Creating corpora with semi-automatic middle Chinese phonetic glosses

Alexander Delaporte, Aliou Badara Diagne

▶ **To cite this version:**

Alexander Delaporte, Aliou Badara Diagne. Creating corpora with semi-automatic middle Chinese phonetic glosses. The 14th International Conference on Digital Archives and Digital Humanities, ; ; ; ; ; ; , Dec 2023, Tainan, Taiwan. hal-04337445

HAL Id: hal-04337445
https://hal.science/hal-04337445

Submitted on 21 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Creating corpora with semi-automatic Middle Chinese phonetic glosses

Extended abstract

劉詩豪 Delaporte Alexander[1] and Diagne Aliou Badara[2]

[1]Centre for Linguistic Research on East Asia (Centre de Recherches Linguistiques sur l'Asie Orientale, CRLAO)
[2]Centre d'études supérieures de la Renaissance (CESR), Université de Tours
alexander.delaporte@cnrs.fr, *beugdabakh92@gmail.com*

November 29, 2023

We describe a semi-automatic method for adding Middle Chinese phonetic glosses to a set of Archaic Chinese texts, in order to create a fully annotated corpus. Our aim is to make sound glosses available for non-specialists in Chinese historical phonology which would thus facilitate study by linguists and philologists.

1 Context

The project takes its origins in our colleague Jacques Guillaume's works, notably his paper titled *On the nature of morphological alternations in Archaic Chinese and their relevance to morphosyntax* (Jacques 2022). Jacques states that the opacity to non-specialists of *Jingdian shiwen* 經典釋文, a widely studied collection of sound glosses authored by the Tang dynasty scholar Lu Deming 陸德明, may lead to the neglect of relevant phonological features in linguistic analysis. Increasing accessibility to these glosses would then benefit researchers from various fields, and may also provide a first step to an Archaic Chinese reconstruction.

A way of achieving such accessibility would be to produce corpora including full annotation of all characters. However, the process of building a fully transcribed corpus cannot be completely automated: "the alignment of the glosses to the text is not trivial, notably because some sound glosses strand over several sentences [...], and also because readings judged 'obvious' by the compiler of the *Shiwen* have not been systematically indicated" (ibid.).

2 Annotation methodology

Given this need to allow human intervention while processing a whole set of texts into structured documents, we adopted a semi-automatic approach. A first step of automatic preannotation determines whether a character has one or multiple possible readings, and provides the list of all possible readings for each character. Then, in the case of multiple-reading characters, human annotators select the right reading in its context.

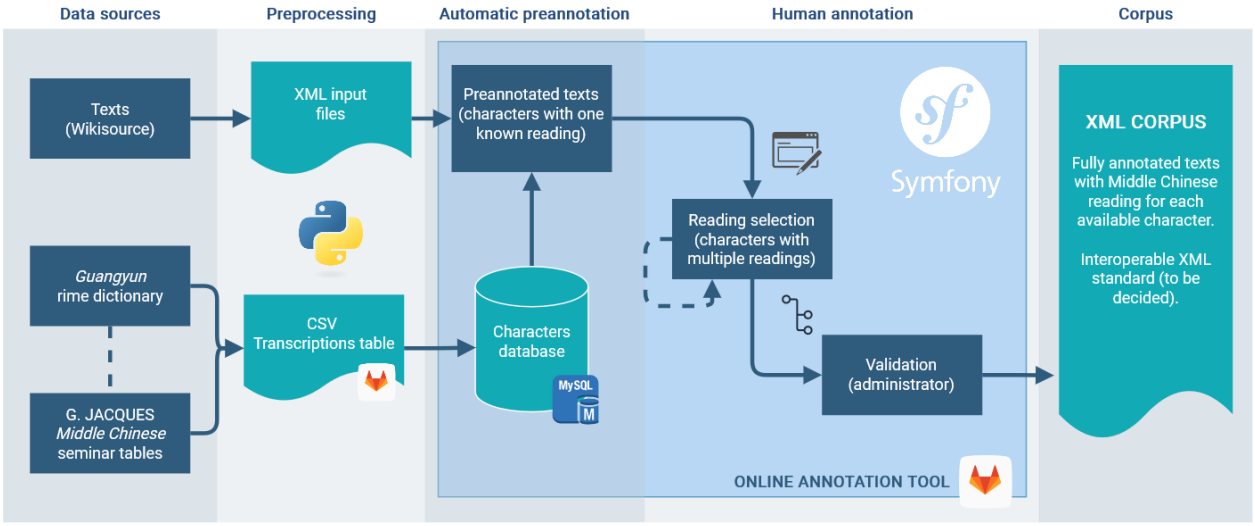

Figure 1: Processing chain overview

In practical terms, the first stages of data processing are handled by a set of Python scripts, leading to input files in XML and CSV format. Further stages are handled by an *ad hoc* online tool, developed in PHP Symfony 5. The processing chain as a whole is shown on Figure 1.

2.1 Data sources

We used data for the following two purposes:

1. Building a preliminary XML corpus, for which we extracted texts from Wikisource.
2. Creating a characters database, associating characters and their possible readings. The set of characters and their properties were taken from *Guangyun* 广韵 rime dictionary. The transcriptions are following the principles described in Baxter (1992): in order to produce Middle Chinese transcription(s) for each character, we used tables extracted from Jacques' doctoral seminar in Middle Chinese, in which these principles are put into practice (Jacques 2021).

As previously stated, this data was processed using several Python scripts, ultimately producing a repository containing 20 XML files and a CSV table providing transcriptions for approximately 26000 characters.

2.2 Automatic preannotation

Preannotation and annotation stages will all occur in an online platform, currently under development using the framework Symfony 5. The XML input files are included as is, while the CSV transcription table has been converted into an SQL database for more efficient data handling.

The preannotation stage consists in producing and displaying a dedicated page for each text, as follows:

- The page shows text content taken from the corresponding XML input file. A differienciate formatting allows viewers to quickly identify the different components: text, commentary and glosses.

- For the parts that need transcription, each character's possible transcriptions are fetched from the database. When only one reading has been fetched, it is displayed. If several transcriptions are found, they are displayed in a select list from which the human annotator will have to determine which is the most accurate.

This preparatory stage limits the actual annotation task to selecting the appropriate item in a list of suggestions.

2.3 Human annotation

For this project we intend to set up annotation as an iterative, collaborative process.

- As an authenticated user, an annotator should be able to fill in the form in full or in part, which means it is not mandatory to annotate a whole text at once.
- Submitted annotations will be reviewed and validated (or discarded) by a small preselected group of expert users/annotators.
- Modifications on validated annotations will be disabled in the form. Completed and validated texts will be taken out of the annotation platform and made available for download as XML files.

It should be noted that, as this process will be based on a version control system [1], the intermediate states of the corpus may also be downloadable.

3 Current state of progress and further developments

This project is still undergoing development and has not yet produced publicly available deliverables. Input files have been created and most of the preannotation process is ready to use. Our work now focuses on the implementation of annotation forms and a version control system. Git seems to be the best candidate, as long as there are no technical limitations to its use. This remains to be verified.

Additional features are also needed, such as online database consultation (already implemented) or research of awaiting annotations based on a given character (under construction).

When completed, the source code for the online platform will be shared, along with the characters database [2].

4 Contributions to digital sinology and digital humanities

The production of corpora fully annotated with sound glosses will provide researchers with a valuable resource, allowing non-specialists to take into account otherwise inaccessible phonological parameters in their study of Archaic Chinese texts. Other deliverables will also benefit digital sinology and digital humanities. For instance, the database providing characters transcriptions in Middle Chinese is as such a highly reusable data source. The online platform's source code can also be used as a starting point for similar projects, for any language requiring a unit-by-unit transcription.

[1] To be decided, cf. section 3.

[2] The files are currently hosted on Huma-Num's instance of Gitlab, on a private visibility level. They will subsequently be made public.

References

Baxter, William H. (1992). *A Handbook of Old Chinese Phonology*. Berlin, New York: De Gruyter Mouton. ISBN: 9783110857085. DOI: doi:10.1515/9783110857085. URL: https://doi.org/10.1515/9783110857085.

Jacques, Guillaume (2021). "Middle Chinese." In: *Historical Phonology Doctoral School, Gent*.

— (2022). "On the nature of morphological alternations in Archaic Chinese and their relevance to morphosyntax." In: *Bulletin of the School of Oriental and African Studies* 85.3, pp. 475–494. DOI: 10.1017/S0041977X22000854.