From Semi-Digital Edition to Historical NLP Resource:Constructing and Annotating Historical Multilingual Parallel Text Collections on the TEITOK Platform

Maarten Janssen, Anna Jouravel, Piroska Lendvai
Faculty of Mathematics and Physics, Dept. of Slavic Languages and Literatures, Dept. of Digital Humanities
Charles University, University of Freiburg, Bavarian Academy of Sciences
janssen@ufal.mff.cuni.cz, anna.jouravel@slavistik.uni-freiburg.de, piroska.lendvai@badw.de

Abstract

We present a workflow for transforming a semi-digital scholarly edition into a multilingual historical NLP resource on the TEITOK web-based platform. As a case study, we compile a parallel collection centered on Methodius of Olympus' *De Lepra*, comprising its reconstructed Greek text, Church Slavic translation witnesses, diplomatic transcriptions, and modern German and English translations. The sources are converted to a common TEI/XML representation, aligned at sentence and word level, and enriched with Universal Dependencies annotations. We describe the workflow and integrated tools for ingesting, aligning, parallelizing and morphosyntactically annotating these materials. A central methodological contribution of the paper is the use of TEI-embedded shared identifiers for many-to-many alignment across heterogeneous historical witnesses, which makes it possible to represent non-isomorphic segmentation patterns (omissions, expansions, transpositions, etc.). In addition, we describe how automatic morphosyntactic annotation for historical Greek and Slavic was manually corrected and partially carried over to closely related witnesses. The resulting collection is searchable and visualizable in TEITOK both as individual annotated documents and as a synoptic parallel resource. Beyond the specific case study, the paper argues that TEITOK can serve not only as an interface for display and search, but as an environment for constructing reusable historical parallel corpora from philological editions and manuscript-derived transcriptions, providing support for digital humanities and historical NLP projects via transforming the input texts into parallel NLP resources, and enabling cross-fertilization and new insights by multiple research communities.

Keywords: historical NLP, parallel corpus, textual alignment, dependency parsing, web-based annotation, pre-modern Slavic, multilingual historical collection, low-resourced languages

1. Introduction

Parallel structuring of (historical) texts across their preserved versions and (modern) translations can provide the basis for applied research in many disciplines, such as Historical Linguistics, Digital Humanities (DH), and Natural Language Processing (NLP). For Historical Linguistics and DH, parallel ('synoptic') rendering of corresponding content parts across texts provides a rich visual presentation of textual witnesses in their context, and also illustrates translation solutions or editorial choices side-by-side across time and geographical space. For NLP, applying textual alignment and linguistic annotation practices on historical material pushes new advances in the field, since working with such data raises well-known challenges, including heavy orthographic and script variation, (undocumented) editorial normalization, gaps, the presence of marginal and interlinear glosses, resegmentation, translation shifts, and so on.

In this paper, we present a workflow for constructing a multilingual historical parallel text collection in TEITOK, using Methodius of Olympus' *De Lepra* as a case study. We convert the source materials to a common TEI format, and align, correct and enrich these in the TEITOK environment[1] that supports the creation, maintenance, visualization and searching of corpus data (Janssen, 2016).

The contribution of the paper is threefold. First, we provide a reusable workflow for transforming a semi-digital philological edition and related textual witnesses into a searchable historical parallel corpus. Second, we show how TEI-embedded, many-to-many alignment can represent non-isomorphic correspondences more adequately than stricter 1:1 alignment models. Third, we create a manually corrected, historically multilingual dataset with sentence-level and word-level alignment as well as Universal Dependencies (UD) paradigm token annotation for Ancient Greek and Pre-Modern Church Slavic material.

The newly created collection aims to deliver a novel gold standard dataset for sentence and word alignment and morphosyntactic parsing that facilitates interdisciplinary research. It can be accessed online at the following URL, which provides links to all deliverables, as well as additional analytical data: https://lindat.mff.cuni.cz/services/teitok-live/methodius_lepra/

Digital humanities projects increasingly publish synoptic editions of comparable sources, involving

[1] http://www.teitok.org

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1553–1561
11–16 May 2026.

historical and philological material, providing structurally closely rendered witnesses and translations. For Byzantine Greek, for instance, the *Akindynos and Palamas* site offers a digital synoptic edition of two versions of Palamas' Third Letter with side-by-side comparison and search.[2] More broadly, the University of Oslo's *Bibliotheca Polyglotta* hosts multilingual historical texts (e.g., Hebrew-Greek-Latin-English Bibles) in parallel full-text views designed for cross-version reading,[3] and the *Multilingual Parallel Bible Corpus* offers parallel representations of Bible translations in over 900 languages.[4] For Slavic historical alignment, curated parallel treebanks such as PROIEL and TOROT (notably, Old Church Slavic aligned to Greek) provide sentence- and token-level links (Eckhoff et al., 2017), which underpin empirical Old Church Slavic studies such as Eckhoff (2022).

Modern multilingual parallel corpora are able to rely largely on automatic sentence alignment to achieve parallelization. For web-scale resources, fully automatic pipelines are exemplified by OPUS (Tiedemann, 2012). Also in Slavic NLP practice, sentence alignment for contemporary parallel corpora can be done largely automatically with subsequent manual verification in core datasets; *InterCorp*, for instance, explicitly distinguishes a core with manually checked alignments from automatically processed collections, which may contain more misaligned segments, as documented.[5] The Russian National Corpus likewise presents a dedicated parallel subcorpus aligned at the sentence level.[6].

What is still less well supported, however, is a workflow that starts from heterogeneous philological source materials – including scholarly editions, diplomatic transcriptions, and modern translations – and turns them into a single, richly structured historical parallel resource that remains anchored in TEI/XML throughout. This is particularly relevant for pre-modern textual traditions, where correspondence between witnesses is sometimes partial, non-linear, and unevenly segmented. Our work is aimed precisely at this intersection between digital philology and historical NLP.

The parallel text collection presented by us differs from most of the synoptic digital editions mentioned above; in our approach, each witness is a fully fledged TEI/XML document, which can contain any complex annotation required to properly represent the witness. In TEITOK, each text is a tokenized and annotated TEI/XML file. Since TEITOK has a modular composition, this allows us to extend its main functionalities by custom scripts, which facilitates the creation of tailored solutions for the set of eight texts, featuring multiple languages and language stages, that we want to compile into a parallelized collection.

Our workflow encompasses preparing both sentence and word level alignment. Sentence level alignment is created using an automatic alignment method with posterior manual correction, employing the method of Janssen et al. (2025). XML alignment in TEITOK is created by linking nodes in the various witnesses when they represent the same text. These links can subsequently be used to generate all synoptic side-by-side views on-the-fly, as specified by the needs of the user. We subsequently enrich the parallelized texts also in terms of dependency parses, using the UD framework and its tools natively in TEITOK. We show that the resulting parallel resource is searchable and visualizable in alignment view, correction mode, as well as by fine-grained rendering of specific information, such as translation units.

The paper is structured as follows. We first describe the source materials included in the parallel text collection. Subsequently, we give a brief overview of the annotation steps that were applied to the texts to enrich them with dependency parsing information as well as with various levels of alignment between the texts. Next, we present how the parallel text collection can be used for text visualization, displaying as a synoptic edition, and search both with and without use of the parallel alignments. Finally, we draw conclusions and outline future directions.

2. Source Material

Our focus text is the treatise *De Lepra ad Sistelium* by Methodius Olympius (Migne, 1857), an early Christian bishop and theologian who was active in Lycia (Asia Minor). The treatise allegorically interprets the Old Testament regulations on leprosy (Leviticus 13), using the disease as a metaphor for spiritual and moral afflictions of the soul. The original text was written in the Old Greek language (ISO symbol: `grc`) at the turn of the 4th century CE.

This original Greek text is preserved only in excerpts within the Byzantine anthology *Florilegium Coislinianum* that gathers quotations from church fathers, assumed to be compiled in southern Italy in the 9th–10th centuries. It is organized alphabetically by books; *De lepra* appears in book "L".

[2] https://akindynos-and-palamas.ch/digital-edition/ab/

[3] https://www2.hf.uio.no/polyglotta/

[4] https://textgridrep.org/project/TGPR-d862e14d-4df7-052b-00fe-661cb242231c#README, cf. Mayer and Cysouw (2014).

[5] https://wiki.korpus.cz/doku.php?id=en:cnk:intercorp&rev=1727772179

[6] https://ruscorpora.ru/en/page/corpora-structure/

Identifier	Language	Description	Sentences	Tokens
A	grc	Edition by Sieber	52	1762
B	deu	German translation of A by Sieber	213	2225
C	orv	Edition by Jouravel	267	5068
D	deu	German translation of C by Jouravel	270	6552
E	orv	Diplomatic transcription of manuscript used for C	702	4561
F	orv	Diplomatic transcription of Great Menaion Reader manuscript	493	4596
G	deu	Translation of Slavic parts of G by Bonwetsch	241	6584
H	eng	English translation of Bonwetsch's edition by Cleminson & Eastbourne	298	9076

Table 1: Source material overview

Its textual witnesses belong to different recensions. Besides the long version in Paris, BnF, *Coislin 294* (recensio I, late 11th c.), shorter forms occur in *Parisinus graecus 924*, *Atheniensis, Bibliothecae Nationalis 464* (both 10th century), and *Bruxellensis IV 881* (1542; all three recensio II) as well as in a mixed version *Metochion 243* (18th century).

By contrast, the oldest complete witness to *De lepra* is a Church Slavic translation. Linguistic features indicate that the first translation was likely written in an East-Bulgarian translation milieu in the 10th century. The surviving copies, however, are from much later. Hence, we use the `orv` language identifier for this text content (ISO 639-3 for Old East Slavic, traditionally called Old Russian), rather than `chu` (Old Church Slavic), because this witness belongs to the Old East Slavic recension: despite its in part Church Slavic morphosyntax and lexicon, its orthography is mostly Old East Slavic that, in practice, `orv`-trained taggers parse more reliably than `chu`-based models, cf. its description in Berdicevskis and Eckhoff (2020).

According to the present state of research, *De lepra* is preserved in at least 17 manuscripts, mainly from the 16th–17th centuries, mostly in Russian collections; their relative uniformity suggests the mechanical copying of a single Church Slavic archetype. Within the Slavic *Corpus Methodianum* the treatise occupies a stable position.

The text survives in abbreviated form in relation to the Greek archetype, yet is substantially fuller than the Greek florilegial excerpts. The Slavic version preserves sections lost in Greek – notably a narrated "speech of a Lycian wise woman" which Methodius embeds as a fictive authoritative text as well as the dialogue frame; both almost certainly present in the Greek original. For an overview of the history of Greek and Church Slavic versions of *De lepra*, and its critical edition including German translation cf. Jouravel et al. (2024), Jouravel and Sieber (2024), and Maksimczuk (2024).

For western scholarship the treatise became accessible through the German theologian G. N. Bonwetsch, who at the end of the 19th century prepared a German translation of the entire Slavic text, based primarily on the manuscript *RNB Q.I. 265*, St. Petersburg, 16th c., providing the deviating readings from three further witnesses, and supplementing it with Greek portions where available (Bonwetsch, 1891). In 1917 (GCS) he printed the Greek text from *Coislin. 294* and used his earlier German rendering of the Slavic to bridge perceived gaps Bonwetsch (1917). This pragmatic approach made the treatise available to theologians, but remained provisional for philological research.

Recently, a new, semi-digital, German edition became available Jouravel et al. (2024) that was natively produced in the Classical Text Editor (CTE)[7], cf. Hagel (2007). From this edition we obtained the following five (derived) texts of *De lepra*, and exported them as structured XML:

A the Old Greek (`grc`) version

B the modern German translation of the Old Greek version

C the Church Slavic (`orv`) version based mainly on manuscript *RNB Q.I. 265*

D the German translation of the Church Slavic version

E the diplomatic transcription of the Church Slavic text as preserved in C

On top of those texts, the following additional material was included in our collection:

F a diplomatic transcription of the Church Slavic text as preserved in a 16th century copy in another manuscript, namely in *GIM, Sin. 995* (Great Menaion Reader, Uspenskiy copy, `orv`)

G a German translation of the Greek and Church Slavic texts provided in the *editio princeps*. It is published within the *Perseus Digital Library*[8] among other works by Methodius, but it does not provide the Greek original but only the German translation. However, the Greek text[9] is almost identical to the text in A, which is why we have not included it separately as a parallel representation.

[7] https://cte.oeaw.ac.at

[8] https://scaife.perseus.org/reader/urn:cts:greekLit:tlg2959.tlg004.opp-ger1:1.1-1.2/

[9] https://archive.org/details/methodiusherausg00meth/page/450/mode/2up

H an English translation of the Greek and Church Slavic texts provided in the *editio princeps*. It is taken from an online publication prepared by Pearse et al. (2015).[10]

An overview of the texts is given in Table 1, along with an indication of the extent of each text in terms of sentences and tokens. Importantly, the Bonwetsch edition is non-trivial to align, since as mentioned above, it does not provide the original texts with their corresponding translations, but rather resembles a patchwork in which the missing Greek passages have been filled in with the German translation of the corresponding sections preserved in the Church Slavic version.

3. Processing Workflow on the TEITOK Platform

The various source texts came in different formats and were all converted to TEI/XML for use in TEITOK: the English version *H* was converted from HTML to TEI using a custom script, while parts of the edition-derived material were already available in XML/TEI-like form and were normalized further for TEITOK (*A*-*E*); and the manuscript transcription material was converted from PageXML (*F*). The exported TEI/XML was subsequently manually adapted for use in the TEITOK environment, including structural normalization, tokenization checks, and the insertion of alignment-ready identifiers.

Once the various documents were added to TEITOK, they were enriched with annotations inside the TEITOK platform, with both parsing and alignment information.

3.1. Dependency Parsing

All documents were adorned with linguistic information resulting from lemmatization, part-of-speech tagging, and dependency parsing, following the Universal Dependencies annotations. The modern texts were automatically annotated using the TEITOK internal NLP pipeline that utilizes UDPipe[11]. For English, the employed model is based on the EWT treebank (english-ewt-ud-2.15-241121 Silveira et al. (2014)), and for German on the GSD treebank (german-gsd-ud-2.15-241121)[12].

For historical languages, the accuracy of UDPipe is lower than for modern languages, in part due to orthographic differences – not only in the source documents themselves, but also in the conventions that are employed for their transcription. There can be large differences between the training data that had been used to train UDPipe and the transcript to be annotated.

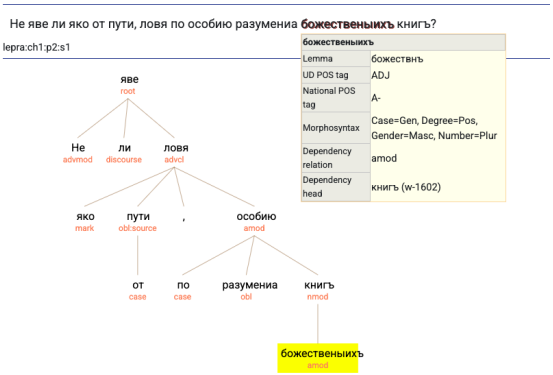


Figure 1: Corrected dependency tree in TEITOK.

Four of our historical texts were automatically annotated using the models based on the PROIEL and TOROT treebanks (Haug and Jøhndal, 2008) for both `orv` and `grc` (old_east_slavic- torot-ud-2.15-241121, resp. ancient_greek-proiel-ud-2.15-241121). In contrast with the modern texts in our data, the annotation of the historical Church Slavic and Greek texts was manually corrected by us, leading to more gold standard UD-annotated data. Manual correction covered tokenization where needed, lemma assignment, UPOS tags, selected morphological features, and dependency relations. An example of a corrected dependency tree from the (C) version is given in Figure 1. Manual correction is extremely labor intensive for historical languages, aggravated by the lack of expert annotators, which we attempted to alleviate by utilizing close text copies present in our collection, using the following strategy: selected witnesses were automatically annotated and then manually corrected. After that, word-level alignment was used to copy the annotation from these versions to their copies, and the copied annotation was manually corrected again in cases where the orthographic form was different between the source and the copies.

At the current stage, the manually corrected collection consists of the core historical witnesses – the Greek edition (A) and the normalized Church Slavic edition (C). We do not yet report intrinsic tagging or parsing scores, since the resource is still being finalized; instead, the emphasis of the present paper is on the construction workflow and on the representation of corrected annotation in a reusable parallel environment. A full quantitative evaluation of parsing and alignment quality can be found on the corpus website.

[10] https://archive.org/details/MethodiusDeLepra2015/

[11] https://lindat.mff.cuni.cz/services/udpipe/

[12] https://github.com/UniversalDependencies/UD_German-GSD

3.2. Textual Alignment

The alignment of the texts was done semi-automatically, using automatic pre-alignment with posterior manual corrections. The alignment setup was created for the purpose of this collection, based on a method described in (Janssen et al., 2025).

The representation of the alignment in TEITOK is carried out inside the TEI/XML documents of each text version, by means of a shared attribute (`@tuid`) on XML nodes. Alignment in this way can be represented on multiple levels within the same documents; in the present collection, alignment is represented at the levels of document, chapter, paragraph, sentence, and token. For example, two sentences across two TEI/XML documents are translations of each other if they share a `@tuid`.

A major challenge of this material is that diplomatic transcriptions – a usual editorial praxis in Slavic Studies – preserve historical punctuation that is not suitable for sentence- (i.e. punctuation-) based NLP processing. In addition, sentence boundaries do not align consistently across versions. Modern translations may split long historical periods into several shorter sentences; also, historical witnesses may omit textual material, while others expand or paraphrase it. For this reason, the alignment model must support not only 1:1 correspondences, but also 1:n, n:1, and n:m relations.

Since literary translation might not always involve sentence-by-sentence pairs, a single sentence can comprise a list of `tuid`s, creating a many:many mapping between sentences, or across other units. A `tuid` element can be used to link not only sentences but any other elements across versions of the same text in the same language, where they hence do not correspond to translation units, but rather to copies, creating what is typically called an *apparatus structure*. This setup is different from the traditional setup for representing text alignment, which can roughly be divided into two types: alignment by translation units , as well as pairwise alignment. Let us illustrate this difference for alignment at the sentence level.

Figure 2: Automatic alignment view in TEITOK.

Alignment by translation unit is used in formats such as TMX[13], which has a separate node for each sentence in the text (the translation unit), with inside that node the text in each language for that sentence. Any TMX file can easily be translated into the TEITOK setup, and the *teitok-tools* repository[14] contains a script that does exactly that. But TMX can only have alignments at a single level. More crucially, TMX and similar formats require a strict 1:1 mapping between translation units, which is often not the case in existing translations or copies. In contrast, the use of multiple `tuid`s per sentence makes the TEITOK setup more flexible and usable for not strictly corresponding translation data.

Pairwise alignment, used for instance by automatic alignment tools such as *hunalign* (Varga et al., 2007), is done by listing the sentences in the source text, and then for each sentence the sentence(s) in the target text that correspond to it. This pairwise alignment setup is equally expressive as the setup used in TEITOK, since in TEITOK, we could in theory prefix each pairwise alignment with a unique prefix to get a pair-based list of `tuid`s . But in practice, the TEITOK setup uses single `tuid`s or small lists of `tuid`s , creating new `tuid`s only where needed. This makes the assignment of the correct `tuid` somewhat more complicated when dealing with many non-aligning sentences, but creates an end result that is much easier to maintain in case alignments need to be corrected, which makes the resulting data somewhat easier to use in visualization and search.

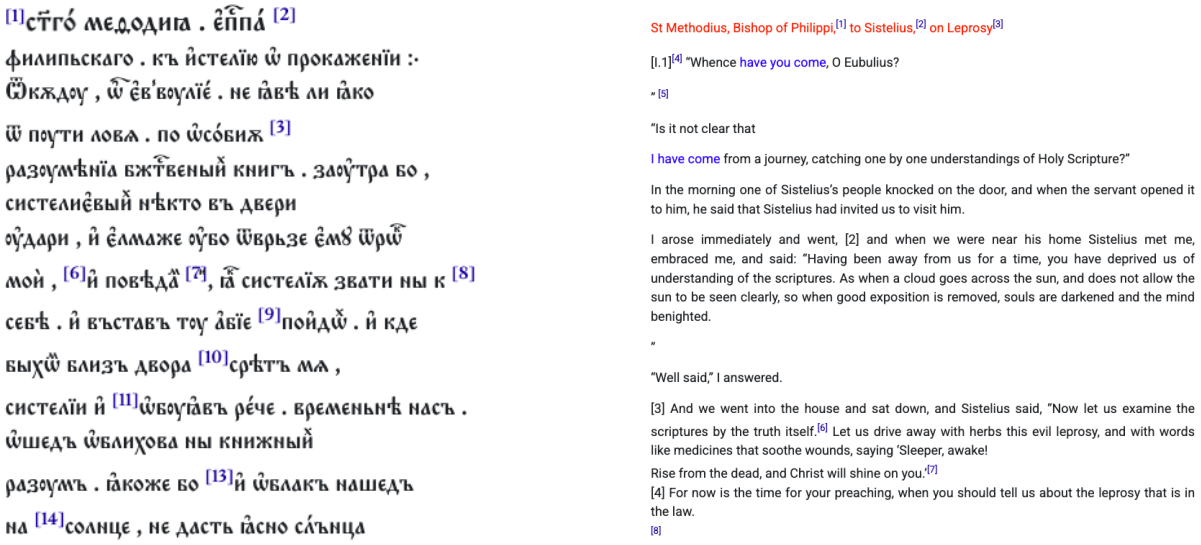
In practice, the many-to-many setup proved particularly important for three cases: (i) modern translations that regularize punctuation and sentence segmentation, (ii) witnesses with abbreviated or expanded content, and (iii) passages whose correspondence is philologically secure but structurally non-isomorphic. A stricter sentence-to-sentence model would have ignored these relations.

4. Data Display and Search Functions in the TEITOK Interface

The TEITOK platform provides a number of different interactions with the TEI/XML files. These can be divided into three different categories: the visualization of a single document, the parallel display of multiple version making use of the alignment between the versions, and searching through the collection, either a classical search or a search that makes use of alignments.

[13] https://en.wikipedia.org/wiki/Translation_Memory_eXchange

[14] https://github.com/ufal/teitok-tools



(a) Church Slavic Version (E)

(b) English Version (F)

Figure 3: Visualization with different fonts

4.1. Text Display

The standard text display in TEITOK visualizes each document by putting the raw TEI/XML into the browser, and letting the browser handle the correct visualization, using a combination of CSS and JavaScript. This visualization can handle typesetting from the original document, editor notes, line breaks, and most other information present in the TEI/XML documents. This makes for a complete digital edition of the various versions of the text.

Text visualization enables displaying the annotations for each token, either below each word, or on mouse-over. This shows all annotations that were added by UDPipe, including the lemma, the part-of-speech tag (both universal and custom), and dependency relations with their head.

One of the modifications made to the standard TEITOK view for this project was to allow for document-specific font selection (and rendering in general), since Church Slavic texts require using a font that supports the full Church Slavic character set (which is not supported most standard fonts), while the texts in modern languages are to be presented in a modern font. The result can be seen in Figure 3, where the Church Slavic text (left) is rendered differently from the English text (right).

4.2. Parallel Display

The TEITOK interface can exploit textual alignments in a number of different ways: it can show full TEI documents in parallel, it can create a table of aligned elements, and it can collect all occurrences of a given translation ID in a corpus.

4.2.1. Parallel TEI

The main visualization of parallel aligned text is to display multiple versions in different columns. Each document is displayed in full, as they are in the full document visualization, although with less options to customize the display. The parallel visualization is shown in Figure 4. The example shows only two documents in parallel, but there can be any number of versions displayed in parallel at the same time.

When moving the mouse over the text, instead of showing information about the token, it shows information about the alignment. It highlights the segments in all other versions that correspond to the part of the text we hover over in the following way: it looks for all parents of the node the mouse is over that have a `@tuid` attribute, which depending on what the user hovers over can be a word, a sentence, a paragraph, or the entire text. It then highlights the smallest of those elements, and takes the `tuid`, splitting it into individual `tuid`s if the element has more than one `tuid`, subsequently going through all the other versions that are being displayed, and highlights all the matching elements, scrolling the element into view where necessary. If no direct alignment is found in a specific version, it looks through the list of alternative `tuid`s, and highlights the smallest one for which there is a match. To indicate that the element is not directly aligned to the element highlighted in the first column, it is highlighted in a different color. In this way it is intuitive to see the alignments in their original context.

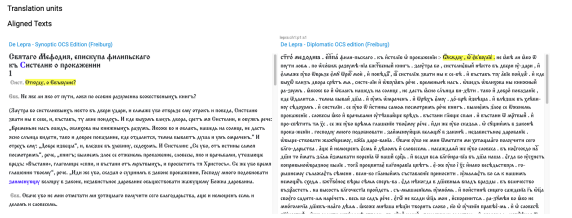


Figure 4: TEITOK: Alignment view for normalized Church Slavic text (left) and the diplomatic Church Slavic transcript (right).

4.2.2. Alignment Table

An alternative visualization creates a table from parallelized text units, where each row represents a translation unit, and each column represents the text in each version of that translation unit. This creates a visualization similar to the aforementioned *polyglotta*, but where the user has more control over which versions to display in the table, and how. This visualization makes it is easier to see the alignments, and the data model becomes close to that of TMX; these data can also be exported to TMX. It makes it intuitive to observe versions of each sentence or paragraph . This visualization is illustrated by Figure 2.

However, as explained in Section 3.2, a TMX style representation cannot capture everything that is present in the TEI/XML. In the alignment table, what is represented is alignment at a single level (by default sentence level), where the rows are taken from the first (left-most) version. This means that the sentence order is that of the specific version (where sentences might appear in a different order in other versions), and only the sentences that are present in that version are listed, while the other versions only list an equivalent if there is an aligned sentence. Any text in the first version that is not inside a sentence is not included either. The problem of sentence alignment that is not 1:1 is handled with some heuristics, so that grouped `tuid`s are correctly visualized.

4.2.3. Display by Translation Unit

The last type of visualization of parallel alignment is to search for a given `tuid` throughout the collection, and then display the full text in all the versions of that translation unit. Each text is shown as full TEI/XML fragment, with a link back to the original context in the full document visualization. This visualization is exemplified in Figure 5.

Especially for synoptic editions with many witnesses, as in our case, this is a more manageable format, since a table with many columns quickly becomes unreadable.

Figure 5: TEITOK: Display of elements by `tuid`.

4.3. Corpus Search

The TEITOK functionalities not only allow for the display of a document, but also allow searching through the collection, whereby each result is rendered as a fragment of the original TEI/XML, with all annotations inside made available. Given that this is a parallel aligned collection, there are two different search methods: a traditional corpus search using the Corpus WorkBench (CWB) (Evert and Hardie, 2011), and dedicated search tools that make use of the alignments.

4.3.1. Corpus Query Language Search

The collection can be searched using the CWB corpus search engine natively in TEITOK. The Corpus Query Language (CQL) of CWB allows for powerful and complex searches, making it possible to search for words by form, lemma, or part-of-speech tag, search for sequences of tokens, and restrict searches by metadata, e.g. to only search in texts in a specific language, or to make temporal constraints.

In our collection, each query result shows the keyword in context (KWIC) list, as well as the `tuid` of the actual sentence for the keyword. Figure 6 shows the results of a simple search for the German lemma *Aussatz* ('leprosy') in the collection, which provides a list of all the occurrences in the collection in their context. To the left of the the context fragment the link *context* enables moving the mouse over the word and displaying relevant information about the witness from which the result originates, e.g., language, author, title. Clicking on it opens a new tab that displays the full text as described in Section 4.1, with the resulting word highlighted. On the right, it shows the `tuid` of the context sentence, while clicking on that opens a window showing that sentence and all sentences in other witnesses that are aligned to it, using the interface described in Section 4.2.3.

The CQP search hence makes it easy to search through the collection to quickly find examples of words, lemmas, sequences of words or POS tags, and to navigate from any of the results to its context document, as well as to observe the returned sentence in each of its witnesses.

In order to fully exploit the alignment in the

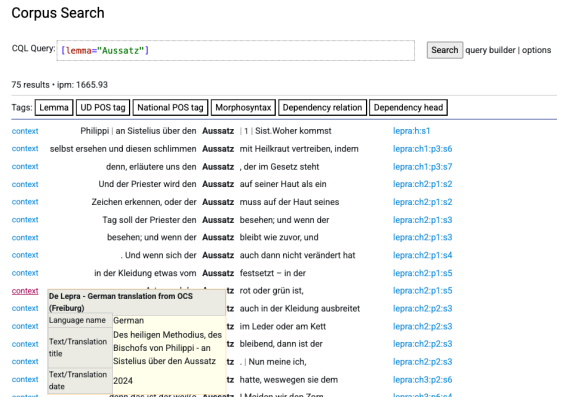


Figure 6: TEITOK: KWIC search view.

search, the collection motivated the creation and use of a new search interface in TEITOK. This interface allows searching through parallel aligned corpora, where search terms can be specified in both the source and the target document. The interface is not yet released, but the parallel search functionality should be available soon, and integrated into the corpus website.

4.4. Added value of the new resource

The resulting collection supports several types of analysis and text representation that are difficult to carry out on the source materials in isolation. From a technical perspective, it enables comparison across witnesses, diplomatic transcriptions, and modern translations within a unified environment, while retaining links to the full TEI/XML context.

From an editorial perspective, the aligned collection makes it possible to relate transmitted form, normalized representation, and editorial intervention within a single structured environment.

From a linguistic perspective, it supports contrastive querying across languages and textual layers, for instance by combining lemma- or POS-based searches with aligned context retrieval.

From a philological perspective, it creates a basis for future empirical work on historical translation techniques, lexical correspondences, and annotation transfer across closely related witnesses.

From the perspective of the history of textual transmission the collection provides a basis for investigating textual variation across witnesses and editorial stages; the resource facilitates the study of omissions, expansions, and other forms of textual displacement.

From a literary- or cultural-historical perspective, it can support the investigation of reception, rewriting, censorship, and other forms of text reshaping in different historical and social contexts.

For historical NLP, the resource is particularly useful because it combines manually corrected annotation with explicit alignment links. This makes it suitable not only for corpus exploration, but also for downstream experiments on alignment projection, parser adaptation, and low-resource multilingual modeling.

5. Conclusion and Future Work

We have shown how a recently published, semi-digital, multilingual CTE edition can be converted into a fully digital, synoptic resource. We aligned TEI/XML sources and served those in TEITOK, with UD-style token annotation and many:many links across textual witnesses and layers.

Our novel technical contribution is the implementation and description of a workflow that bridges the printed book and the digital platform, with amended functionalities. The tools support reliable character conversion, correction of linguistic annotations for pre-modern Greek and Church Slavic, as well as their faithful display so that all witnesses can be visualized, compared and searched. We extended and customized the interface to make the resulting edition as useful as possible for our target users and potentially across the Historical Linguistics, DH, and NLP communities.

Our key methodological contribution is that the conversion of the existing transcriptions of the witnesses to a common format (TEI), their correction and enrichment in an existing tool (TEITOK) lead to a versatile digital synoptic edition. For philologists, this yields a trustworthy, searchable material for exploring textual transmission, for corpus and computational linguists it offers parallel, well-documented data suitable for alignment and modeling, with clean provenance back to TEI.

Our key data contribution is an emerging gold standard data collection; we are releasing the data as a new aligned collection of historical multilingual documents and plan its transformation into a UD treebank.

In future work, we plan to extend the collection with additional witnesses, incorporate further layers from the scholarly edition such as apparatus material and introductory matter, and release that extended dataset in a versioned public repository together with documentation of the conversion and alignment workflow.

6. Bibliographical References

Aleksandrs Berdicevskis and Hanne Eckhoff. 2020. A diachronic treebank of Russian spanning more than a thousand years. In *Proceedings of*

the Twelfth Language Resources and Evaluation Conference, pages 5251–5256, Marseille, France. European Language Resources Association.

Gottlieb Nathanael Bonwetsch. 1891. *Methodius von Olympus*. Deichert, Erlangen, Leipzig.

Gottlieb Nathanael Bonwetsch. 1917. *Methodius*. J.C. Hinrich, Leipzig.

Hanne Eckhoff, Kristin Bech, Gosse Bouma, Kristin Eide, Dag Trygve Truslew Haug, Øyvind B. Haugen, and Marius L. Jøhndal. 2017. The proiel treebank family: a standard for early attestations of indo-european languages. *Language Resources and Evaluation*, 52(1):29–65.

Hanne Martine Eckhoff. 2022. First attestations. an old church slavonic sampler. In Imke Mendoza and Sandra Birzer, editors, *Diachronic Slavonic Syntax: Traces of Latin, Greek and Church Slavonic in Slavonic Syntax*, pages 255–302. De Gruyter Mouton, Berlin, Boston.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Corpus Linguistics 2011*.

Stefan Hagel. 2007. The Classical Text Editor. An attempt to provide for both printed and digital editions. In A. Ciula and F. Stella, editors, *Digital Philology and Medieval Texts*, pages 77–84. Pacini Editore, Pisa.

Dag Trygve Truslew Haug and Marius L. Jøhndal. 2008. Creating a parallel treebank of the old indo-european bibletranslations.

M. Janssen, P. Lendvai, and A. Jouravel. 2025. Alignment of historical manuscript transcriptions and translations. In *Proc. of RANLP 2025, September 2025, Varna, Bulgaria*.

Maarten Janssen. 2016. TEITOK: Text-faithful annotated corpora. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 4037–4043.

Anna Jouravel and Janina Sieber. 2024. The Greek and Slavonic Transmission of Methodius' De lepra. In Katharina Bracht, Anna Jouravel, and Janina Sieber, editors, *Methodius of Olympus: De lepra*, pages 11–30. De Gruyter, Berlin, Boston.

Anna Jouravel, Janina Sieber, and Katharina Bracht, editors. 2024. *Methodius von Olympus: De lepra*. De Gruyter, Berlin, Boston.

José Maksimczuk. 2024. The florilegium coislinianum and the greek text of methodius' de lepra. In Katharina Bracht, Anna Jouravel, and Janina Sieber, editors, *Methodius of Olympus: De lepra*, pages 31–54. De Gruyter, Berlin, Boston.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jacques Paul Migne. 1857. *Tou En Agiois Patros Emon Methodiou, Episkopou Kai Martyros, Ta Euriskomena Panta*, volume 18 of *Patrologiae Cursus Completus. Series Graeca. Patrologiæ Græcæ*. J.-P. Migne, Parisiis.

Roger Pearse, Ralph Cleminson, and Andrew Eastbourne. 2015. *Methodius of Olympus, De Lepra (On Leprosy).*

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. *Parallel corpora for medium density languages*, pages 247–258. John Benjamins Publishing Company.