# Creating corpora with semi-automatic middle Chinese phonetic glosses

## 製作半自動中古音韻標註的語料集

劉詩豪 DELAPORTE Alexander [1]  DIAGNE Aliou Badara [1 2]

[1] 國家科學研究中心 8563 綜合研究單位 CNRS-UMR 8563 | 東方語言研究中心 CRLAO

[2] 圖爾大學 Université de Tours | 文藝復興研究中心 CESR

December 2nd, 2023

# Outline

# Outline

# Context and objectives

## Research context

### Project background

This project is a collaboration with our colleague, Research Director JACQUES Guillaume 向柏霖 who is fully in charge of the theoretical aspects of the project.

# Context and objectives

## Research context

- 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.

(Jacques 2022)

# Context and objectives

## Objectives

### Main goal

Our main goal is to produce fully annotated texts, in which each character is annotated with its Middle Chinese transcription.

# Context and objectives

## Objectives

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” (Jacques 2022)

# Context and objectives

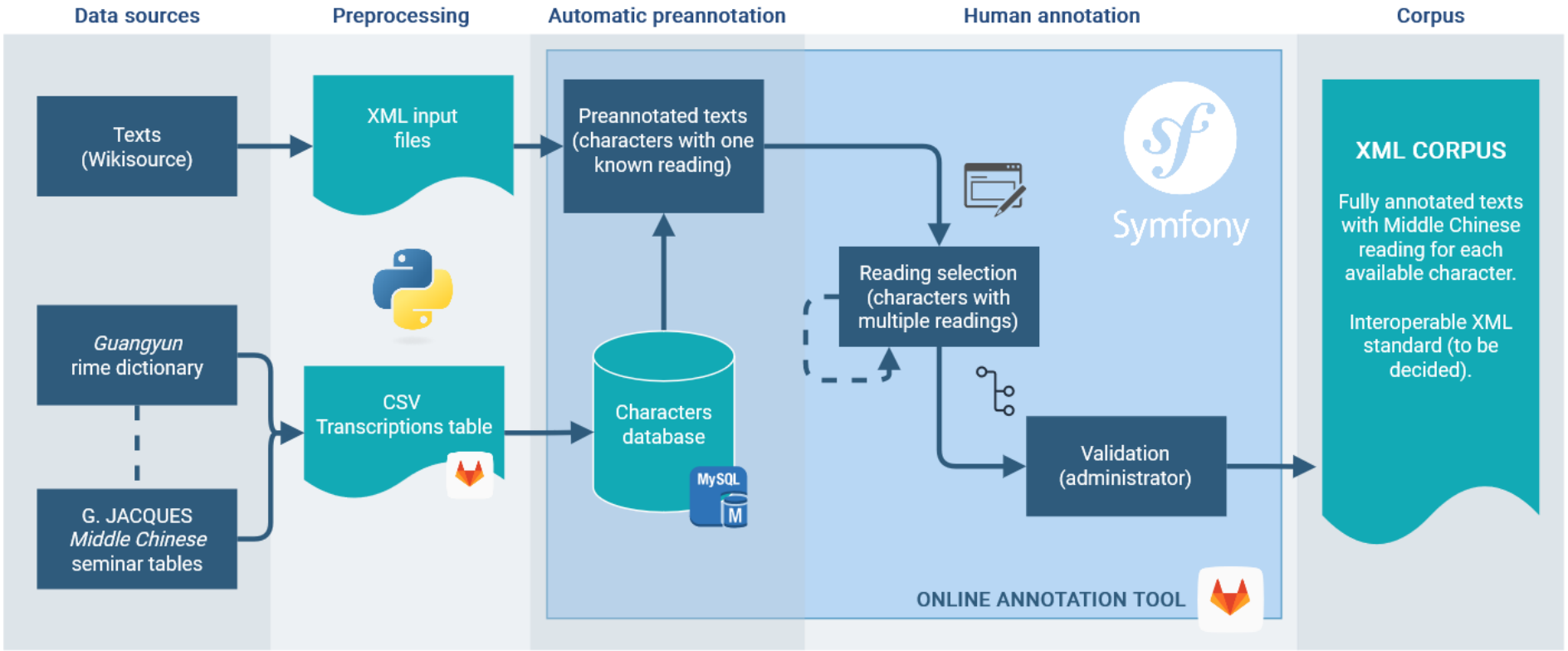
## Proposed deliverables

The following deliverables are planned:

1. At least one XML corpus of fully annotated texts.
   - "Fully annotated"= A Middle Chinese transcription is provided for each character.
2. A character database containing a large set of characters and their possible reading(s).
   - Current coverage : approximately 20 000 characters
3. An online semi-automatic collaborative annotation tool.
   - The source code will be distributed under an open license.

# Outline

# Processing chain overview



# Outline

# Automatic preannotation

## Producing input files

A set of Python scripts has been used to produce the following input files:

1. XML text+metadata files
2. CSV characters and transcriptions table
   - Transcriptions = Baxter-Sagart system [1]

All subsequent processing stages are/will be managed via an online platform developed in PHP Symfony 5.

[1] As described in Baxter (1992), reconstructed using notably Jacques (2021).

# Automatic preannotation

## Presenting the user/annotator with preannotated texts

Using data from input files:

- Texts are displayed by reading data and metadata directly from XML input files.
- Transcriptions are fetched from a character database.
  - Database built using data from the CSV file containing characters and their respective possible reading(s).

# Automatic preannotation

## Presenting the user/annotator with preannotated texts

### Annotation task set-up

For each character:

- If the character has only one known transcription = the transcription is displayed, no additional task needed.
- If the characters has multiple known transcriptions =
  - Its transcriptions are displayed as a select list.
  - User/annotator input is then needed to select the appropriate transcription.

# Outline

# Human annotation and review

**An iterative and incremental process**

- The user/annotator selects the suitable transcriptions for one, several, or all characters within a given text.
  - This task can be repeated.
  - Multiple annotators can contribute.
- Changes are registered, then reviewed by an expert user or administrator.
  - A version control system will be implemented (currently under development).
- Validated transcriptions are taken out from the annotation process.

# Human annotation and review

## A simple task for the user/annotator

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

# Outline

# State of progress

**Main features and deliverables**

- Completed or very advanced development:
  - Character database
  - Online platform architecture
  - Preannotation process
  - Data and users administration dashboard
- Under development:
  - Annotation forms
  - Version control system (Git via Huma-Num's Gitlab instance)

# State of progress

**Additional features**

- Completed:
  - Character database consultation page
- Under development:
  - Research of awaiting annotations based on a given character

# Outline

# Contributions to Digital Humanities

## Main contribution

Fully annotated corpora with sound glosses will provide researchers a valuable resource, enabling exploration of otherwise inaccessible phonological parameters in the study of Archaic Chinese texts by non-specialists.

All proposed deliverables can benefit Digital Sinology and Digital Humanities:

- Annotated corpora (as stated before).
- Database providing character transcriptions in Middle Chinese.
- Source code of the online platform.
  - Can serve as a starting point for similar projects involving any language that requires unit-by-unit transcription.

# 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.

# Contact

劉詩豪 **DELAPORTE Alexander** alexander.delaporte@cnrs.fr
**DIAGNE Aliou Badara** beugdabakh92@gmail.com

**Source code repository**
https://gitlab.huma-num.fr/projet-annotation-chinois-moyen
(files will be made public at v1.0 release)