Rosa, J., Pérez, Á., Hernández, L., Ros, S., & Gonzalez Blanco, E. (2020). PoetryLab as Infrastructure for the Analysis of Spanish Poetry. Proceedings of CLARIN Annual Conference 2020, 82-87. https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf
- Hernández Lorenzo, L., De Sisto, M., Pérez, Á., De la Rosa, J., Ros, S., & González-Blanco, E. (2021). Automatic quantitative metrical analysis of Spanish Poetry with Rantanplan: A first approach. En Tackling the Toolkit. Plotting Poetry through Computational Literary Studies.
- Pérez Pozo, Á., Rosa, J., Ros, S., González-Blanco, E., Hernández, L., & Sisto, M. (2021). A bridge too far for artificial intelligence?: Automatic classification of stanzas in Spanish poetry. Journal of the Association for Information Science and Technology, asi.24532. https://doi.org/10.1002/asi.24532
- Rosa, J. de la, Pérez, Á., Hernández, L., Ros, S., & González-Blanco, E. (2020). Rantanplan, Fast and Accurate Syllabification and Scansion of Spanish Poetry. Procesamiento del Lenguaje Natural, 65(0), Article 0.

References

- Delor, M., Martí, A., & Recasens, M. (2008). AnCora: Multilevel annotated corpora for Catalan and Spanish. *Proceedings of 6th International Conference on Language Resources and Evaluation*, 96-101.
- Domínguez Caparrós, J. (2014). *Métrica española* (Primera edición). Universidad Nacional de Educación a Distancia.
- Honnibal, M., & Montani, I. (2017, enero 1). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. Sentometrics Research. https://sentometrics-research.com/publication/72/.
- Mestre, A. R. (1998). *La transcripción fonética automática del diccionario electrónico de formas simples flexivas del español: Un estudio fonológico en el léxico* [Http://purl.org/dc/dcmitype/Text, Universitat Autònoma de Barcelona]. https://dialnet.unirioja.es/servlet/tesis?codigo=182922.
- Navarro Tomás, T. (1918). *Manual de pronunciación española*. Junta para ampliación de estudios e investigaciones científicas. Centro de estudios históricos.

- Navarro-Colorado, B. (2017). A metrical scansion system for fixed-metre Spanish poetry. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx009.

Mapping Arthur Schnitzler in space and time

URL

The showcase describes a set of related applications gathered under a common umbrella: https://schnitzler.acdh.oeaw.ac.at/.

Repository

Datasets and applications described in the showcase are maintained in multiple repositories under: https://github.com/arthur-schnitzler.

Status / publication date

The materials and applications have been continuously developed since 2018. In the context of the CLS INFRA showcase these materials have been further consolidated and described.

Creators / developers / authors

Authors responsible for the showcase description:

- Matej Ďurčo (Austrian Centre for Digital Humanities and Cultural Heritage, Vienna)
- Vera Maria Charvát (Austrian Centre for Digital Humanities and Cultural Heritage, Vienna)

Creators of the underlying datasets and applications presented in the show case:

- Martin Anton Müller (Austrian Centre for Digital Humanities and Cultural Heritage, Vienna)
- Laura Untner (Austrian Centre for Digital Humanities and Cultural Heritage, Vienna)
- Peter Andorfer (Austrian Centre for Digital Humanities and Cultural Heritage, Vienna)
- Gerd-Hermann Susen
- and others (detailed credits can be found in the about pages of the respective projects and in the contribution lists of the underlying repositories)

Target audience

The target audience of this showcase can be divided into two main groups:

- scholars as well as broader public interested in the life and work of the author Arthur Schnitzler;
- DH practitioners interested in the methodological approaches for compiling, analysing and presenting non-literary material of a person in novel ways enabling richer means for exploration and thus deeper/better understanding of given person's life.

Executive summary

This showcase presents the extensive work of Martin Anton Müller and Laura Untner, aimed at elucidating the life of the writer Arthur Schnitzler. By analyzing a comprehensive body of historical textual material and making it available through a set of interconnected applications, different aspects of Schnitzler's life can be explored, mainly along the natural primary dimensions of time, space, and social relations.

This undertaking offers great value to literary scholars, historians, and anybody interested in the life and work of Arthur Schnitzler or Viennese social life around 1900. Beyond the particular scope, it is also a prime example of digital humanities scholarship combining scholarly insight, innovative technology, as well as careful data preparation and curation to offer novel and comprehensive ways of exploring historical material.

The value of this contribution for the DH and CLS community is further enhanced by the use of well-established technologies, extensive, elaborate documentation, as well as strict adherence to the principles of Open Science, by making all the underlying data and tooling available for reuse under permissive licenses.

This showcase gives a concise overview of the interconnected collection of resources by way of contextualising. It then describes the main underlying datasets, their technical parameters and coverage, as well as the methods, formats and tools applied to create these datasets and to make them available through the various web applications. Finally, the rich bouquet of visualisation functionalities is described, which allow to explore the different aspects of the datasets.

Research questions and objectives

The research question / objective of the described showcase is to explore the ways how non-literary historical textual material, such as correspondences or diaries, can be used to offer a comprehensive representation of a person's life, exemplified through the author Arthur Schnitzler.

Data

The data underlying this showcase consists of several datasets, continuously edited and developed by the team since 2018. These datasets are made available for rich exploration through dedicated applications, gathered on the signpost page: https://schnitzler.acdh.oeaw.ac.at/.

At the same time, the underlying raw data, mostly encoded in TEI/XML, is available in dedicated GitHub repositories, gathered under https://github.com/arthur-schnitzler/.

In the following, selected datasets from this extensive collection will be described.

Diary

Arthur Schnitzler maintained a diary from 1879 to 1931, comprising over 16,400 (almost daily) entries. The diary entries provide insights into the life of the Viennese writer, his encounters, and conversations. The diary was initially published in ten volumes from 1981 to 2000.[1] In 2018, an initial version of a digital edition was released, with gradual improvements over upcoming years. This digital edition is based on the printed edition's texts but has been significantly improved, especially regarding the named entities such as persons, places and works. It includes a comprehensive index of individuals mentioned in the text, along with additional information (date and place of birth and death, as well as (and most importantly) links to other semantic resources. Additionally, the digital edition now features facsimiles of the individual diary entries.

The digital edition of Arthur Schnitzler's diaries is available via https://schnitzler-tagebuch.acdh.oeaw.ac.at/.

Next to this interactive application, the data is made available in TEI format on ARCHE, the platform for long-term archiving at the ACDH-CH.[2] The entire edition can also be obtained via GitHub (data, application).

The individual diary entries in the online edition can be accessed and viewed by using the calender format or via a searchable table of contents list. Additionally, the edition features indices of all persons, places and works Schnitzler mentioned in his diary entries. Those indices are all available in the GitHub repository including concordance lists of the mentioned entities sorted by day.

[1] Schnitzler, A. (1981). Arthur Schnitzler Tagebuch 1879-1931. Gesamtwerk (W. Welzig, M. Neyes, R. Miklin, P. M. Braunwarth, K. Fliedl, W. Ruprechter, R. Urbach, & S. Perlik, Eds.). Verlag der Österreichischen Akademie der Wissenschaften. See also: https://www.austriaca.at/arthur_schnitzler_tagebuch.

[2] see: https://hdl.handle.net/21.11115/0000-000E-0FF1-2.

Places

The project "Arthur Schnitzler Orte" aimed to create a comprehensive list of Schnitzler's whereabouts, i.e. places he lived in or visited during his lifetime. The data is based on diary entries, a list of his travels, and other sources.

The data source file can be downloaded from GitHub.

The project's main page (https://schnitzler-orte.acdh.oeaw.ac.at) includes several visualisations, which are described in the Visualisation section of this showcase.

Correspondences

The project "Arthur Schnitzler Briefe" features current TEI encoded editions of Arthur Schnitzler's correspondence from 1888 to 1931 with his colleagues[3]. The main website (https://schnitzler-briefe.acdh.oeaw.ac.at/) features 45 exchanges of letters and records 3.618 letters in total - many of them are being published for the first time. All letters have been scrutinized and corrected according to the originals in archives in Europe and America.

The project's website provides lists of the individual letters, correspondences, postal routes and archives, which store the original letters. Additionally, indices for persons, works, places and institutions can be extracted from the TEI data.

Authority File - PMB

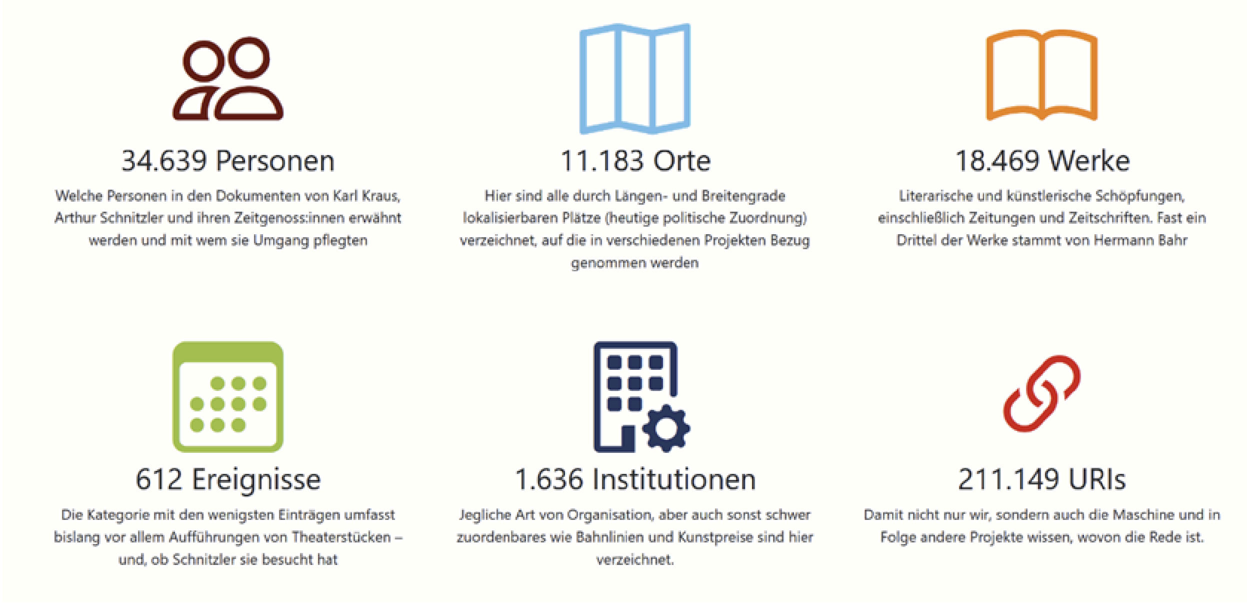
The processing of the various resources like the correspondences or the diary, prompted the need for an overarching local authority file – a central registry of the named entities mentioned in the various texts. Both to ensure consistency, to avoid redundancy and especially to establish reliable semantic links between the datasets. To this end a dedicated database application, PMB (Personen Moderne Basis) has been introduced, based on the framework APIS, used to manage prosopographical and biographical data.[4] The application, available under https://pmb.acdh.oeaw.ac.at/, also exposes a REST API allowing programmatic access from other applications.[5]

It serves as a central registry of people, works, locations and institutions specifically for the period around 1900 with focus on Vienna, which can be used by various edition projects, reducing the work needed to identify and curate entries for named entities.

[3] A full list of all correspondences currently included in the edition as well as the number of edited correspondence items can be found here: https://schnitzler-briefe.acdh.oeaw.ac.at/tocs.html.
[4] see: https://www.oeaw.ac.at/acdh/tools/apis-app.
[5] see: https://pmb.acdh.oeaw.ac.at/apis/api/.

Moreover, the entries in PMB are linked to corresponding entries in global reference resources and authority files like Wikidata, GND or Geonames, thus serving as a Linked Open Data hub.

Recently PMB has gained special recognition, through the introduction of a dedicated Wikidata-property PMB-id (similar to the GND-id and other dedicated authority file specific identifier properties featured in Wikidata).



34.639 Personen

Welche Personen in den Dokumenten von Karl Kraus, Arthur Schnitzler und ihren Zeitgenoss:innen erwähnt werden und mit wem sie Umgang pflegten

11.183 Orte

Hier sind alle durch Längen- und Breitengrade lokalisierbaren Plätze (heutige politische Zuordnung) verzeichnet, auf die in verschiedenen Projekten Bezug genommen werden

18.469 Werke

Literarische und künstlerische Schöpfungen, einschließlich Zeitungen und Zeitschriften. Fast ein Drittel der Werke stammt von Hermann Bahr

612 Ereignisse

Die Kategorie mit den wenigsten Einträgen umfasst bislang vor allem Aufführungen von Theaterstücken – und, ob Schnitzler sie besucht hat

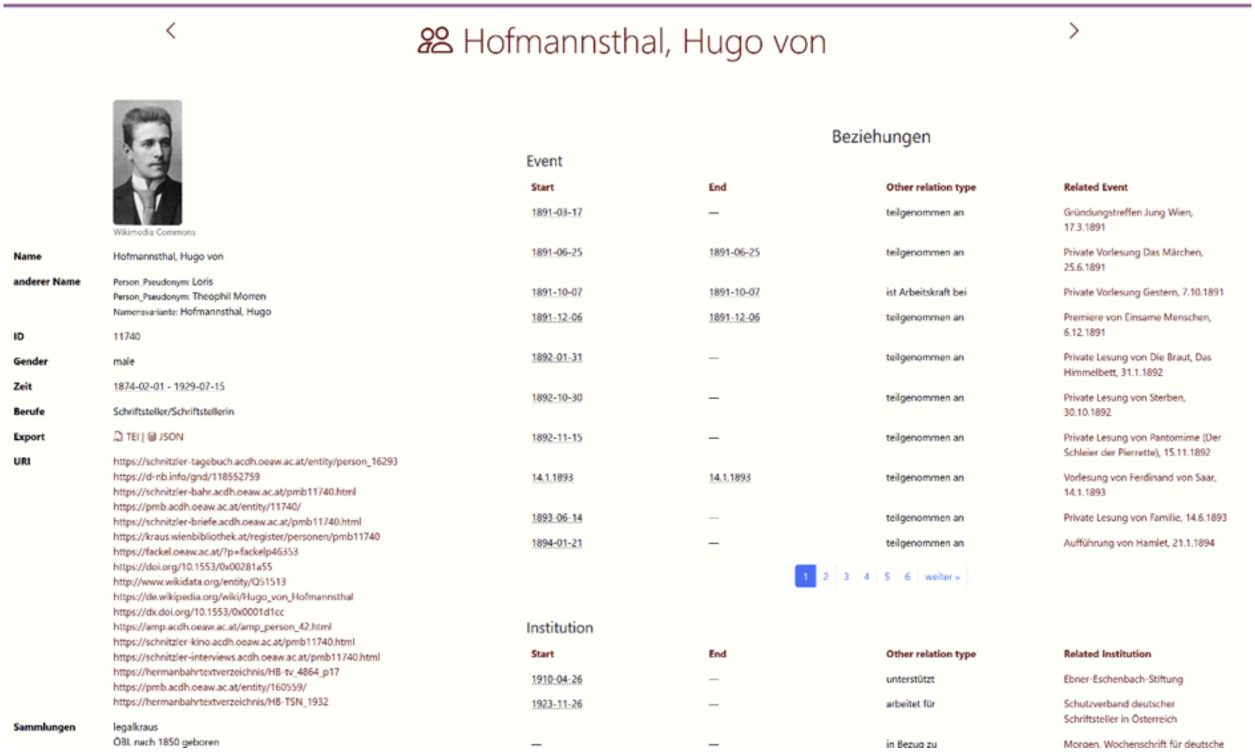
1.636 Institutionen

Jegliche Art von Organisation, aber auch sonst schwer zuordenbares wie Bahnlinien und Kunstpreise sind hier verzeichnet.

211.149 URIs

Damit nicht nur wir, sondern auch die Maschine und in Folge andere Projekte wissen, wovon die Rede ist.

Fig. 1: Screenshot of the frontpage of PMB.



Hofmannsthal, Hugo von

Name	Hofmannsthal, Hugo von
anderer Name	Person_Pseudonym: Loris Person_Pseudonym: Theophil Morren Namensvariante: Hofmannsthal, Hugo
ID	11740
Gender	male
Zeit	1874-02-01 - 1929-07-15
Berufe	Schriftsteller/Schriftstellerin
Export	TEI | JSON
URI	https://schnitzler-tagebuch.acdh.oeaw.ac.at/entity/person_16293 https://d-nb.info/gnd/118552759 https://schnitzler-bahr.acdh.oeaw.ac.at/pmb11740.html https://pmb.acdh.oeaw.ac.at/entity/11740/ https://schnitzler-briefe.acdh.oeaw.ac.at/pmb11740.html https://kraus.wienbibliothek.at/register/personen/pmb11740 https://fackel.oeaw.ac.at/?p=fackelp46353 https://doi.org/10.1553/0x00281a55 http://www.wikidata.org/entity/Q51513 https://de.wikipedia.org/wiki/Hugo_von_Hofmannsthal https://dx.doi.org/10.1553/0x0001d1cc https://amp.acdh.oeaw.ac.at/amp_person_42.html https://schnitzler-kino.acdh.oeaw.ac.at/pmb11740.html https://schnitzler-interviews.acdh.oeaw.ac.at/pmb11740.html https://hermanbahrtextverzeichnis/HB-tv_4864_p17 https://pmb.acdh.oeaw.ac.at/entity/160559/ https://hermanbahrtextverzeichnis/HB-TSN_1932
Sammlungen	legalkraus ÖBL nach 1850 geboren

Beziehungen

Event

Start	End	Other relation type	Related Event
1891-03-17	—	teilgenommen an	Gründungstreffen Jung Wien, 17.3.1891
1891-06-25	1891-06-25	teilgenommen an	Private Vorlesung Das Märchen, 25.6.1891
1891-10-07	1891-10-07	ist Arbeitskraft bei	Private Vorlesung Gestern, 7.10.1891
1891-12-06	1891-12-06	teilgenommen an	Premiere von Einsame Menschen, 6.12.1891
1892-01-31	—	teilgenommen an	Private Lesung von Die Braut, Das Himmelbett, 31.1.1892
1892-10-30	—	teilgenommen an	Private Lesung von Sterben, 30.10.1892
1892-11-15	—	teilgenommen an	Private Lesung von Pantomime (Der Schleier der Pierrette), 15.11.1892
14.1.1893	14.1.1893	teilgenommen an	Vorlesung von Ferdinand von Saar, 14.1.1893
1893-06-14	—	teilgenommen an	Private Lesung von Familie, 14.6.1893
1894-01-21	—	teilgenommen an	Aufführung von Hamlet, 21.1.1894

Institution

Start	End	Other relation type	Related Institution
1910-04-26	—	unterstützt	Ebner-Eschenbach-Stiftung
1923-11-26	—	arbeitet für	Schutzverband deutscher Schriftsteller in Österreich
—	—	in Bezug zu	Morgen. Wochenschrift für deutsche

Fig. 2: The example above features the PMB entry for Hugo von Hofmannsthal.[6]

Chronicle

The "Chronik" (Chronicle) is the most recent addition to the Schnitzler collection. Its purpose is to gather materials from the different sources, primarily the diary and the correspondences, but also others, and collocate them along the time dimension, similar to what "Schnitzler Orte" delivered for the spatial dimension.

For each day of Arthur Schnitzler's (adult) life, the application, available at https://schnitzler-chronik.acdh.oeaw.ac.at, offers the significant events as witnessed by these various source materials, both as structured data encoded in JSON and TEI/XML as well as in the form of a HTML page with corresponding links to the source material.

[6] see: https://pmb.acdh.oeaw.ac.at/apis/entities/entity/person/11740/detail.

Methods

Data preparation / preprocessing

Though the collection consists of distinct datasets based on different historical material, the workflow for creating the datasets has been mostly the same (the following description uses the correspondences dataset as example):

1. Obtaining digitized versions/scans of the resources from the holding library or archive.

2. Using the text recognition service Transkribus to produce digital textual representation either using existing OCR models, or training own models.[7]
   Note: Some of the newly trained OCR models were also made available for reuse.[8]

3. Automated process (implemented as GitHub action) to retrieve the data from Transkribus via API and convert it into TEI files and store these in a dedicated github repository, e.g.: https://github.com/arthur-schnitzler/schnitzler-briefe-data

Publication of data via a web application

Most of the web applications of the Schnitzler collection are based on the dse-cookie-cutter utility software aimed at easing the process of publishing XML/TEIs encoded files as static sites. *dse-cookie-cutter* features a set of XSLT scripts to generate a generic HTML representation of the texts encoded in XML/TEI, as well as PDF and Epub format (including the critical apparatus of the edition). The scripts as well as the look and feel and the functionality can be further customized, but the *dse-cookie-cutter* offers a very quick path to a baseline prototype.

For the correspondences dataset, the code of the application is located at: https://github.com/arthur-schnitzler/schnitzler-briefe-static, and the application itself can be accessed at https://schnitzler-briefe.acdh.oeaw.ac.at/.

Identifying named entities and establishing semantic links (LOD)

Due to the peculiar form of most of the examined resources, especially the diary, where individual persons are often referred to by just their forename or even just one letter, automated named entity recognition was only of limited applicability. In the diary, the annotations and linking of named entities were based on a digitized index of individuals from the printed edition, sorted by their mentions on a specific day. The persons were then matched to the PMB entries.

[7] see: https://readcoop.eu/transkribus/.
[8] see: HTR goldmann: https://readcoop.eu/model/paul-goldmann_german-kurrent_1889-96_v1/ and HTR salten: https://readcoop.eu/model/german-letters-felix-salten-1890-1931/.

In cases where the datasets to be integrated already contained references to standard authority files, such as GND, geonames.org, automated matching was possible thanks to the set of external identifiers collected in PMB.

The enrichment of the information about the named entities is done centrally in PMB, ideally by fetching relevant information automatically from external reference resources where available. (As of recently, this includes also the possibility to fetch images for persons from Wikimedia Commons.)

Example: Modelling snippet for a person in TEI from a listPerson index in the Schnitzler correspondences dataset.[9]

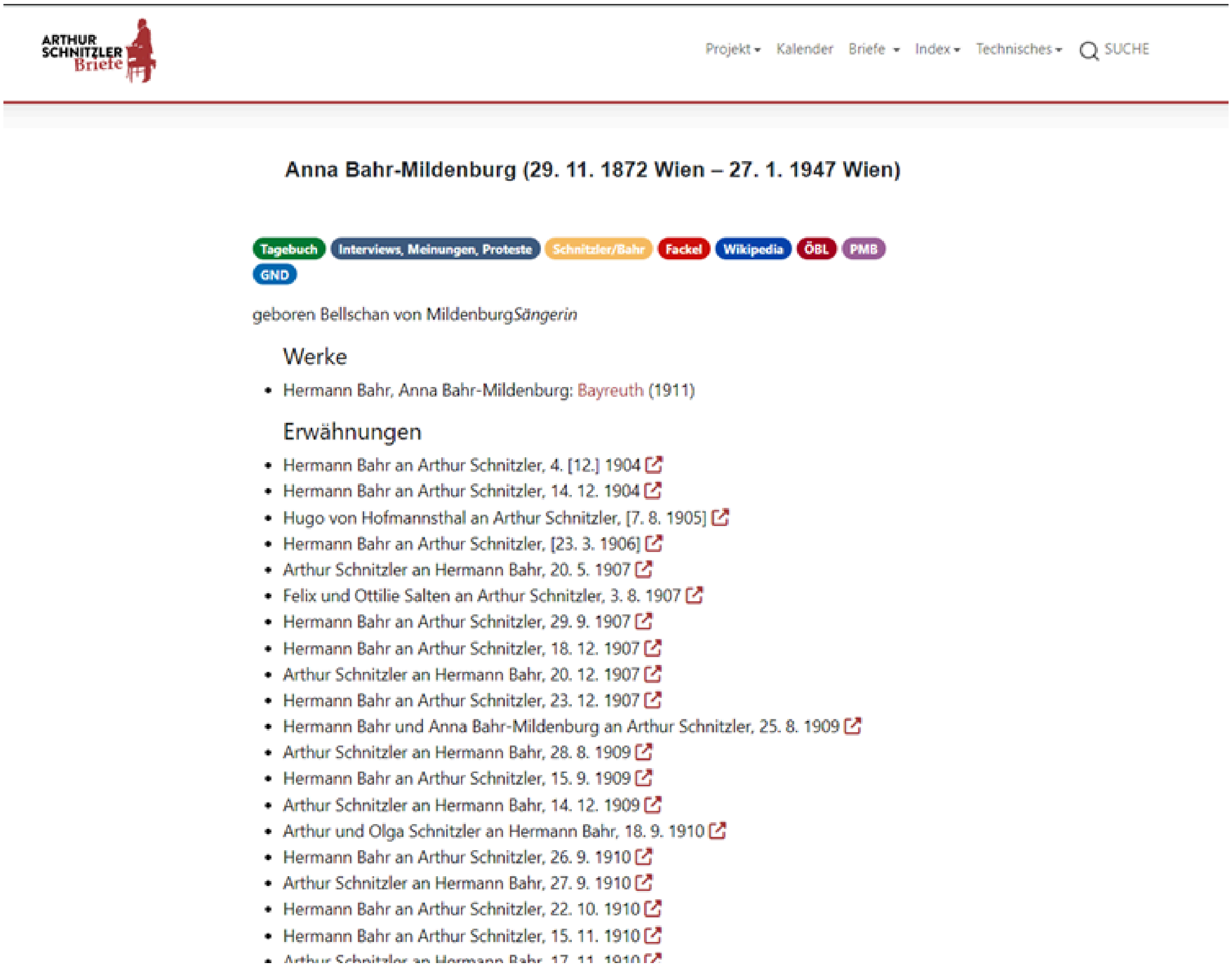
```
<person xml:id="person__10818">
  <persName>
<forename>Anna</forename>
    <surname>Bahr-Mildenburg</surname>
  </persName>
  <persName type="person_geburtsname-vorname">Bellschan</persName>
  <persName type="person_geburtsname_nachname">von Mildenburg</persName>
  <persName type="legacy-name-merge">Bahr-Mildenburg, Anna</persName>
  <birth>
<date when-iso="1872-11-29">29.11.1872</date>
<settlement key="50">
        <placeName type="pref">Wien</placeName>
        <location>
        <geo>48,2066 16,37341</geo>
        </location>
</settlement>
  </birth>
  <death>
<date when-iso="1947-01-27">27.01.1947</date>
<settlement key="50">
        <placeName type="pref">Wien</placeName>
        <location>
        <geo>48,2066 16,37341</geo>
        </location>
</settlement>
  </death>
  <sex value="female"/>
  <occupation key="107">Sänger/Sängerin</occupation>
</person>
```

Example (continuing from the previous example): References to the corresponding PMB entry as well as links to global semantic reference resources (Wikidata, GND, etc.).

```
<person xml:id="person__10818">
  <persName>
    <forename>Anna</forename>
    <surname>Bahr-Mildenburg</surname>
  </persName>
```

[9] see: https://github.com/arthur-schnitzler/schnitzler-briefe-data/blob/main/data/indices/listperson.xml.

```
    <idno type="URL" subtype="schnitzler-tagebuch">https://schnitzler-
      tagebuch.acdh.oeaw.ac.at/entity/person_13376</idno>
    <idno type="URL" subtype="gnd">https://d-nb.info/gnd/118646370</idno>
    <idno type="URL" subtype="schnitzler-bahr">https://schnitzler-
      bahr.acdh.oeaw.ac.at/pmb10818.html</idno>
    <idno type="URL" subtype="pmb">https://pmb.acdh.oeaw.ac.at/entity/10818/</idno>
    <idno type="URL" subtype="schnitzler-briefe">https://schnitzler-
      briefe.acdh.oeaw.ac.at/pmb10818.html</idno>
    <idno type="URL" subtype="oebl">https://doi.org/10.1553/0x00280ef9</idno>
    <idno type="URL" subtype="fackel">https://fackel.oeaw.ac.at/?p=fackelp50201</idno>
    <idno type="URL" subtype="wikidata">http://www.wikidata.org/entity/Q79028</idno>
    <idno type="URL"
      subtype="wikipedia">https://de.wikipedia.org/wiki/Anna_von_Mildenburg</idno>
    <idno type="URL" subtype="oeml">https://dx.doi.org/10.1553/0x0001d9a6</idno>
    <idno type="URL" subtype="schnitzler-interviews">https://schnitzler-
      interviews.acdh.oeaw.ac.at/pmb10818.html</idno>
    <idno type="URL" subtype="pmb">https://pmb.acdh.oeaw.ac.at/entity/159210/</idno>
    <idno type="URL" subtype="hermanbahrtextverzeichnis">https://hermanbahrtextverzeichnis/HB-
      TSN_202</idno>
  </person>
```



Anna Bahr-Mildenburg (29. 11. 1872 Wien – 27. 1. 1947 Wien)

geboren Bellschan von Mildenburg*Sängerin*

Werke

- Hermann Bahr, Anna Bahr-Mildenburg: Bayreuth (1911)

Erwähnungen

- Hermann Bahr an Arthur Schnitzler, 4. [12.] 1904
- Hermann Bahr an Arthur Schnitzler, 14. 12. 1904
- Hugo von Hofmannsthal an Arthur Schnitzler, [7. 8. 1905]
- Hermann Bahr an Arthur Schnitzler, [23. 3. 1906]
- Arthur Schnitzler an Hermann Bahr, 20. 5. 1907
- Felix und Ottilie Salten an Arthur Schnitzler, 3. 8. 1907
- Hermann Bahr an Arthur Schnitzler, 29. 9. 1907
- Hermann Bahr an Arthur Schnitzler, 18. 12. 1907
- Arthur Schnitzler an Hermann Bahr, 20. 12. 1907
- Hermann Bahr an Arthur Schnitzler, 23. 12. 1907
- Hermann Bahr und Anna Bahr-Mildenburg an Arthur Schnitzler, 25. 8. 1909
- Arthur Schnitzler an Hermann Bahr, 28. 8. 1909
- Hermann Bahr an Arthur Schnitzler, 15. 9. 1909
- Arthur Schnitzler an Hermann Bahr, 14. 12. 1909
- Arthur und Olga Schnitzler an Hermann Bahr, 18. 9. 1910
- Hermann Bahr an Arthur Schnitzler, 26. 9. 1910
- Arthur Schnitzler an Hermann Bahr, 27. 9. 1910
- Hermann Bahr an Arthur Schnitzler, 22. 10. 1910
- Hermann Bahr an Arthur Schnitzler, 15. 11. 1910
- Arthur Schnitzler an Hermann Bahr, 17. 11. 1910

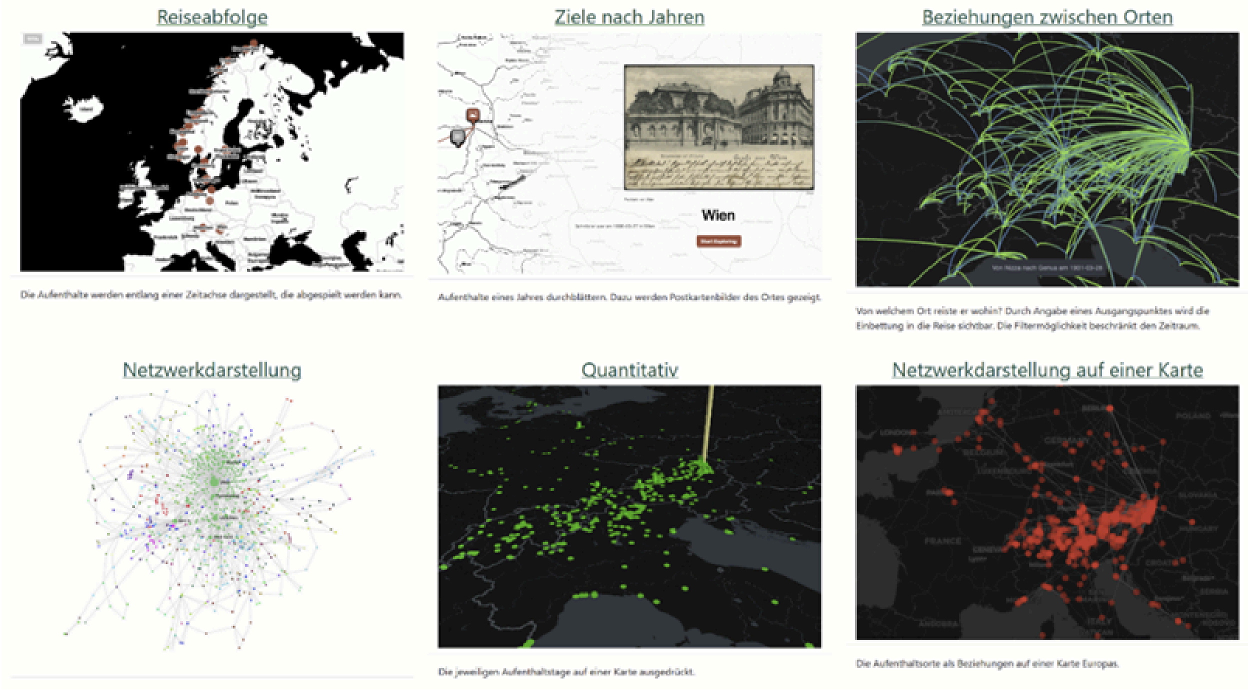
Fig. 3: Detailed view of the above-mentioned references in the frontend (the individual linked references are displayed in form of multi-colored buttons).

Visualization

The various applications of the Schnitzler collection offer different views on the datasets, chiefly along the dimensions of space and time.

Space

The project "Schnitzler Orte" (Schnitzler Places) focuses exclusively on the spatial aspect of Schnitzler’s life. The project’s main page (https://schnitzler-orte.acdh.oeaw.ac.at) includes several visualizations that allow measuring the geographical space in which Schnitzler has moved since he came of age. Below is a screenshot of the main page offering access to the different visualizations:



Die Aufenthalte werden entlang einer Zeitachse dargestellt, die abgespielt werden kann.

Aufenthalte eines Jahres durchblättern. Dazu werden Postkartenbilder des Ortes gezeigt.

Von welchem Ort reiste er wohin? Durch Angabe eines Ausgangspunktes wird die Einbettung in die Reise sichtbar. Die Filtermöglichkeit beschränkt den Zeitraum.

Die jeweiligen Aufenthaltstage auf einer Karte ausgedrückt.

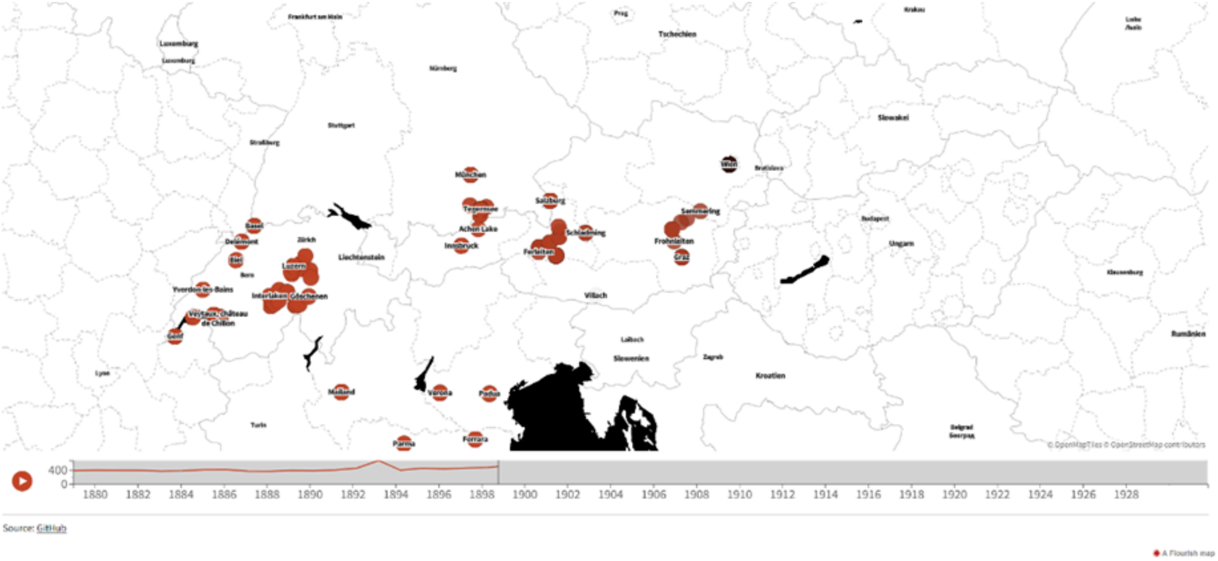
Die Aufenthaltsorte als Beziehungen auf einer Karte Europas.

Fig. 4: Screenshot of the main page offering access to the different visualizations.

Animated travel sequence

Website: https://schnitzler-orte.acdh.oeaw.ac.at/flourish.html

This visualization displays Arthur Schnitzler's stays/whereabouts during his travels along a timeline which can be played back.



Source: GitHub

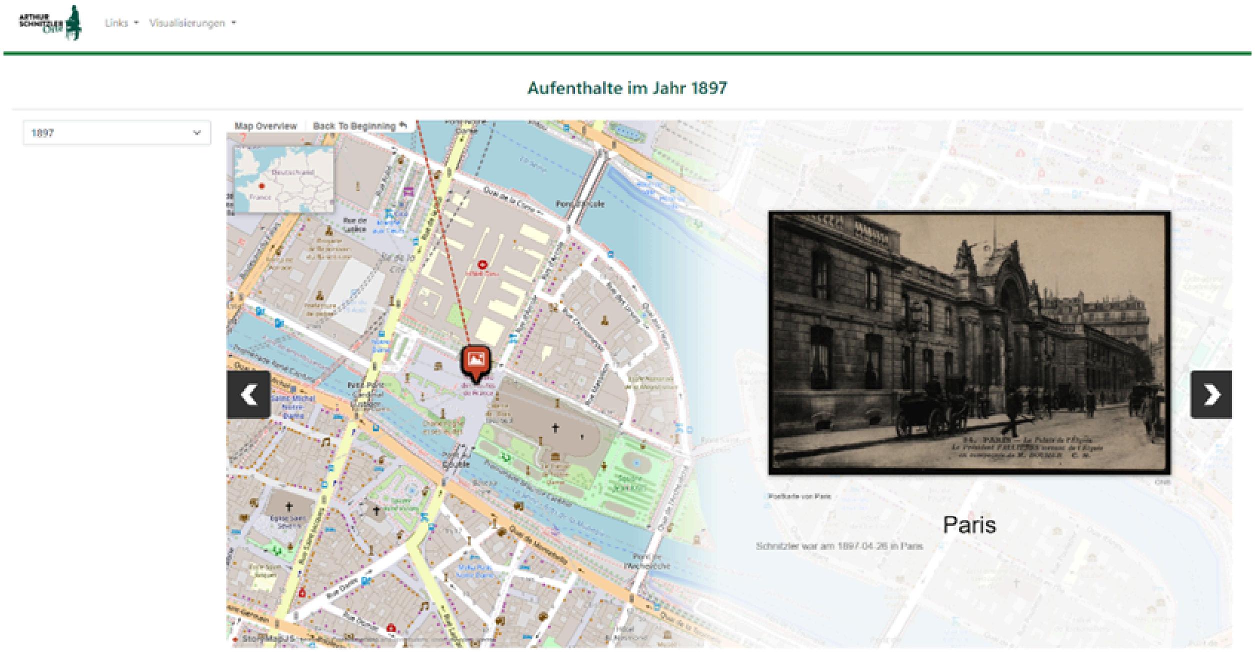
Fig. 5: Arthur Schnitzler's stays/whereabouts.

Travel destinations by year

Website: https://schnitzler-orte.acdh.oeaw.ac.at/itineraries.html

This visualization allows browsing through Schnitzler's travel destinations/stays of one year. Images of postcards of the place around 1900 accompany the experience.[10]

[10] The images are made available via the IIIF server of "AKON", the postcard portal of the Austrian National Library.



Aufenthalte im Jahr 1897

Paris

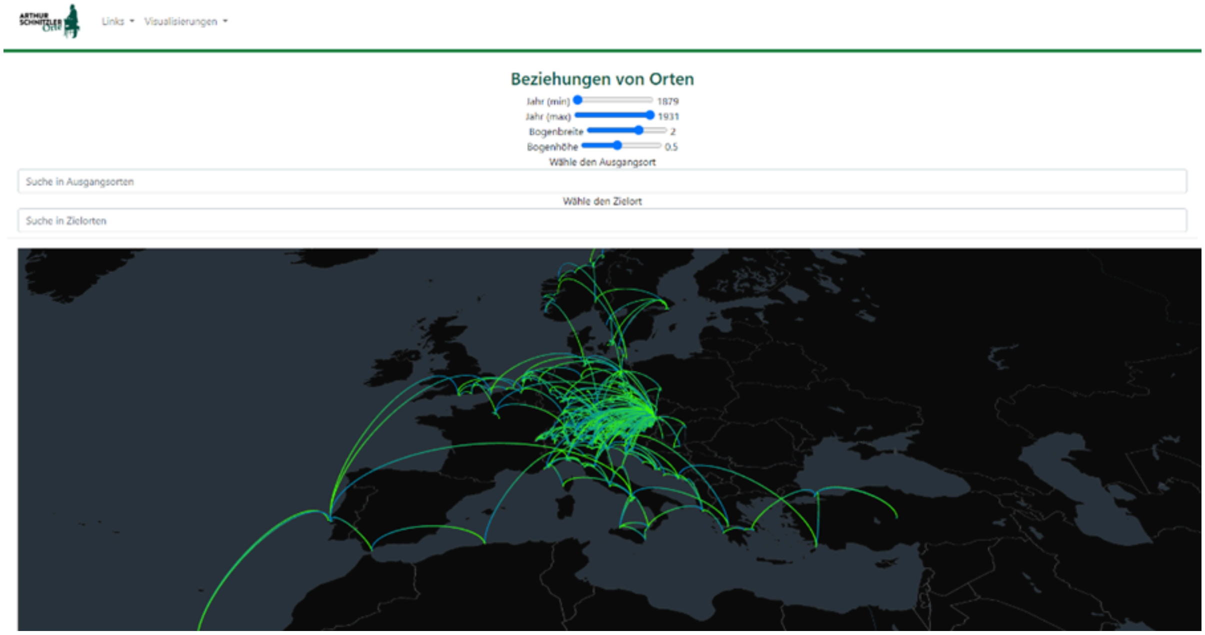
Schnitzler war am 1897-04-26 in Paris

Fig. 6: Browsing through Schnitzler’s travel destinations/stays of one year.

Relationships between the locations

Website: https://schnitzler-orte.acdh.oeaw.ac.at/arcs.html

This visualization tries to answer the question: "At which place did Arthur Schnitzler start his journey and where did he travel to?" By specifying a starting point, the embedding in the journey becomes visible. The filter option limits the time period.



Beziehungen von Orten

Jahr (min) 1879
Jahr (max) 1931
Bogenbreite 2
Bogenhöhe 0.5
Wähle den Ausgangsort
Suche in Ausgangsorten
Wähle den Zielort
Suche in Zielorten

Fig. 6: At which place did Arthur Schnitzler start his journey and where did he travel to?

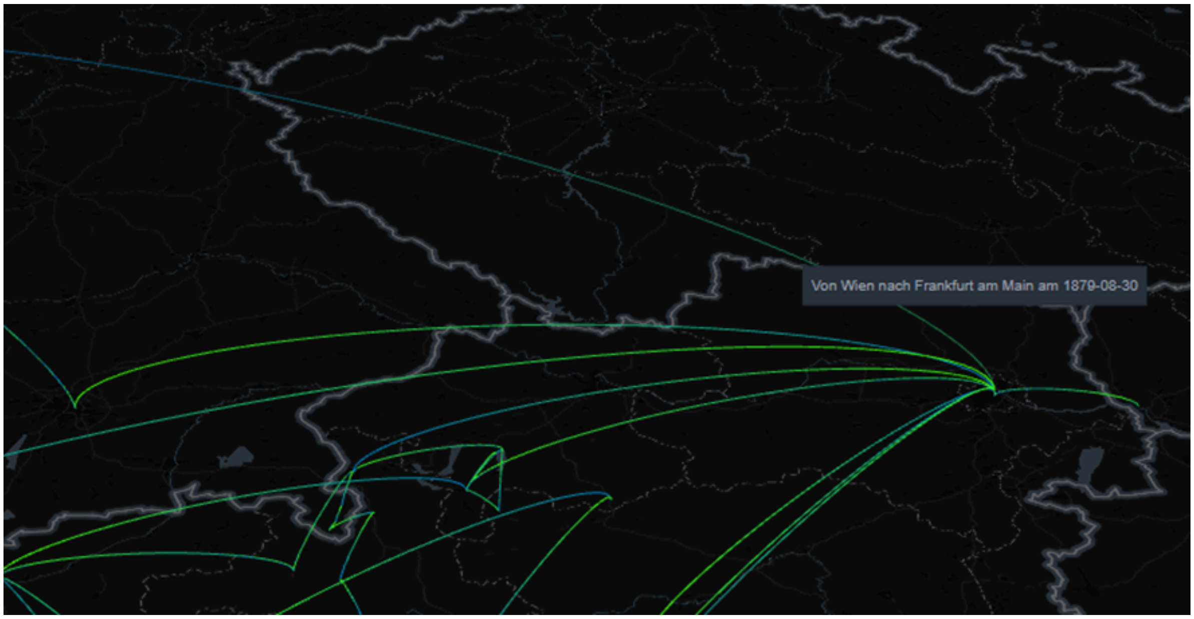


Fig. 7: Zoom in on the map.

Network representation

Website: https://schnitzler-orte.acdh.oeaw.ac.at/network.html

This visualization presents Schnitzler's travels from one place to another in the form of a network.

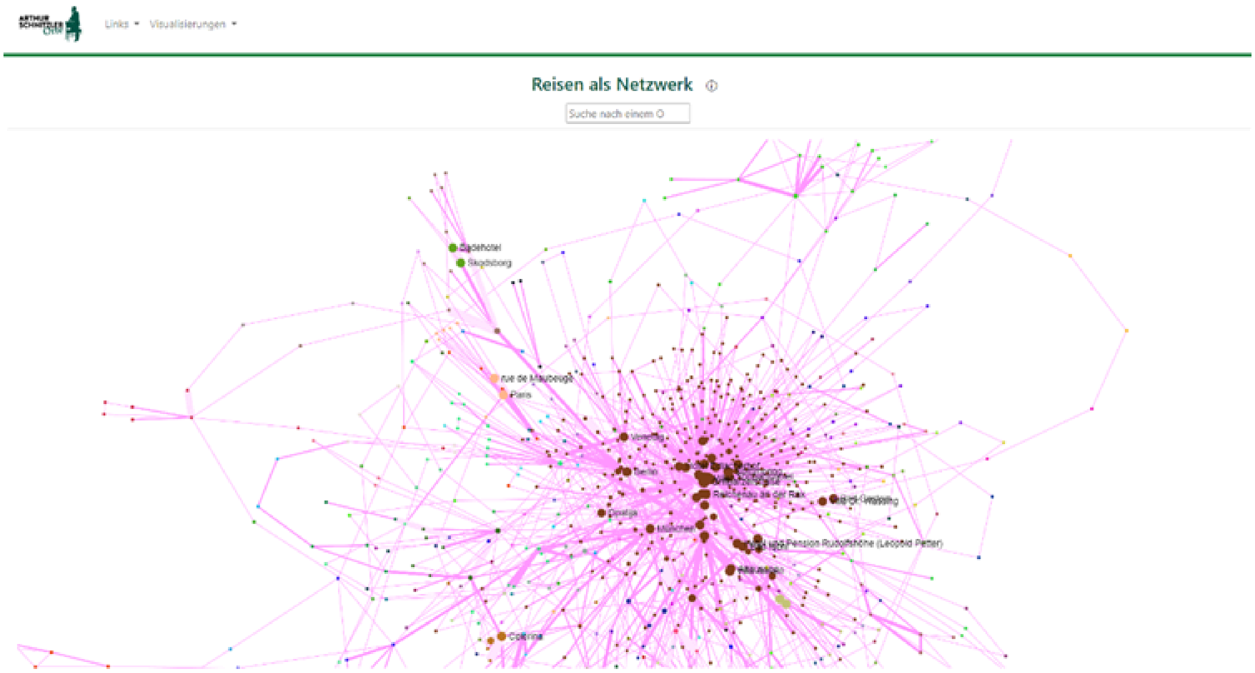


Fig. 8: Schnitzler's travels from one place to another displayed as a network.

When a location is selected, a pop-up window provides the user with the image of a postcard from the period and links to mentions of that location in Schnitzler's correspondence and diary.[11]

[11] The images are made available via the IIIF server of "AKON", the postcard portal of the Austrian National Library.

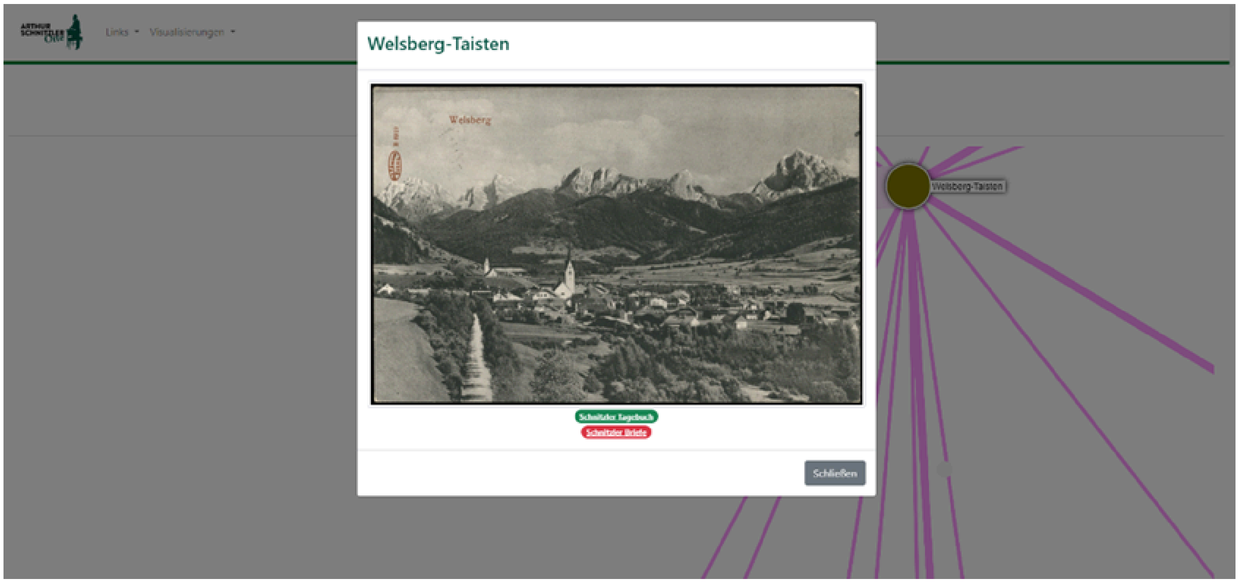


Fig. 9: Pop-up window.

Quantitative representation on a map

Website: https://schnitzler-orte.acdh.oeaw.ac.at/towers.html

This visualization provides a quantitative analysis of Schnitzler’s whereabouts (how many days he has spent in one place during his life) around the world on a map.

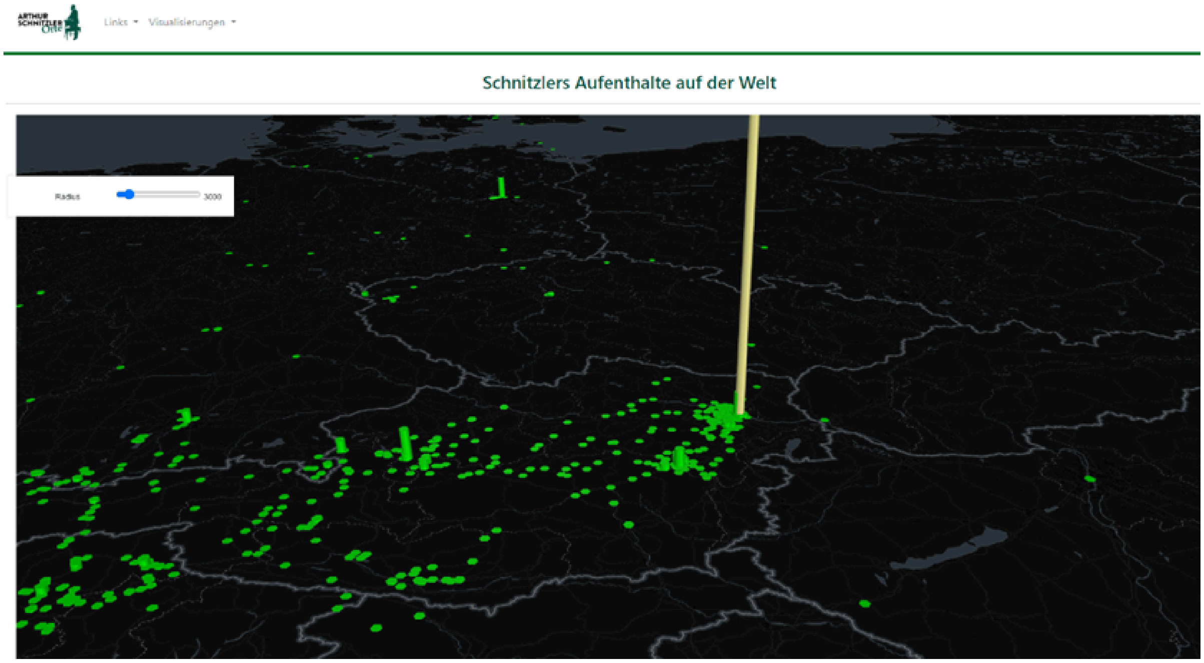
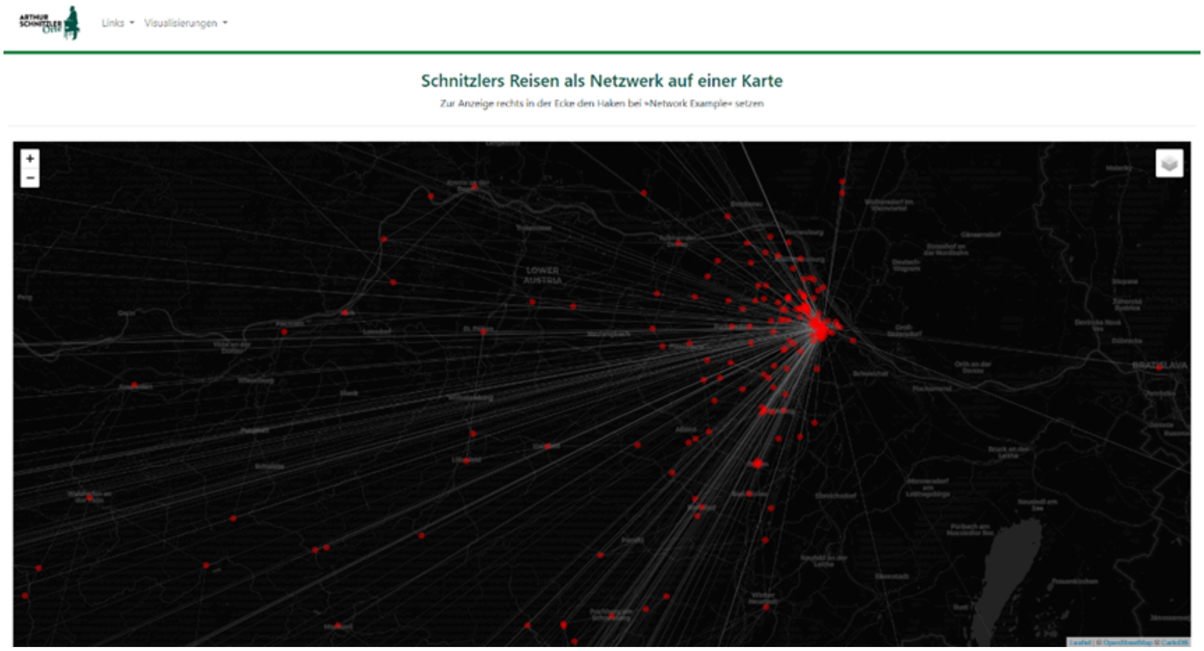


Fig. 10: A quantitative analysis of Schnitzler’s whereabouts around the world displayed on a map.

Network representation on a map

Website: https://schnitzler-orte.acdh.oeaw.ac.at/travel-net-map.html

This visualization depicts Schnitzler’s travel locations as relationships on a map of Europe.



Schnitzlers Reisen als Netzwerk auf einer Karte

Zur Anzeige rechts in der Ecke den Haken bei »Network Example« setzen

Fig. 11: Schnitzler’s travel locations as relationships displayed on a map.

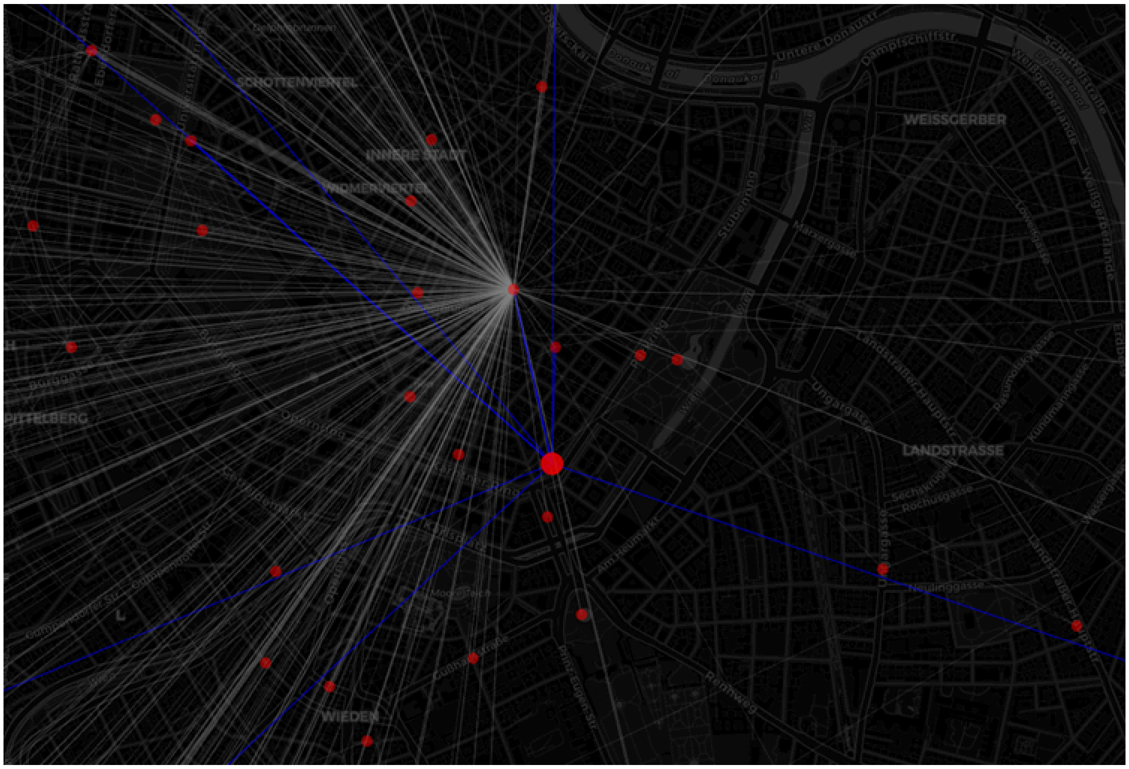
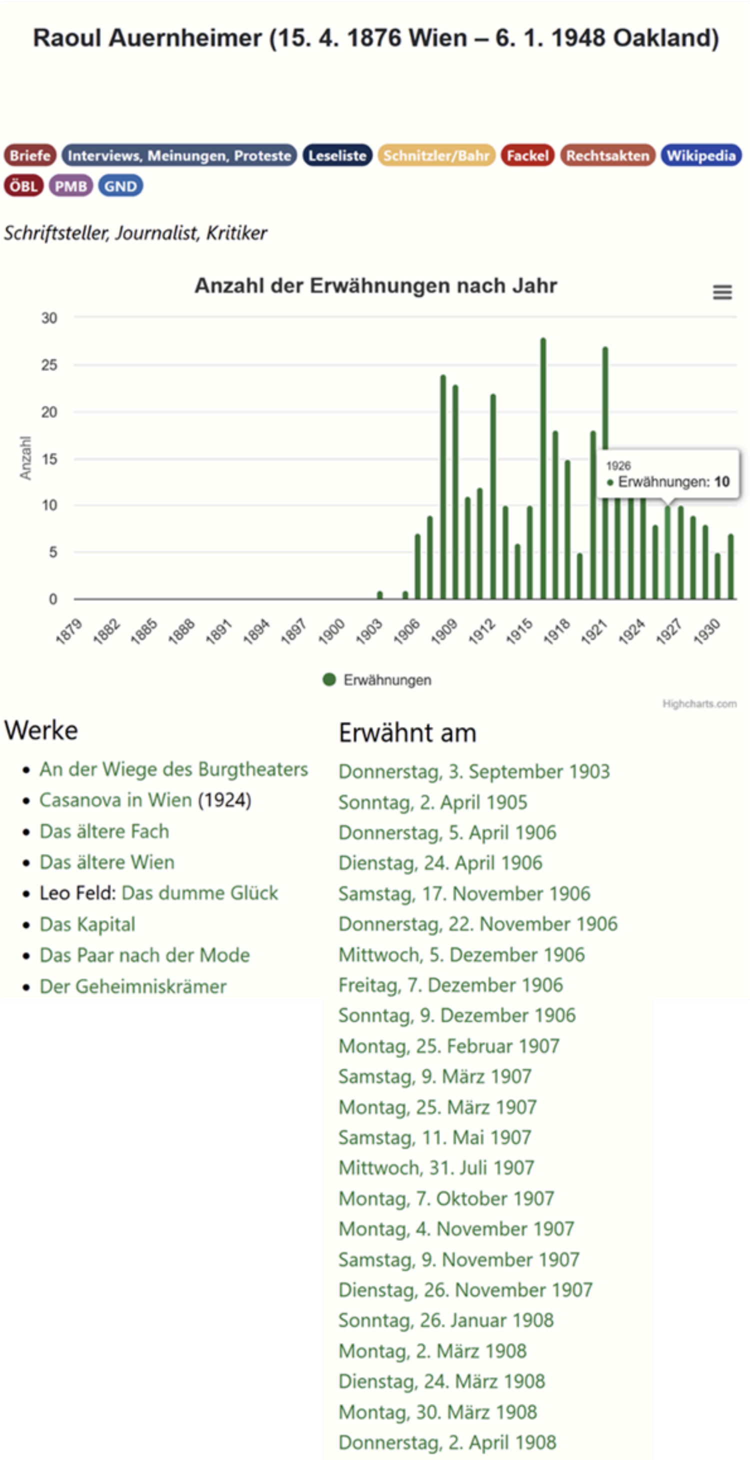


Fig. 12: Zoom in on the map.

Time

The obvious primary resource focusing on the dimension of time is the diary, which contains entries for each day of Arthur Schnitzler's adult life. The application correspondingly features a calendar as a default mode of access, allowing users to jump to any specific day. However, the semantic interlinking with persons and places mentioned on individual days, allows for an inverted view of all days on which a specific person or place was mentioned. Through integrating information from the diary and PMB, a dedicated detail page for a person features not only basic information such as birth, death and occupation, but also links to other interconnected applications. Additionally, it links to all the days a given person was mentioned, which are summarized in a chart showing the frequency of mentions per year.



Raoul Auernheimer (15. 4. 1876 Wien – 6. 1. 1948 Oakland)

Schriftsteller, Journalist, Kritiker

Werke

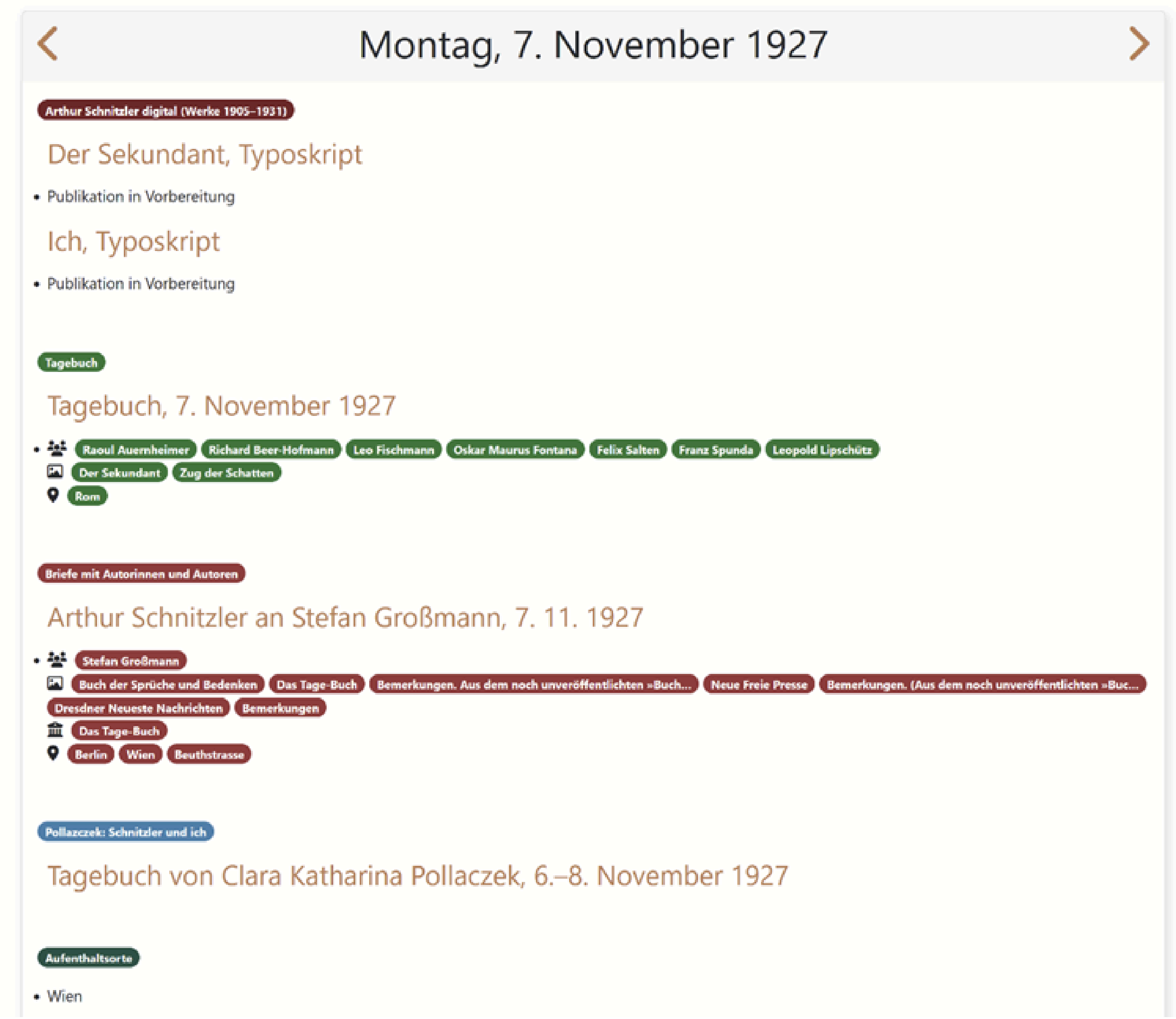
- An der Wiege des Burgtheaters
- Casanova in Wien (1924)
- Das ältere Fach
- Das ältere Wien
- Leo Feld: Das dumme Glück
- Das Kapital
- Das Paar nach der Mode
- Der Geheimniskrämer

Erwähnt am

Donnerstag, 3. September 1903
Sonntag, 2. April 1905
Donnerstag, 5. April 1906
Dienstag, 24. April 1906
Samstag, 17. November 1906
Donnerstag, 22. November 1906
Mittwoch, 5. Dezember 1906
Freitag, 7. Dezember 1906
Sonntag, 9. Dezember 1906
Montag, 25. Februar 1907
Samstag, 9. März 1907
Montag, 25. März 1907
Samstag, 11. Mai 1907
Mittwoch, 31. Juli 1907
Montag, 7. Oktober 1907
Montag, 4. November 1907
Samstag, 9. November 1907
Dienstag, 26. November 1907
Sonntag, 26. Januar 1908
Montag, 2. März 1908
Dienstag, 24. März 1908
Montag, 30. März 1908
Donnerstag, 2. April 1908

Fig. 13: Detailed view of the person Raoul Auenheimer in Schnitzler’s diary, with links to the daily entries in which he is mentioned.

Going beyond the dataset of the diary, recently the dataset "Chronik" (with corresponding application) was developed and released, which compiles and shows information for individual days from different sources.



Montag, 7. November 1927

Arthur Schnitzler digital (Werke 1905–1931)

Der Sekundant, Typoskript

- Publikation in Vorbereitung

Ich, Typoskript

- Publikation in Vorbereitung

Tagebuch

Tagebuch, 7. November 1927

- Raoul Auernheimer, Richard Beer-Hofmann, Leo Fischmann, Oskar Maurus Fontana, Felix Salten, Franz Spunda, Leopold Lipschütz
  Der Sekundant, Zug der Schatten
  Rom

Briefe mit Autorinnen und Autoren

Arthur Schnitzler an Stefan Großmann, 7. 11. 1927

- Stefan Großmann
  Buch der Sprüche und Bedenken, Das Tage-Buch, Bemerkungen. Aus dem noch unveröffentlichten »Buch..., Neue Freie Presse, Bemerkungen. (Aus dem noch unveröffentlichten »Buc..., Dresdner Neueste Nachrichten, Bemerkungen
  Das Tage-Buch
  Berlin, Wien, Beuthstrasse

Pollaczek: Schnitzler und ich

Tagebuch von Clara Katharina Pollaczek, 6.–8. November 1927

Aufenthaltsorte

- Wien

Fig. 14: An example of a calendar entry for 7.11.1927 compiling information for a given day from multiple sources.

And finally, in the application on correspondences, the pages representing some correspondence with a distinct person also features a chart visualizing the intensity of the exchange (number of letters in each direction) per year (example: Correspondence Arthur Schnitzler – Paul Goldmann), as well as more elaborate quantitative analysis for the especially extensive correspondence of Arthur Schnitzler with Hugo von Hofmannsthal comprising 645 items.

The graphic below features a quantitative analysis of the correspondence of Arthur Schnitzler with Hugo von Hofmannsthal:

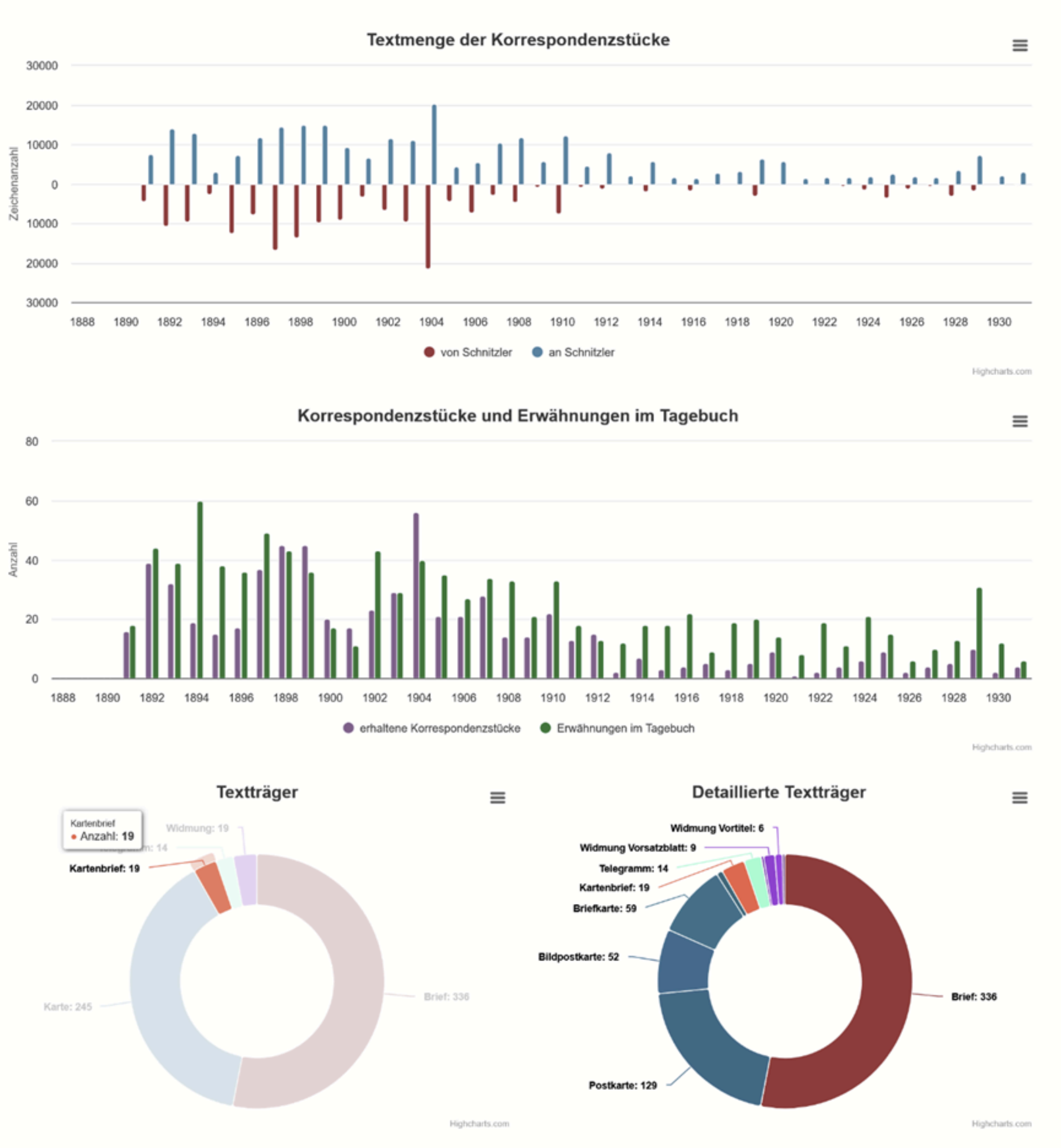


Fig. 15: A quantitative analysis of the correspondence of Arthur Schnitzler with Hugo von Hofmannsthal.

Outcomes

This showcase presents how historical non-literary textual material can be processed and made available through a set of interconnected applications, allowing the exploration of different aspects of a person's life. It was exemplified through the famous writer Arthur Schnitzler concentrating on the dimensions of time, space.

Thanks to the exemplary work of the creators – by using well-established technologies, extensive, elaborate documentation, and strict adherence to the principles of Open Science, by making all the underlying data and tooling available for reuse under permissive licenses – both the data as well as the principal approach can be easily reused.

References

- Schnitzler, A. (1981). Arthur Schnitzler Tagebuch 1879-1931. Gesamtwerk (W. Welzig, M. Neyes, R. Miklin, P. M. Braunwarth, K. Fliedl, W. Ruprechter, R. Urbach, & S. Perlik, Eds.). Verlag der Österreichischen Akademie der Wissenschaften.