

Using Cross-Linguistic Data Formats to Enhance the Annotation of Ancient Chinese Documents Written on Bamboo Slips

Michele Pulini^{1,2} and Johann-Mattis List^{2,3}

¹ Department of Asian and North African Studies, Ca' Foscari University, Venice, Italy

² Chair for Multilingual Computational Linguistics, University of Passau, Passau, Germany

³ DLCE, MPI-EVA, Leipzig, Germany

Abstract

Ancient Chinese documents written on bamboo slips more than 2000 years ago offer a rich resource for research in linguistics, paleography, and historiography. However, since most documents are only available in the form of scans, additional steps of analysis are needed to turn them into interactive digital editions, amenable both for manual and computational exploration. Here, we present a first attempt to establish a workflow for the annotation of ancient bamboo slips. Based on a recently rediscovered dialogue on warfare, we illustrate how a digital edition amenable for manual and computational exploration can be created by integrating standards originally designed for cross-linguistic data collections.

1 Introduction

Most computational approaches on ancient stages of the Chinese language restrict themselves to the classical canon of writings that have been handed down for several thousand years now. These sources are written in a standardized character system called the *regular script* (楷書 *kǎishū*). From the time when it was introduced more than 2000 years ago until now, this system does not seem to have been modified in any significant form. As a result, one often thinks of Chinese writing as a unified endeavor, unlikely to change, and unlikely to have changed radically throughout most of its history. When going back deeper in time, however, this picture changes drastically. Thanks to an increasing amount of documents written on bamboo slips that have been archeologically excavated in recent times, scholars are now learning more and more about the immense degree of variation in writing that was characteristic for China long before the *regular script* became adopted as the standard. In order to understand the history of the Chinese language, it is indispensable to pay attention to its variety in writing reflected in these sources.

Most bamboo manuscripts date back to the period from the mid-late Warring States Period (*Zhànguó* 戰國, late fourth century to early third century BCE) up to the Hàn dynasty (*Hàn cháo* 漢朝, 206 BCE to 220 CE). They represent a true wealth of new data and evidence for linguistics, paleography, philology, and historiography. However, putting this treasure of knowledge to use in research bears two major challenges that have not been sufficiently addressed so far. A first challenge consists in the *analysis* of the characters observed in bamboo slips. While these follow the general building structure of Chinese characters, allowing us to identify phonetic and semantic components that usually also find their counterparts in the *regular script*, ancient writing shows a much greater variation with respect to the combinations that are possible here (see Figure 1 as an example). As a result, it is often impossible to find the modern counterparts – along with their Unicode values – of characters observed on bamboo slips. A second challenge consists in the digital *curation* of modern editions of excavated texts. Given that scholars often lack digital training, the vast majority of editions is restricted to scans accompanied by comments, while a deeper integration of data is lacking.

Taking one particular text – available in two different bamboo manuscript versions – as example, we illustrate how these problems could be addressed in the future. Using basic concepts from corpus linguistics, Natural Language Processing, and computer science, we show how the original documents can be annotated, how individual characters can be analyzed with respect to their composition, and how an entire digital edition can be constructed, providing scholars interested in manual data exploration with interactive access to the original documents, while the data in standardized, machine-readable form to scholars interested in computational analysis.







Original						
Kǎishū	𩇛	𩇛	既	氣	氣	氣
Source	GD *Lǎozǐ A 35	BS *Būshì jidào 223	SB *fán wù liú xíng B 3	QH *Yuègōng qí shì 20	QH *Nǎi míng yī 9	SB Cáo Mò zhī zhèn 46b

Figure 1: Illustration of variation in bamboo script. The table lists variants of the character *qì* 氣 “breath, energy”, as it can be found across manuscripts, along with projected conversions of the character structure to the modern regular script.

2 Materials

The text we use to illustrate our workflow is called *Cáo Mò zhī zhèn* 曹沫之陳 (*Cáo Mò’s Battle Formations*). It is a long-lost Chinese philosophical dialogue on ethics and warfare between Duke Zhuāng of Lǔ (Lǔ Zhuāng Gōng 魯莊公, reign 693–662 BCE) and his general Cáo Mò, which has resurfaced after millennia in the form of two manuscript copies on bamboo. The Shanghai Museum manuscript (henceforth SB) was first published in print in 2004 in the fourth volume of the collection (Mǎ, 2001–2012), while the Anhui University copy (henceforth AD) was first published in print in 2022 in the second volume of the collection, whose publication is still ongoing (Huáng and Xú, 2019–2022). Though coming from illegal excavations, their authenticity has been proven through radiocarbon dating of the bamboo slips, which has confirmed a mid-third century BCE dating for SB and a late fourth century BCE dating for AD. SB counts 65 bamboo slips (45 intact and 20 broken, average length 47.4 cm, average width 0.6 cm). The manuscript contains 1778 characters written in the Chǔ orthography and carries the title of the text on the back of slip no.2. On the other hand, AD counts 46 slips (slips 4 and 5 are entirely missing, average length 48.2 cm, average width 0.6 cm). The manuscript contains a total of 1623 character entries. Unlike SB, AD lacks a clear title indication on its back, but some inscriptions as well as some oblique lines useful for correct ordering of the slips are present on the verso. Both *folia* were originally held together by three binding cords (*biānshéng* 編繩), secured to the bamboo slips through binding notches (*qikǒu* 契口). The binding cords, however, have been irreparably lost. A solid study on some codicological aspects, such as scribal hands, manuscripts production, and use of punctuation as been recently pub-

lished by Zheng (2024) to whom we redirect the reader. The high-quality scans of both AD and SB bamboo slips provided in the the second volume of the Anhui University manuscripts collection has served as starting point and preliminary material for the digital edition.

3 Methods

Our workflow for the digital annotation, curation, and publication of bamboo script documents consists of four stages. In the first stage, we carry out a detailed *digital annotation* of the original data. In the second stage, we conduct an extensive *analysis* of the texts by analyzing characters, identifying words, and glossing sentences semantically. In the third stage, we model the data according to the formats proposed by the Cross-Linguistic Data Formats initiative (CLDF, <https://cldf.cldf.org>, Forkel et al. 2018). In the fourth stage, finally, we *deploy* the data in the form of an interactive CLLD application (<https://pypi.org/project/cldd>, Forkel 2014).

3.1 Digital Annotation

In order to allow for a flexible reuse of the analyzed data, our workflow starts from the digital annotation of the original documents. The core of this annotation consists in marking individual characters with boundary boxes and annotating the boxes in such a way that the characters can be identified at later stages. For this task, we used Recogito (<https://recogito.pelagios.org/>, Barker et al. 2019), an interactive tool for semantic annotation that greatly facilitates this task. Other tools could have been used for this step as well, but we selected Recogito for its shallow learning curve and its general openness. Once finished, Recogito’s boundary boxes can be easily exported to various file formats and accessed from computer programs, al-

lowing us to cut individual characters out of the original scans in order check how much they vary.

3.2 Linguistic Analysis

The linguistic analysis consists in two parts that may go hand in hand, namely *character analysis* and *text analysis*. During character analysis, all characters annotated in the previous step of our workflow must be analyzed carefully, identifying their external and internal structure, – where possible – their modern counterparts, as well as establishing their pronunciation through different stages of Chinese (regarding the distinction between external and internal structure, see List et al. 2016, 50). Since the identification of words and the assignment of readings to the characters attested on the bamboo slips constitute a comprehensive philological enterprise open to criticism and debate, each character was first reproduced as faithfully as possible using *ideographic description sequences* (for details, see Kordek 2013, 62) and then – where possible – assigned a counterpart in standard Chinese characters. The assignment of the modern counterparts was grounded in the extant scholarly literature on the *Cáo Mò zhī zhèn* manuscripts. Aside from the original critical editions of 2004 and 2022 (Mǎ, 2001–2012; Huáng and Xú, 2019–2022), the editions by Yú and Zhāng (2019) and Sūn (2023) proved essential. In addition, it turned out that a basic understanding of word families and language-internal cognates in Old Chinese phonology (compare Pulini and List, 2024) can be of great help during the analysis. This two-level analysis of character structures – one stricter and one broader – ensures transparency and prevents the loss of information during the annotation process. In identifying ancient readings for the characters, we followed the tradition of research on Old Chinese phonology in providing Middle Chinese readings (a language variety documented in rhyme books published around the 6th century CE) in the system of Baxter (1992), and Old Chinese reconstructions, following the system proposed by Baxter and Sagart (2014).

During text analysis, both manuscripts are unified to form one coherent text. Given the status of the sources, the preference is here given to the AD version of *Cáo Mò's Battle Formations*, with apparent gaps being filled in from the SB version. The analysis starts from the identification of words (which may consist of two and more characters at times) and phrases. Once identified, these are

glossed semantically, using basic techniques for *interlinear morphemic glossing* (Lehmann, 2004). In this stage, we also handle characters that may be missing entirely from both versions. These are not only marked specifically but can also be identified easily, as they have no image data attached to them. At the end, translations for phrases and entire passages in English are suggested.

Both character analysis and text analysis are carried out in tabular form, using common spreadsheet editors. In character analysis, each row in a table corresponds to a character that is itself linked to the original scan via the boundary boxes that were added during digital annotation, with separate information being placed into different columns of the table. In text analysis, the basic unit assigned to each row is the word, consisting of one or more characters. Both tasks require a detailed knowledge of bamboo slips, Chinese paleography, and Old Chinese etymology, and were therefore exclusively carried out by the first author of this study, while the role of the second author consisted in the design of formal tests of annotation consistency. While this means that tests on inter-annotator agreement (McDonald et al., 2019) are lacking, we hope to improve the data in the future through comments by our colleagues.

3.3 CLDF Integration

The CLDF specification has been developed for various forms of cross-linguistic data, including wordlists, dictionaries, and feature collections. More recently, CLDF has been extended to integrate corpus data, offering additional functionality to handle interlinear-glossed texts (List et al., 2021). While CLDF has not been specifically designed to handle Chinese text collections, the format offers many advantages about alternative data formats. First, data provided in CLDF can be easily queried computationally, using the dedicated PyCLDF Python package (Forkel et al., 2025), as well as SQLite (<https://sqlite.org>, see Shcherbakova and List 2023 for an example querying lexical data in CLDF, and Blum et al. 2024 for an example query in corpus data). Second, data available in CLDF can be easily imported into CLLD applications (<https://cldd.org>, Forkel, 2014), thus offering facilitated ways to deploy data collections to the web where they can be conveniently inspected by interested users (see § 3.4 for details).

With our main data available in tabular form, it is straightforward to convert the data to CLDF,

Original						
Analysis	魯	莊公	將	為	大	鍾
Character	魯	莊公	將	為	大	鍾
Pinyin	lǔ	zhuāng gōng	jiāng	wèi	dà	zhōng
Middle Chinese	luX	tsrjang-kuwng	tsjang	hjwe	dajH	tsyowng
Old Chinese	r.ŋʰaʔ	tsraŋ-C.qʰoŋ	tsaŋ	ɣʷ(r)aj	lʰat-s	toŋ
Gloss	Lu.(place.name)	duke.Zhuang.(personal.name)	aspect.marker	do	big	bell

Duke Zhuang of Lu was in the process of having a massive bell cast.

Figure 2: Interlinear-glossing example from the web application of the digital edition.

since the CLDF format specification itself is mostly based on tabular data. In CLDF, our data is modeled as a generic dataset, consisting of a language table (linking only to one language, Old Chinese), an entry table that stores the words in the text, and an example table consisting of the individual phrases in interlinear-glossed form (following the *Leipzig Glossing Rules*, see Comrie et al. 2015), where Middle Chinese and Old Chinese reading are offered as additional glossing layers. An additional table is used to store individual characters linked to their location in the scans of the original two editions. The CLDF conversion was carried out with the help of CLDF-Bench (<https://pypi.org/project/cldfbench>, Forkel and List 2020), a tool that facilitates the conversion to Cross-Linguistic Data Formats. Additional data handling was conducted with the help of the SinoPy package (<https://pypi.org/project/sinopy>, List 2019).

3.4 Deployment with CLLD

The cllD toolkit (Forkel, 2014) is a Python library that facilitates the deployment of data provided in the form of Cross-Linguistic Data Formats by providing researchers with an interactive web framework that can be interactively explored. The cllD application that we created on top of the CLDF data provides the digital edition of the *Cáo Mò zhī zhèn* in the form of an integrated web application in which original characters can be explored in a unique way that integrates the graphemic, phonetic, and semantic analysis underlying the edition.

4 Examples

With the data assembled both in the form of CLDF and an interactive web application, our digital edition of the *Cáo Mò zhī zhèn* allows for both computational and manual exploration. Since we provide the data in the form of interlinear-glossed text, standard approaches from corpus linguistics can be used to query the data in order to investigate the text. As a first example, Table 1 shows the ten most frequently recurring words in the source. As can be seen, we do not only find grammatical markers in this list, but also words like *yuē* 曰 “say” that point to the fact that the text is written as a dialogue, and *zhàn* 戰 “war”, pointing to the major topic of the text. While the source itself is limited, this example shows the potential for extended computational analysis once more annotated texts become available.

C	G	OCH	OCC
之	conjunction	tə	76
曰	say	ɣʷat	57
不	negative adv.	pə	56
有	have	ɣʷəʔ	52
以	preposition	ləʔ	40
而	conjunction	nə	30
於	preposition	ʔʰa	29
其	poss. pron.	gə	27
戰	war	tar-s	27
莊公	duke Zhuang	tsraŋ C.qʰoŋ	22

Table 1: Most frequent words and particles in the *Cáo Mò zhī zhèn*. C refers to the character (in standardized form), G refers to the gloss (abbreviated for reasons of visibility), OCH refers to Old Chinese readings, and OCC provides information on occurrences in the text.





ID	Word Form	Middle Chinese	Old Chinese	Original Characters	Phrase ID
<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>
word-4-1	為	hjwe	ɣʷ(r)aj		1
word-4-10	為	hjwe	ɣʷ(r)aj		151
word-4-11	為	hjwe	ɣʷ(r)aj		165
word-4-12	為	hjwe	ɣʷ(r)aj		174

Figure 3: Comparing individual occurrences of words in the digital edition. The table provides the first 6 out of a total of 16 entries in the digital edition for 為 *wéi* “do, act” MCH *hjwe*, allowing scholars to inspect differences in the writing of the word inside and across editions.

Figure 2 gives a direct example of one interlinear-glossed phrase in the web application, showcasing how the original images are integrated with the two-level analysis of the Chinese characters, the readings in Middle Chinese and Old Chinese, and the semantic glosses. The example also shows how we handle those cases in which the *internal character structure* that we observe in bamboo slips cannot be matched with a counterpart in modern *kāishū* writing. Since the components of Chinese characters are usually limited and can be easily detected, we use *ideographic description sequences* – a system that allows to systematically analyze characters into their components (Skala, 2015) – to display the way in which the individual components are arranged. Thus, the sequence 戕口, that we identify with the personal name *zhuāng* 莊, refers to a character consisting of *qiāng* 戕 “to kill” with the character radical *kǒu* 口 “mouth” inside.

An additional web view is shown in Figure 3, allowing users to compare individual variants of the same word or character, thus offering direct and convenient ways to assess the variation in writing. In addition, the example not only illustrates how the image annotations are directly integrated with the web application, but also shows how the different backgrounds going back to the original publications of the scans, allow us quickly to see to which of the two sources (AD and SB) the text parts belong, with AD being based on colored images, while SB is given in gray scale.

All in all, we hope that these examples illustrate the advantage of integrated web applications

in which data are stored separately, providing access to computers and humans at the same time. Complex web applications often deliberately keep users proficient in computing from accessing the resources with the help of software tools. Resources that focus on providing data exclusively for computational analyses, on the other hand, often underestimate the important role that direct inspection can play in spotting potential errors in annotations.

5 Conclusion and Outlook

With this study, we have presented a digital edition of an ancient text on warfare in China that can be accessed both computationally and manually. While computational approaches to bamboo script are still meeting a large number of obstacles, we think that our example application could offer a solution for future studies, by increasing the amount of digitally annotated data that could be used in several kinds of computational studies. Thus, using the annotated images with the associated character structures, one could develop initial models to test the limits of current tools for hand-written text recognition (Kahle et al., 2017). Using tools from corpus linguistics, one could analyze the text in various ways (Hunston, 2022), using it also as the basis for the creation of corpus-based dictionaries (Bowker, 2010). Finally, by providing textual data along with reconstructed pronunciations, such data collections as the one shown here can help to improve and consolidate our knowledge about the ancient pronunciation of Chinese throughout different times and places.

Supplementary Material

All data and code necessary to replicate the study presented here are freely available for download. The code for the curation of the data is hosted with Codeberg (<https://codeberg.org/cldf-datasets/caomozhizhen>) and archived with Zenodo (<https://doi.org/10.5281/zenodo.15039078>, Version 1.0). The application is available online (<https://cmzz.digling.org>), and the code underlying the application is also hosted with Codeberg (<https://codeberg.org/digling/cmzz>).

Acknowledgements

This research was supported by the ERC Consolidator Grant *ProduSemy* (JML, Grant No. 101044282, see <https://doi.org/10.3030/101044282>). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them. We thank Robert Forkel for providing important help tips in the preparation of the data in the form of a CLDF package and the deployment of the data as CLLD application.

Limitations

The workflow we propose is still preliminary. While the extension of *Cáo Mò zhī zhèn* allows for manual transcription and annotation of the data, additional challenges may arise when dealing with longer texts. Moreover, methods for further integrating the critical apparatus into the edition will need to be developed in the future. There are also some concrete points that we hope to improve in the current annotation in the nearer future. Among these are a much more detailed check of glosses and translations, as well as a direct reference to the sources that were employed in the analysis of each individual character and word.

However, the edition we propose serves as a proof of concept and a model for the further development of digital critical editions of early Chinese manuscripts. We thus hope that despite its preliminary status, our work may prove useful for colleagues working on similar research questions and open a broader discussion on the need for consistent annotation and digitalization workflows for data on ancient historical languages.

References

- Elton Barker, Leif Isaksen, Rebecca Kahn, and Rainer Simon Valeria Vitale. 2019. *Recogito [Software Library, Version 3.3]*. Pelagios Network, Vienna.
- William H. Baxter. 1992. *A Handbook of Old Chinese Phonology*. Trends in Linguistics: Studies and Monographs. Mouton de Gruyter, Berlin.
- William H. Baxter and Laurent Sagart. 2014. *Old Chinese: A New Reconstruction*. Oxford University Press, Oxford.
- Frederic Blum, Ludger Paschen, Robert Forkel, Susanne Fuchs, and Frank Seifart. 2024. **Consonant lengthening marks the beginning of words across a diverse sample of languages**. *Nature Human Behaviour*, 8(11):2127–2138.
- Lynne Bowker. 2010. **The contribution of corpus linguistics to the development of specialised dictionaries for learners**. In Pedro A. Fuertes-Olivera, editor, *Specialised Dictionaries for Learners*, page 155–170. De Gruyter.
- Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2015. *Leipzig Glossing Rules. Conventions for interlinear morpheme-by-morpheme glosses*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Robert Forkel. 2014. **The Cross-Linguistic Linked Data project**. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, pages 61–66, Reykjavik. ELRA.
- Robert Forkel and Johann-Mattis List. 2020. **CLDF-Bench. Give your Cross-Linguistic data a lift**. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, pages 6997–7004, Luxembourg. European Language Resources Association (ELRA).
- Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymiski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. **Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics**. *Scientific Data*, 5(180205):1–10.
- Robert Forkel, Christoph Rzymiski, and Sebastian Bank. 2025. *PyCLDF [Software, Version 1.40.4]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Susan Hunston. 2022. *Corpora in applied linguistics*, 2 edition. Cambridge University Press, Cambridge.
- Dékuān Huáng and Zàiguó Xú, editors. 2019–2022. *Ānhuī Dàxué Cáng Zhànguó Zhújiǎn* 安徽大學藏戰國竹簡. Zhōngxī Shūjú 中西書局, Shanghai. 2 volumes.

- Philip Kahle, Sebastian Colutto, Gunter Hackl, and Gunter Muhlberger. 2017. [Transkribus – A service platform for transcription, recognition and retrieval of historical documents](#). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE.
- Norbert Kordek. 2013. *On Some Quantitative Aspects of the Componential Structure of Chinese Characters*. Wydawnictwo Rys, Poznań.
- Christian Lehmann. 2004. [Interlinear morphemic glossing](#). In Geert E. Booij, Christian Lehmann, Joachim Mugdan, and Stavros Skopeteas, editors, *Morphology. An international handbook*, volume 2, pages 1834–1857. De Gruyter, Berlin and New York.
- Johann-Mattis List. 2019. *SinoPy: A Python library for quantitative tasks in Chinese historical linguistics [Software Library, Version 0.3.4]*. Max Planck Institute for the Science of Human History, Jena.
- Johann-Mattis List, Nathaniel A. Sims, and Robert Forkel. 2021. [Towards a sustainable handling of interlinear-glossed text in language documentation](#). *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(2):1–15.
- Johann-Mattis List, Anselm Terhalle, and Daniel Schulzek. 2016. Traces of embodiment in Chinese character formation. A frame approach to the interaction of writing, speaking, and meaning. In Liane Ströbel, editor, *Sensory-motor concepts. At the crossroad between language & cognition*, pages 45–62. Düsseldorf University Press.
- Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. [Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice](#). *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23.
- Chéngyuán Mǎ, editor. 2001–2012. *Shànghǎi Bówùguǎn Cáng Zhànguó Chǔ Zhúshū* 上海博物館藏戰國楚竹書. Shànghǎi gǔjí chūbǎnshè 上海古籍出版社, Shanghai. 9 volumes.
- Michele Pulini and Johann-Mattis List. 2024. [Finding language-internal cognates in Old Chinese](#). *Bulletin of Chinese Linguistics*, 17(1):53–72.
- Olena Shcherbakova and Johann-Mattis List. 2023. [Retrieving and analyzing taste colexifications from Lexibank](#). *Computer-Assisted Language Comparison in Practice*, 6(2):73–86.
- Matthew Skala. 2015. [A structural query system for Han characters](#). *International Journal of Asian Language Processing*, 23(2):127–159.
- Yǒngbō Sūn. 2023. *Ānhuī Dàxué Cáng Zhànguó Zhūjiǎn (Èr) Jí Shì*, volume 38 of *Gǔdiǎn Wénxiàn Yánjiū Jíkān*. Huāmùlán wénhuà shìyè yǒuxiàn gōngsī, New Taipei. 安徽大學藏戰國竹簡(二)集釋.
- Shàohóng Yú and Qīngsōng Zhāng, editors. 2019. *Shànghǎi Bówùguǎn Cáng Zhànguó Chǔ Jiǎn Jí Shì* 上海博物館藏戰國楚簡集釋. Social Sciences Academic Press, Beijing. 10 volumes.
- Oscar Qiu Jun Zheng. 2024. Scribal hands, errors, and intervention – manuscript production approaches of the cao mie zhi zhen 曹蔑之陣 manuscripts. *Bamboo and Silk*, 7(2):155–201.