Multilingual Workflows in *Bullinger Digital*: Data Curation for Latin and Early New High German

**ABSTRACT**

This paper presents how we enhanced the accessibility and utility of historical linguistic data in the project *Bullinger Digital*. The project involved the transformation of 3,100 letters, primarily available as scanned PDFs, into a dynamic, fully digital format. The expanded digital collection now includes 12,000 letters, 3,100 edited, 5,400 transcribed, and 3,500 represented through detailed metadata and results from handwritten text recognition. Central to our discussion is the innovative workflow developed for this multilingual corpus. This includes strategies for text normalisation, machine translation, and handwritten text recognition, particularly focusing on the challenges of code-switching within historical documents. The resulting digital platform features an advanced search system, offering users various filtering options such as correspondent names, time periods, languages, and locations. It also incorporates fuzzy and exact search capabilities, with the ability to focus searches within specific text parts, like summaries or footnotes. Beyond detailing the technical process, this paper underscores the project’s contribution to historical research and digital humanities. While the *Bullinger Digital* platform serves as a model for similar projects, the corpus behind it demonstrates the vast potential for data reuse in historical linguistics. The project exemplifies how digital humanities methodologies can revitalise historical text collections, offering researchers access to and interaction with historical data. This paper aims to provide readers with a comprehensive understanding of our project’s scope and broader implications for the field of digital humanities, highlighting the transformative potential of such digital endeavours in historical linguistic research.
1 INTRODUCTION

The extensive corpus of roughly 12,000 preserved letters of the Swiss reformer Heinrich Bullinger (1504–1575) offers more than a glimpse into the past; it serves as a vital link to understanding the intricate web of interactions which Bullinger fostered with scholars, theologians, royalty, and other notable figures across Europe during the Reformation era (Campi 2004). These letters, primarily penned in Latin – the prevailing lingua franca of the time – also include a significant number in Early New High German (ENHG). 1 Our project, Bullinger Digital,2 made this rich, multilingual heritage accessible and interactive through advanced digital curation methods.

The primary challenge in curating such a diverse and historic collection lies in its multilingual nature. Common language identification tools, like langid3 (Lui & Baldwin 2012) and the fastText-based classifier4 (Joulin, Grave, Bojanowski & Mikolov 2016), though adept at differentiating between modern languages, fall short when faced with 16th-century German. Addressing this gap, our project develops and implements bespoke methodologies for language identification, setting the stage for further text processing, including topic modelling and semantic indexing (McGillivray, Buning, & Hengchen 2019; van den Heuvel 2019).

Our work transcends mere data preservation in the spirit of van Miert, Hotson, and Wallnig’s5 vision for the Republic of Letters. We increased the accessibility and interoperability of Bullinger’s correspondence, enhancing its utility for a broad spectrum of researchers. This paper presents a comprehensive analysis of our two-year journey in the Bullinger Digital project, outlining the processes and challenges encountered in managing a multilingual corpus and offering insights into future enhancements aligned with the latest developments in digital humanities.

Section 3 delves into the intricacies of managing a multilingual corpus, covering aspects like code-switching, text normalisation, machine translation, and handwritten text recognition. The following sections detail our approach:

1. General Workflow: We provide an overview of the project’s workflow, focusing on handling multilingual aspects within the Bullinger corpus.

2. Language Identification and Code-Switching: This section highlights the critical role of accurately identifying languages and detecting code-switching, a crucial precursor to further processing.

3. Normalisation: We discuss our approach to normalising ENHG text, a key component in enhancing the search system’s effectiveness.

4. Machine Translation: To facilitate access for scholars not proficient in Latin, we developed a system translating Latin to German, thus broadening the reach and understanding of Latin letters.

5. Handwritten Text Recognition: Addressing the challenge of reading and transcribing handwritten texts in Latin and ENHG, this section elaborates on developing our recognition systems.

Moreover, we make all the data that went through the aforementioned workflows publicly available for reuse by the digital humanities community.

---

1 The corpus also encompasses Greek, Hebrew, English, Italian, and French texts.
3 Refer to https://pypi.org/project/langid.
5 The Republic of Letters refers to the correspondence networks, book exchanges, and intellectual discourse across Europe during the Enlightenment. The Bullinger correspondence (and other letter exchanges during that time) can be seen as a precursor of the Republic of Letters.
2 THE GENERAL WORKFLOW

The cornerstone of the Bullinger Digital project is its comprehensive workflow, designed to transform the expansive Bullinger correspondence into a digitally accessible and analysable format. This section offers an overview of our general workflow, setting the stage for the more detailed, language-specific methodologies discussed in Section 3. A thorough description of the entire process will be detailed in future publications, focusing on its implications for digital humanities and potential for application in similar projects.

2.1 THE INDEX CARD CATALOGUE

Our journey began with digitising an index card catalogue compiled by the Swiss Reformation Studies Institute at the University of Zurich (highlighted in blue box (1) in Figure 1). This physical catalogue, encompassing approximately 10,000 index cards, represents each preserved letter from or to Heinrich Bullinger, which has not yet been edited. Each card details key metadata such as author, recipient, date, signature, and incipit. The process began with the scanning and OCR (Optical Character Recognition) of these index cards, followed by manual correction through a crowd-sourcing campaign. This enabled us to build a relational database (DB) and a content management system (CMS), which acted as the “master hub” for all subsequent steps.

2.2 SEARCHABLE PDF LETTERS (teoirgsed.ch)

The Bullinger Digital progression from the metadata-rich index catalogue to the digital representation of the actual letter texts required attention to detail. From 1974 to the present, 20 volumes (Gabler et al. 1974–2020) containing 3,100 letters from the Bullinger correspondence have been published. These editions feature both the text of the letters and a comprehensive apparatus of textual and factual criticism. This apparatus, manifested as footnotes, includes editorial remarks on textual alterations (additions, deletions, and marginal notes) and factual comments that provide historical and biographical contexts.

---

6 The incipit refers to the initial sentence of a letter, excluding the greeting. It can be used for the identification of a letter.

7 The “master hub” integrated results from later stages, leveraging unique letter IDs created in the DB for efficient data management.

8 The editorial effort over nearly five decades resulted in approximately 82,000 comments.
However, the initial challenge lay in the fact that the earlier editions (volumes 1 to 7) were not digital-born. As the demand for an electronic version rose, the Swiss Reformation Studies Institute sought the expertise of a company specialising in information systems for historical documents. Their role was to digitise the 20 printed editions using OCR technology to extract the texts (represented by the green box (2) in Figure 1). The outcome was an online search system, allowing users to navigate through both the PDF scans and the OCRed texts.

A critical aspect of this digitisation process was preserving historical characters, particularly those in Early New High German (ENHG) and Latin. During the OCR, characters such as over u (˚u), frequently found in ENHG, and the e-caudata (ę) in Latin were lost (as illustrated in Figure 2). We employed the ABBYY Recognition Server to tackle this issue, effectively recovering the e-caudata but still missing many ENHG-specific characters. We generated custom training data from the digital-born editions (volumes 8 to 20) where these characters were present to address this shortfall. The process involved sampling 70 pages from each volume and creating clean PDFs devoid of metadata, footnotes, lines, and page numbers. This data was then fed into the Transkribus system (Mühlberger, Seaward, Terras et al. 2019), where we trained a multilingual model on approximately 40,000 lines and evaluated it on an additional 4,000 lines. With its impressive 0.09% character error rate (CER) on the validation set, we then applied the model to the earlier volumes (1 to 7) of the Bullinger edition.

Combining the strengths of ABBYY and Transkribus allowed us to reintroduce the lost historical characters into the letter texts, a crucial step in transforming the existing PDF edition into a truly digital edition, conforming to common XML standards. This enhancement not only preserved the authenticity of the texts but also laid a robust foundation for the multilingual workflows presented in the next section.

2.3 XML OUTPUT AND DATA AVAILABILITY

This section elucidates the output format and availability of the Bullinger Digital corpus consisting of 12,000 letters (of which 3,100 are edited, a further 5,400 transcribed, and the remaining 3,500 available only as metadata entries). After successful digitisation and text recovery, the project’s enriched data were systematically structured into an XML format. This format ensures uniformity and compatibility with digital humanities standards and enhances the corpus’ accessibility and interoperability with other projects.

Crucially, the entire corpus, in its XML format, is openly accessible on GitHub, fostering collaborative scholarly engagement and further research. Additionally, the project’s dedicated download site provides direct links to the letter corpus in the TEI format (see the dataset description at the end of this article), simplifying access for researchers interested in exploring or reusing the data. This open availability underscores our commitment

---

9 See https://www.arpa.ch.
10 The system is accessible at http://teoirgsed.uzh.ch.
11 Currently known as ABBYY FineReader Server (version 14.0), see https://www.abbyy.com/de/finereader-server.
12 Throughout the project, we employed a custom XML format for each letter, managed via a GitHub repository to facilitate collaborative editing.
14 See https://www.bullinger-digital.ch/about/project/downloads.
to supporting digital humanities research. It encourages the integration of the Bullinger correspondence with other similar historical letter collections, potentially through platforms like TEI Publisher, thereby contributing to a broader understanding of historical correspondence networks.

### 2.4 DATASET DESCRIPTION

**Object name** bullinger-korpus-tei  
**Format names and versions** TEI XML  
**Creation dates** 2023-04-03  
**Dataset creators**  
- Martin Volk, Department of Computational Linguistics, University of Zurich, Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Validation, Supervision.  
- Eyal Dolev, Department of Computational Linguistics, University of Zurich, Data curation, Investigation. Lukas Fischer, Department of Computational Linguistics, University of Zurich, Data curation, Investigation, Formal analysis, Software, Resources.  
- Raphael Müller, Department of Computational Linguistics, University of Zurich, Data curation, Investigation, Visualisation, Software.  
- Patricia Scheurer, Department of Computational Linguistics, University of Zurich, Project administration, Supervision, Data curation, Resources.  
- Bernard Schroffenegger, Department of Computational Linguistics, University of Zurich, Data curation, Investigation, Software.  
- Raphael Schwitter, Swiss Reformation Studies Institute, University of Zurich, Data curation, Resources, Conceptualisation.  
- Phillip Benjamin Ströbel, Department of Computational Linguistics, University of Zurich, Data curation, Investigation, Methodology, Formal analysis, Resources.  
- Benjamin Suter, Data curation, Investigation, Formal analysis.  
**Language** Latin, Early New High German, Greek, Hebrew, French  
**License** CC BY-NC 2.0  
**Repository name** GitHub  
**Publication date** Last update: 2023-08-11

### 3 TEXTUAL WORKFLOWS AND MULTILINGUALITY

Building on the foundation laid in Sections 2.2 and 2.3, this section delves deeper into the multilingual workflows of the Bullinger Digital project. Our efforts go beyond merely reintroducing special characters to multilingual OCR; they encompass a series of inherently multilingual processes contributing to the subsequent stages of text curation and analysis.

Given the language diversity of the Bullinger correspondence (mostly Latin and Early New High German, but also Greek, Italian, French, and little Hebrew and English), it was imperative to develop robust workflows capable of handling such linguistic complexity. These workflows ensure the accurate representation of the original texts and enhance their accessibility and interpretability.

By addressing the challenges of multilinguality, our project opens up new avenues for exploration in historical linguistics, digital humanities, and computational text analysis. The following subsections detail the methodologies and technologies employed in managing this multilingual corpus to advance digital scholarship and foster a deeper understanding of historical texts.
3.1 LANGUAGE IDENTIFICATION AND DETECTING CODE-SWITCHING

3.1.1 Language Identification

In the Bullinger corpus, which contains edited letters and scholarly transcriptions as indicated in boxes (2) and (3) in Figure 1, we encounter texts in several languages comprising a total of approximately 5.5 million tokens. After sentence boundary detection, as illustrated in the purple box (5) in Figure 1, language identification and code-switching detection become paramount. This necessity stems from the corpus’s linguistic value for historical linguistics, offering insights into the evolution of Late Latin and Early New High German (ENHG).

Despite the prowess of common language identification tools, their efficacy in historical contexts remains limited, often misclassifying languages such as Latin. For instance, evaluations of langid on Latin texts like Caesar’s *De Bello Gallico* showed only an 86.6% accuracy level (Volk et al. 2022). The challenge intensifies with ENHG, where even the existence of a classifier was not a given, and misclassifications with Dutch or Nordic languages were likely.

Although some classifiers allow languages to be excluded in the classification process, they still have difficulties with languages on which they were not trained. To address this challenge, we developed a custom classifier to differentiate between Latin and ENHG. We trained our classifier on 150 sentences of Latin and ENHG each. We noticed a remarkable accuracy, correctly classifying 100% of Caesar’s *De Bello Gallico*. Our classifier identified 165,500 Latin sentences and 39,600 ENHG sentences in the corpus, as highlighted in our publication detailing the ENHG data (Scheurer et al. 2022). This text corpus, amounting to 800,000 tokens, provides a substantial resource for research into this Germanic language variant, with expectations of growth to 1.2 million tokens with the inclusion of automatic transcriptions from the 3,500 non-transcribed letters (see “Original Letters” in the orange box (3) and below in Section 3.4).

Our language identification also enables us to analyse and visualise the linguistic composition of the letters. Figure 3 showcases this capability, where users can see a colour-coded representation of the text, indicating the presence of Latin, ENHG, and other languages like English, French, Greek, or Hebrew. This visualisation, accessible through the “Sprachen markieren” (EN: “highlight languages”) feature, enhances understanding of the multilingual nature of the correspondence, presenting a mix of Latin and ENHG in the displayed example.

3.1.2 Code-Switching Detection

The challenge of code-switching, where the language changes within a sentence, but also once or several times in the letter, is a prominent feature in the Bullinger correspondence, particularly when authors transition from discussing ecclesiastical or theological matters in Latin to daily life or personal affairs in ENHG.

Figure 3 exemplifies such instances. For instance, the sentence Ein g’uter Veltlyner mag hundert jaar ligen, *et nos tale bibimus*. (EN A good Veltlyner may lay (around) one hundred years, and we drink such.) seamlessly changes from ENHG to Latin, showcasing the need for word-level language classification. We addressed this by creating language-specific vocabularies and assigning a language to each word based on its occurrence in these vocabularies. Sentences with at least two consecutive tokens in a different language than the rest were marked for code-switching. This method allowed us to annotate code-switches in our XML files effectively. The reclassification of languages and updates to language boundaries in our pipeline are automatically reflected in the XML files. These files are then integrated into the database, which informs the front end to display languages in distinct colours.

Furthermore, this granularity in language tagging enhances user experience. Users can filter the corpus by language. We also implemented a feature allowing users to search for letters based on a minimum language percentage. For example, we allow searches for letters containing the word wyn (EN wine) and say that the letter must be written to at least 50% in ENHG. An envisioned future enhancement includes a filter to isolate letters exhibiting code-switching, a valuable tool for linguists.

---

15 E.g., there are words which occur in both languages, such as ENHG *dies* (EN this) and Latin *dies* (EN day).
16 The performance of language identifiers varies with the string length as noted in Volk and Clematide (2014).
17 Veltliner is a type of wine.
3.2 NORMALISATION OF EARLY NEW HIGH GERMAN

Normalisation improves search engine efficacy within the Bullinger Digital corpus. Techniques like stemming (e.g., reducing the Latin word dominus to domin), and lemmatisation are typically employed. Lemmatisation reduces various forms of a word to its basic lemma, enhancing the searchability of related articles. For example, reducing the occurrences of DE Häuser, Häusern, Hauses, House to the lemma Haus, EN house ensures that all articles in which the aforementioned variations of Haus occur can be found by simply entering the query Haus.

Historical texts, particularly ENHG, pose challenges in normalisation due to the lack of standardised spelling rules. For instance, users searching for occurrences of wine might not know its various spellings in ENHG. They might use the modern German term Wein, unaware of historical variants like win or wyn. Without normalisation, these variants would lead to missed results. Our goal was to enable modern German terms to be used effectively for searching through ENHG texts.

Given the complexity of normalising ENHG, our approach was akin to a translation task, inspired by the perspective proposed by Sahle (2013) and employing a Transformer-based sequence-to-sequence model (Vaswani et al. 2017). The lack of parallel training data led us to create a synthetic corpus, applying character substitutions to modern German to simulate ENHG. While not producing authentic ENHG, this approach generated a practical basis for our normalisation model. For example, and sticking to our wine example, the modern German word for wine is Wein. Knowing that in ENHG it was spelt wyn (among others), we define a rule that substitutes ei → y and lowercases W → w, thus generating Wein → wyn.

The normalisation model, trained on 84 MB of text from the period between 1900 and 1999 from the Deutsche Text Archiv, achieved a word error rate of 14.2%, aligning with other state-of-the-art methods (Makarov & Clematide 2020). Figure 4 provides an example from the Bullinger letters, showing both the original ENHG form and its normalised version.

Further, we applied this model to the ENHG sentences identified in the previous language identification step (see Section 3.1 and purple box (5) in Figure 1). In Figure 4, we note a few errors in the normalisation process. We see that the model commits four errors (in red):

18 For an in-depth exploration of this issue, Bollmann (2019) provides a comprehensive review.
19 See https://www.deutschestextarchiv.de.
20 Access the repository for ENHG normalisation at https://github.com/besou/textnormalization.

Figure 3 Visualisation of code-switching on the Bullinger Digital website, illustrated with a letter from Johannes Fabricius Schmid to Heinrich Bullinger (29 November 1563) (Bullinger Digital, 2023). Users can choose to highlight languages in different colours.
Despite the errors, the normalisation increases the number of relevant hits significantly. Figure 5 shows an example of this, where a search for *wein* yields hits for various word spellings, including *win* and *wyn*.

Regarding normalisation, stemming for Latin might be more effective due to fewer variations within word stems than for German. However, we plan to utilise lemmatisation, leveraging the abundant data available for training such models, potentially collaborating with resources like the LiLa project. This promises to further enhance search accuracy and user experience.

### 3.3 Machine Translation and Multilingual Summaries

This section emphasises the reuse potential of the translations and clarifies the methodologies applied. Acknowledging Heinrich Bullinger’s historical influence, we expand on the relevance and accessibility of his works to a global scholarly community, particularly in English-speaking countries.

Given the intricate nature of Bullinger’s Late Latin, we developed a system for dynamic translations, improving as our machine translation model evolves (detailed in the red box (4) in Figure 1). Our approach focused on German translations. We elucidate our data collection, training procedures, and model improvements in Fischer, Scheurer, Schwitter, and Volk (2022). Our methods, including pre-training and normalising Latin segments, achieved notable accuracy, clearly outperforming GoogleTranslate on our test letters (by 2.5 BLEU points).

Recent advancements, like OpenAI’s GPT 3.5 and 4 models (OpenAI 2023), demonstrate superior translation quality plus robustness concerning mixed-language letters. GPT produces fluent English or German output even when translating letters that mix Latin and ENH German sentences.

---

1. *thū gind* should be *tun or tun würden* (EN do, would do).
2. *rü ms* should be *Ruhm* (EN fame).
3. *das* should be *dass* (EN that).
4. *sy* should be *sie* (EN they).

---

21 Refer to the LiLa project at [https://lila-erc.eu/output](https://lila-erc.eu/output).
22 Scholars already translated works by Heinrich Bullinger to English during his lifetime (Opitz, 2004), indicating Bullinger’s influence on the Reformation in, e.g., England.
Initial experiments also indicate the power of recent Large Language Models to summarise the letters in modern languages. For example, a prompt like “Generate a paragraph-by-paragraph summary in English of the following letter by Heinrich Bullinger to Oswald Myconius, which is written in 16th century German and Latin:” results in surprisingly fluent and accurate texts (see Table 1). Such automatically produced summaries serve well for searches over the letters (minor errors notwithstanding) and as a basis for further professional editing. Please note the header-like start phrases for each GPT-generated paragraph.

In more detail, we will evaluate GPT’s accuracy and vocabulary appropriateness, identifying its limitations in historical contexts. The human summary in Table 1 shows the added value of disambiguating comments like “Sanherib = Kaiser Karl V.” which is difficult to achieve automatically. Our future work will include fine-tuning GPT to optimise the historical accuracy and scholarly utility.

<table>
<thead>
<tr>
<th>LETTER IN ENG GERMAN AND LATIN</th>
<th>HUMAN SUMMARY (GERMAN) FROM THE BULLINGER EDITION</th>
<th>GPT-4 SUMMARY (ENGLISH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nostri officii fuerit in tantis calamitatisibus inhortari populum ad preces assiduas et ad poenitentiam. Certo, certo inveniet et nos malum istud!</td>
<td>[4] Im gegenwartigen Unglück muss das Volk zum Gebet und zur Buße angehalten werden, denn bestimmt wird nun auch die Eidgenossenschaft bedroht! [4] Religious Duty and Moral Exhortation: In light of these calamities, Bullinger feels it’s their duty to urge the people to constant prayer and repentance, with a certain belief that these troubles will also befall them.</td>
<td></td>
</tr>
</tbody>
</table>

3.4 TRAINING OF HANDWRITTEN TEXT RECOGNITION FOR MULTIPLE LANGUAGES

Lastly, we summarise our work on automatic handwriting recognition in the Bullinger corpus. In our workflow (orange box (3) in Figure 1), we focus on the automatic recognition of handwriting in the remaining 3,500 untranscribed letters, utilising a large corpus of 8,700 manually transcribed letters as a foundation for training our models. We analysed the texts with our language identifier (see Section 3.1) and found that in our corpus of ~165,000 lines, 81% were in Latin and 19% in ENHG. In total, this amounts to 1.3 million words. Figure 6 presents a line image and its transcriptions. This example is also an instance in which we see a code-switch from Latin to ENHG.

Recognising the prevalence of Latin and Early New High German in the corpus, we employed the Transformer-based TrOCR model (cf. Li et al. (2021)), fine-tuning it for varying epochs and language combinations. Our experiments (detailed in Ströbel (2023)) reveal intriguing insights into the character error rates (CER) across different writer groups and language sets, as depicted in Figure 7.
The key takeaway is the pivotal role of multilingual data in reducing error rates, as a monolingual dataset significantly compromises accuracy. Our findings also underscore that the quantity of training data has a more pronounced impact on HTR performance than the language itself. The TrOCR model, refined for the Bullinger corpus, produced transcriptions with an average character error rate of 7.06%, which we integrated into the searchable Bullinger Digital system.

The TrOCR model, refined for the Bullinger corpus, produced transcriptions with an average character error rate of 7.06%, which we integrated into the searchable Bullinger Digital system.

The TrOCR model, refined for the Bullinger corpus, produced transcriptions with an average character error rate of 7.06%, which we integrated into the searchable Bullinger Digital system.

The TrOCR model, refined for the Bullinger corpus, produced transcriptions with an average character error rate of 7.06%, which we integrated into the searchable Bullinger Digital system.

The model is available in the Hugging Face model hub: https://huggingface.co/pstroe/bullinger-general-model. The data is available on GitHub: https://github.com/pstroe/bullinger-htr.

Figure 6 TrOCR model outputs demonstrating multilingual capabilities with an example of a language switch from Latin to Early New High German.

Figure 7 Comparative analysis of different HTR models, showcasing the efficacy of our multilingual approach in the Bullinger corpus.
4 CONCLUSION

This project underscores the critical role of multilingualism in historical letter editions, extending from character representation to language identification, code-switching detection, normalisation, machine translation, and handwritten text recognition. A crucial insight from our experience is the language-specificity of workflows in dealing with historical multilingual texts. For instance, the normalisation process for Early New High German (ENHG) developed in Bullinger Digital is not transferable to Latin.

While our initial focus was on establishing a robust textual foundation, we now plan to explore more advanced aspects. It is imperative for researchers in similar fields to recognise the importance of meticulous construction (or pre-processing) of textual bases, as these form the cornerstone of subsequent analyses and applications.

Looking ahead, we aim to develop additional tools to enhance the functionality of our web application. This includes text classification for topic identification across languages, aiming for a language-independent tagging system that can accommodate both Latin and ENHG while providing outputs in modern German and eventually English. Another ambitious goal is creating a unified semantic space for Latin and ENHG vocabularies, facilitating intuitive search capabilities. In that way, a search with the modern German word Papst (EN pope) would return letters containing ENHG pabst and Latin papa. Furthermore, integrating machine translation with AI-driven summarisation, such as leveraging GPT models, presents a promising avenue to support human editing process by automating summary generation in multiple languages for each letter, thereby streamlining scholarly research and analysis.

In conclusion, our project not only contributes to computational linguistics but also offers insights and tools for theologians and historians. We have taken steps to ensure the sustainable accessibility and applicability of our work, aiming to inspire and aid future endeavours in the digital humanities.

ACKNOWLEDGEMENTS

We want to thank all the researchers and student assistants who have been involved in this project and have contributed substantially to it. These are: Tobias Hodel, David Schoch, Anna Janka, Myrjam Hostettler, Laura Strub, Beat Wolf, Jonas Widmer, Andreas Fischer, Anna Scius-Bertrand, Martin Spoto, Angela Heldstab, Anna Mogillo, Peter Opitz, Raphael Schwitter, Judith Steiniger, Reinhard Bodenmann, David Mache, Paul Neuendorf.

FUNDING INFORMATION

We gratefully acknowledge project funding provided by various sponsors through the UZH Foundation (see www.bullinger-digital.ch/about).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Phillip Benjamin Ströbel, Data curation, Investigation, Methodology, Formal analysis, Resources, Writing – original draft.

Lukas Fischer, Data curation, Investigation, Formal analysis, Software, Resources. Raphael Müller, Data curation, Investigation, Visualisation, Software.

Patricia Scheurer, Project administration, Supervision, Data curation, Resources. Bernard Schroffenegger, Data curation, Investigation, Software.

Martin Volk, Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Validation, Supervision.
AUTHOR AFFILIATIONS

Phillip Benjamin Ströbel orcid.org/0000-0003-2063-5495
Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

Lukas Fischer
Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

Raphael Müller
Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

Patricia Scheurer
Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

Bernard Schroffenegger
Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

Benjamin Suter
Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

Martin Volk orcid.org/0000-0002-2063-4516
Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

REFERENCES


TO CITE THIS ARTICLE:


Submitted: 16 October 2023
Accepted: 20 December 2023
Published: 24 January 2024

COPYRIGHT:
© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.