Chances and Challenges for Quantitative Approaches in Chinese Historical Phonology

Johann-Mattis List

Abstract

The field of Chinese Historical Phonology is traditionally dealing with a large number of complex and diverse types of data. While the data diversity can be conveniently dealt with in qualitative approaches, computational possibilities that have arisen during the past two decades offer new possibilities and new challenges for the field. In the study, I will summarize the chances and challenges which we face in the discipline and point to some suggestions for future work. While not being able to provide a direct solution for most issues of data handling and standardization, I hope that this study can contribute to a broader discussion about data and standards in the field of Chinese Historical Phonology.

1. Introduction

It is well known to all practitioners of Chinese Historical Phonology that the evidence scholars have to deal with is complex and diverse. The situation is nicely reflected in Li (1971, later translated into English in 1974), who mentions rhyme series, xiéshēng series, and the study of rhyme books as the three major types of evidence which were already traditionally employed by Qing scholars in the study of Chinese language history. These were later complemented by several “new lines of investigation” (ibid.), including loan word researches, researches in comparing Chinese with other Sino-Tibetan languages, variant forms and loan characters in the classical texts and Han time sound glosses (ibid.). All these lines of investigation, mentioned by Li, still constitute the fundament upon which the reconstruction of Old Chinese pronunciation rests, as evidenced clearly by the reconstruction systems published in the past decades (Baxter and Sagart 2014; Zhengzhang 2000; Starostin 1989; Karlgren 1957; Wáng 1980; Pān 2000; Schuessler 2007).

A crucial part of research in the field of Chinese Historical Phonology rests upon
the organization and interpretation of data. Scholars have to sift through ancient texts, understand their organization, memorize their content, identify characters as rhyming with each other, or analyze the structure of individual Chinese characters, which have undergone several changes during the history of the Chinese writing system. As such, Chinese Historical Phonology has always been a very tedious subject of study, given that it demands a huge amount of diligence paired with creativity and wit. With increasing amounts of data which have to be inspected, compared, weighted and finally puzzled together to form a coherent story, the human mind reaches its limit. It is therefore not surprising that scholars have already for a longer time developed methods to organize their evidence outside their brains – not only in the field of Chinese Historical Phonology, but also in historical linguistics in general – ranging from punch card collections (Swadesh 1963) up to sophisticated database systems (Starostin 2000; Streeter 1972).

Since the use of computational systems for the purpose of organizing different kinds of data demanded a high level of computational knowledge and larger financial investments in the past, the use of computational databases to store and analyze data accumulated to study Chinese language history has for a long time been an exception rather than a rule. The past two decades, however, have drastically increased the possibilities for scientists to profit from the use of computers in their work, and it is therefore not surprising that more and more scholars start to build their own digital data collections. Similar to earlier punch card collections, these digital collections are rarely shared openly. Instead, they are treated as an integral part of the intellectual work done by individual scholars. What ends up in linguistic studies is often only the final result of a semi-automated process of data analysis: the part that scholars share to support their views or to reject the views of their peers.

While there is nothing to object in this procedure, as long as all evidence is shared in a given publication and no evidence has been deliberately discarded, the situation changes in those cases where scholars rely on quantitative methods and analyses, reporting only their results but not the code and data they used to achieve them. In these cases, research loses one of its most essential qualities: that it can be replicated by others. Unfortunately, we have seen quite a lot of irreproducible research in the past decades, specifically also reflected in numerous phylogenetic studies where scholars only showed the phylogenetic trees but would not share the cognate judgments from which the trees were derived (see Forkel et al. 2018 for
examples).

While data sharing is crucial in those cases where research could not be reproduced without access to the data, it is also useful to share and publish data where this might not be warranted directly by scientific standards. As I have mentioned before (List 2020a), apart from replication, sharing data transparently also contributes to the education of our colleagues, it makes falsification easier, and it allows for a quicker amplification of our collective knowledge. Specifically the aspect of education is important in this context, since the practice of collecting data in private exacerbates the introduction of standards and may also be harmful for research in those situations where scholars spend a lot of their time to collect data inefficiently or in ways that lead to a drastic information loss.

For this reason, I consider it important to encourage colleagues not only to share their data along with their publications, but also to start to discuss openly how data can best be stored, how they should be modeled, and how they can be analyzed. This study can be seen as a first attempt to initiate a discussion about the different types of data we collect for our research in Chinese Historical Phonology. After listing major types of data that we encounter when investigating the history of the Chinese language and its numerous varieties, I will briefly mention some initial ideas of how data in Chinese Historical Phonology could be standardized, and then discuss the chances and the challenges I see for the future of data in the field.

2. Data in Chinese Historical Phonology

As noted already by Li (1971) and others, types and evidence in Chinese Historical Phonology are numerous and complex. It is therefore difficult to provide a clear-cut and exhaustive classification of the datatypes that are crucial in the field. A very general distinction of data would distinguish primary and secondary sources. Rhyme patterns as they are reflected in the Shijing and other resources, for example, qualify as primary sources, given that they were not originally compiled for the purpose of studying Chinese historical phonology. Rhyme books, on the other hand, such as the Qiyeun, which provide information on the pronunciation, meaning, and form of Chinese characters, are secondary sources, since they were compiled for the purpose of providing explicit information on various aspects of Chinese characters.

Chinese characters are characterized by three major semiotic properties, their form, their meaning, and their reading (List 2009). Primary and secondary sources can
accordingly be distinguished by identifying the major types of information which they provide for the characters. While this may seem trivial at first sight, it is nevertheless important for all attempts to organize the data in Chinese Historical Phonology systematically, given that graphemics, semantics, and phonology all require a specific treatment, and we can profit from more general approaches to handle these specific aspects in general comparative linguistics.

In the following, I will quickly point to the basic types of data that have been popularly used in order to deal with character forms, character readings, and character meanings.

2.1. Character Forms

With respect to the form part of Chinese characters, we find several approaches to describe, analyze, and model it in classical and more recent secondary sources. The most typical way to describe Chinese character forms are beyond doubt those based on the traditional notion of the six character types (liūshū 六書), which were first most prominently demonstrated in Xù Shèn’s 許慎 (58-148 AD) Shuòwén Jièzì 說文解字 and have since then been of great importance throughout the history of Chinese linguistics. The most important character types among the classical liūshū are the xiēshēng characters which show a semantic and a phonetic component. The analysis according to the liūshū practice consists in the identification of the two components. In Xù Shèn’s work, this is usually (but not exclusively) annotated by the formula “A cóng 从 X Y shēng 聲”, which reads as “character A has X as semantic and Y as phonetic component”.

Despite their wide-spread use, it is not always clear to which degree liūshū 六書 and specifically traditional xiēshēng 諧聲 analyses account for the historical development of the Chinese characters. When acknowledging that Chinese characters were not coined de nova from one moment to the other, but rather derived over centuries and even millennia, an analysis should aim at reflecting this derivational process. The liūshū 六書 were most likely created for more practical purposes, allowing for an efficient knowledge organization in a situation where already thousands of distinct characters could be observed.

Starting from Duàn Yūcái’s 段玉裁 (1735-1815) famous detection that the phonetic part of xiēshēng 諧聲 characters often finds a direct correspondence in rhyme patterns and can therefore be used to predict whether characters rhyme in ancient sources or not, character classification reaches a “phonological turn”. Had
the *liūshū* 六書 practice been a mere classification scheme that would specifically emphasize the semantic component for mnemotechnic and organizational reasons, it was now possible to group characters by their abstract, historical sound shape and combine this information with the information that could be retrieved from rhyme analysis. This new kind of analysis would give preference to the phonetic rather than the semantic aspects of Chinese characters, which cannot only be justified by the predictive information which the phonetic elements bear for the rhyme behavior of Chinese characters, but also by the fact that – as all practitioners of historical linguistics have learned early in their career – sound change tends to proceed in much more regularly than semantic change (List 2014). While this does not automatically mean that the same should also hold for the choice of phonetic as opposed to semantic elements during character formation, there seems to be a rather broad consensus among scholars that phonetic components of the Chinese *xiēshēng* 諧聲 characters bear more systematic information than semantic components.

The most influential version of the *tóngbù* 同部 classification can be found in Bernard Karlgren’s *Grammata Serica Recensa* (Karlgren 1957), which grouped Chinese characters systematically by their phonetic parts and provided additional Old Chinese readings for each of them. Up to today, most reconstructions of Old Chinese phonology largely follow Karlgren’s classification, and although revisions of Karlgren’s work have been published (Schuessler 2009), it is still customary to assign characters to Karlgren’s identifiers from the *Grammata Serica Recensa* (Karlgren 1957) when providing their ancient readings.

Despite the importance of the *xiēshēng* 諧聲 series as a character grouping principle, the new grouping of characters based on their phonetic elements bears a certain shortcoming in comparison with Xu Shén’s approach, given that it ignores the fact that characters – as has been mentioned before – have been historically derived from each other, similar to the way in which words are derived from other words in the lexicon of human languages. In order to account for this important characteristics, Hill and List (2019) have proposed to model Chinese characters in the form of *character derivation graphs*. Technically, the approach can be seen as a reconciliation of the ahistorical *liūshū* classification with the more historically oriented classification of *xiēshēng* characters by common phonetic components. Instead of assigning each character to one abstract group (or series) as it has been done by Karlgren, the character formation graph analysis allows individual series
to be hierarchically structured. As a result, groups can be divided into an arbitrary number of subgroups, depending on the size and the internal structure of a given group of xiéshēng characters.

The annotation of Chinese character forms, be it in liùshū 六書 style, in xiéshēng 諧聲 style, or as a character derivation graph can conveniently be carried out with the help of simple tables. In such a table, each character is listed in a separate row, and additional columns provide the information needed to reflect what can be found in the original resources. While one could list both the phonetic and the semantic component of the liùshū 六書 analysis in the same cell, it is advisable to split the information carefully and add it to separate cells, following the general principle of transparent data organization that demands to avoid the mixing of information when working with tabular data (Forkel et al. 2018). This has three major advantages. It is first easier to describe the data, since the column names often speak for themselves, while complex internal structures often lead to confusion. Second, it is less error-prone, since annotation problems are easily spotted, and one can directly see, if, for example, specific values are missing. Third, it helps to test the data using software tools, since the data are essentially machine-readable.

Table 1 gives an example of such a tabular analysis of characters derived from fāng 方 “square”. While the analysis may seem simple, if not even trivial, to most scholars, it is important to note that the majority of data compiled by scholars in tabular form does not strictly follow basic principles of data annotation. As a result, it is very difficult to understand a given dataset when it is provided without further comments. For this reason, I recommend in general that all those who intend to carry out analyses in Chinese Historical Phonology which focus on Chinese character structures, and how they evolved over time, start by making a clear plan on how they intend to model and represent their data, and also share their data in transparent formats (spreadsheet tables rather than word documents with text in prose) in addition to the information they share in their publications.
List

<table>
<thead>
<tr>
<th>Character</th>
<th>MCH</th>
<th>Shuōwén 說文</th>
<th>xièshēng 謝聲</th>
<th>GSR</th>
<th>derivation graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>方</td>
<td>p a ng</td>
<td>象兩舟省，總 方</td>
<td></td>
<td>0740a</td>
<td></td>
</tr>
<tr>
<td>放</td>
<td>p a ng H</td>
<td>从支方聲。 方</td>
<td></td>
<td>0740i</td>
<td>方</td>
</tr>
<tr>
<td>旁</td>
<td>b a ng H</td>
<td>从二、闢方 聲。</td>
<td></td>
<td>0740f</td>
<td>方</td>
</tr>
<tr>
<td>諧</td>
<td>p a ng H</td>
<td>从言勇聲。 方</td>
<td></td>
<td>0740k’</td>
<td>旁</td>
</tr>
</tbody>
</table>

Table 1: Illustrating different approaches to a tabular analysis of characters.

2.2. Character Readings

The necessity to annotate the reading of a given character has led to a large amount of different techniques by which scholars can indicate how a given character should be pronounced in a given time. Starting from the often not very accurate dùruò 讀若 glosses in Xū Shén’s work (Coblin 1983, 5) and collections of loangraph glosses (Coblin 1983, 4), the most popular innovation was the use of fāngqì 反切 glosses (Branner 2000). Even more detailed systems were introduced with the compilation of the first rhyme books, which would not only provide fāngqì 反切 readings for individual characters but also group characters to homophone rhyme groups divided by their tone. While rhyme books can be seen as secondary sources on rhyme behavior, primary sources also play a major role. Here, again, we face the challenge to annotate consistently which words rhyme in a given poem.

Apart from these earlier techniques, more recent linguistic work devoted to the reconstruction of different historical stages of Chinese language varieties has introduced new types of evidence. Here, we find as our most important resources transliterations of Chinese words in other languages with phonologically less opaque writing systems, the spelling of borrowed words in Chinese (as reflected, for example, in Chinese Buddhist glosses, Coblin 1983, 7), and the witness of words that were borrowed from Chinese into neighboring languages.

While the information of the dùruò 讀若 and fāngqì 反切 type can be easily displayed in tables (following the principle of separating information in favor of separating the fāngqì shāngzi and the fāngqì xiàzi in separate columns, see also List 2018, 165–70), the annotation of rhyme patterns in poetry and the handling of transliterations, as well as spellings in Chinese, and borrowings from Chinese into earlier stages of other languages and language families warrants a considerably
enhanced treatment with respect to annotation. Regarding the annotation of rhyme patterns, we have shared first attempts towards a standardization of rhyme annotations across various languages earlier (List, Hill, and Foster 2019), proposing both a tabular stand-off annotation format (Eckart 2012), and a text-based inline-annotation, along with examples of annotation in various languages. The annotation is accompanied by the first draft of a software library that can be used to check if files which are annotated in the formats described in our study conform to the prescribed formats, and to carry out initial analyses (List 2019c). Along with the study, we also share annotated versions of the rhyme judgments by Baxter (1992) and Wáng (1980) for the Book of Odes (Shījīng 诗经), which can also be inspected interactively online (see https://digling.org/shijing/ and https://digling.org/shijing/wangli/).

In the meantime, I have had the chance to develop a first web-based application that uses the inline annotation format we proposed earlier (List, Hill, and Foster 2019) and allows for a convenient interactive annotation of rhymes in various languages. This Rhyme Annotation Tool (RhyAnT, https://digling.org/ryant/, List 2020b) is freely available and can be accessed directly through its project website. Although the tool does not provide any specific functions that would facilitate the processing of Chinese texts, it can already in its very general, language-agnostic stage, become a useful companion for those who wish to code their rhyme data consistently in such a way that they can later be analyzed automatically or manually. Figure 1 provides a screenshot in which the first three stanzas of the first poem in the Book of Odes have been annotated interactively.
With respect to transliterations, spellings in, and borrowings from Chinese, scholars have not only amassed a considerable amount of data from different languages and different epochs, but also provided very thorough analyses of their findings (Coblin 1983). In addition, these resources, and among them specifically borrowings from Chinese into the ancestor of the Hmong-Mien languages, have been playing an increasingly important role in the reconstruction of Old Chinese phonology (Baxter and Sagart 2014).

Given the importance of these specific types of evidence for the field of Chinese Historical Phonology, it would be desirable to include most recent developments in the area of sequence comparison in historical linguistics (List 2014) in order to model and present the evidence more consistently. In the field of historical linguistics, specifically in the relatively young field of computational historical linguistics, new methods propose to treat cognate word forms in different languages as sequences (that is, as sound chains), and to compare them with techniques originally developed for the handling of DNA and protein sequences in evolutionary biology. The most common approach to sequence comparison are
alignment analyses in which two or more sequences are compared by placing them in the rows of a matrix and arranging the individual sound segments in such a way that corresponding segments appear in the same column. While much of the recent work in historical linguistics has concentrated on finding ways to automate these alignment analyses, it is usually recommended to carry them out manually, since humans still tend to outperform machines, not only in linguistics (List 2014) but also in biology (Morrison 2018).

The major advantage of handling borrowings, transliterations, and spellings with the help of alignment analyses is that – once the alignments have been completed – many additional analyses which would cost humans a lot of time can be carried out automatically. One of the most important analysis in this context is the inference of sound correspondence patterns from aligned cognate words in historical linguistics (List 2019b). Although correspondence pattern analyses are a traditional part of the workflow of the comparative method (Ross and Durie 1996), they can also be very important and useful for the handling of borrowings, especially when it comes to the detection of contact layers (Lee and Sagart 2008). Instead of tabulating source words in Chinese and target borrowings in neighboring languages in order to find hints on the pronunciation of earlier stages of Chinese, I would therefore recommend to align source and target words in order to provide a more transparent understanding of the corresponding segments.

With the Etymological Dictionary Editor (EDICTOR, https://digling.org/edictor/), a tool for the interactive annotation of etymological data in historical linguistics has been available for some time now (List 2017). Up to today, the tool has been constantly approved. By now, it offers very sophisticated ways to handle very specific forms of partial cognacy (List 2016a), to align sequences in various ways, and to visualize and tabulate sound correspondences from aligned data (List 2019b). Figure 2 shows how the EDICTOR tool could be used to handle cases such as Buddhist spellings in Chinese, or borrowings from Chinese into neighboring languages.
2.3. Character Meanings

While it is clear that it is notoriously difficult to handle the meaning part of the linguistic sign (Fox 1995), it is also clear that semantics plays an important role in Chinese Historical Phonology. Specifically, we need to know the meaning of individual characters when searching for cognates in Chinese word families (Baxter and Sagart 1998), we also identify borrowings predominantly by searching for similar sound-meaning pairs across different languages, and we may even stratify them by investigating their specific semantics (List 2019a).

As in other branches of historical linguistics, semantics has received far less attention than phonology in Chinese Historical Phonology. This does not only hold for the semasiological perspective on characters and words, which starts from a given character or word form and then tries to explain which meanings the form can express, but also and specifically for the onomasiological perspective. Contributions such as the 250-item list of Old Chinese Vocabulary by Sagart and Ma (2020) or Chinese basic lexicon from a diachronic perspective by Starostin (2019), where authors try to identify the words that were used to express certain concepts in ancient varieties of Chinese are still extremely rare.

If we look at primary sources for character meanings in the history of Chinese, we find two basic types of evidence: direct glosses and tagged lists. While direct glosses are probably well known and do not need any further explanation, tagged lists refer to the typical collections of characters sharing specific semantic properties as we find them first in the Ėryā 碰彝, typically presented in the form A, B, C, D yê 也, which tells us that characters A, B, and C all have a certain
resemblance with D. I call these lists *tagged lists*, because we can treat the explaining character in the end of such a list as the *tag* that is used to describe the preceding characters. Apart from direct definitions, tagging is a widespread technique by which certain semantic aspects are assigned to specific words (Chechuro 2020). Similar to direct glosses, tags can be easily presented in tabular form and displayed in the form of *bipartite networks* or *hypergraphs* (List 2021). Table 2 shows how characters, tagged with *zhī* 止 “foot” in the first section of the Ėryā 爾雅 can be displayed as a tagged list in tabular form, along with additional tags in which they also occur.

<table>
<thead>
<tr>
<th>Character</th>
<th>References</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>廢</td>
<td>3 93 125</td>
<td>大 止 舍</td>
</tr>
<tr>
<td>戾</td>
<td>5 16 78 93</td>
<td>至 罪 待 止</td>
</tr>
<tr>
<td>懷</td>
<td>5 73 93</td>
<td>至 思 止</td>
</tr>
<tr>
<td>徵</td>
<td>8 93</td>
<td>善 止</td>
</tr>
<tr>
<td>替</td>
<td>78 93</td>
<td>待 止</td>
</tr>
<tr>
<td>底</td>
<td>78 93</td>
<td>待 止</td>
</tr>
<tr>
<td>証</td>
<td>93</td>
<td>止</td>
</tr>
<tr>
<td>妥</td>
<td>93 168</td>
<td>止 坐</td>
</tr>
<tr>
<td>安</td>
<td>93 168 170</td>
<td>止 坐 定</td>
</tr>
<tr>
<td>按</td>
<td>93</td>
<td>止</td>
</tr>
<tr>
<td>尼</td>
<td>93 166</td>
<td>止 定</td>
</tr>
<tr>
<td>定</td>
<td>93</td>
<td>止</td>
</tr>
<tr>
<td>易</td>
<td>93</td>
<td>止</td>
</tr>
<tr>
<td>過</td>
<td>93</td>
<td>止</td>
</tr>
</tbody>
</table>

Table 2: Displaying the information on characters in the Ėryā 爾雅 in the form of tags.

3. Chances for Quantitative Approaches in Chinese Historical Phonology

In the preceding sections, I have been trying to illustrate how a more conscious data handling can help us to assemble different types of data important for the field of Chinese Historical Phonology in a way that both machine- and human-readable. Once the data have been assembled in this form, several analyses can be fruitfully carried out and may assist scholars in finding patterns that may help them to analyze concrete hypotheses. In the following, I will point to some very general methods that can be applied to analyze machine-readable data.
3.1. Corpus Studies

Corpus studies have traditionally been playing an important role for traditional studies on Chinese Historical Phonology. Major contributions have been achieved by scholars who consulted huge collections of ancient text material in order to search for patterns to emerge from the data they collected. In addition, corpus studies have also proven useful in the determination of the semantics underlying certain characters (Starostin 2013). Today, digital corpus collections greatly facilitate the search in ancient texts. If corpora are furthermore accessible for queries with regular expressions or provide software application programming interfaces (APIs) that allow to access the corpus data with the help of programming languages, the life of scholars searching for patterns in ancient text collections can be facilitated even more.

As an example, consider the numerous ways in which loangraphs are described in the traditional Chinese commentary literature (Coblin 1983, 12), ranging from expressions such as x 論 y, via x dūdangwei y up to x duyu y tong. Once the most important expressions to indicate that a given character x should be read as another character y have been identified, it is straightforward to conduct an automated search for these patterns on a given corpus collection and to assemble the findings automatically in tables that could then be inspected and evaluated by scholars.

Automated corpus studies have a huge potential for Chinese Historical Phonology. It is probably only a matter of time until first studies will appear which make exhaustive use of various methods for information extraction, as they are commonly used in the field of Natural Language Processing.

3.2. Network Approaches

Network approaches have proven to be very useful for a large variety of tasks in historical linguistics, ranging from initial attempts to detect borrowings in lexical datasets (Nelson-Sathi et al. 2011), via polysemy (or colexification) networks (List, Terhalle, and Urban 2013), up to novel approaches to the automated identification of sound correspondence patterns (List 2019b). Network approaches have also been applied in various tasks in Chinese Historical Phonology, including the application of rhyme networks (List 2016b; List et al. 2017), the analysis of Chinese characters in the form of character derivation graphs (Hill and List 2019), or initial attempts to analyze fānqiè 反切 characters with the help of fānqiè networks (List 2018).
Given that network approaches are widespread in many scientific fields, flexible to model, and easy to analyze – thanks to the fact that they are well supported in many software packages –, it is very likely that they will continue to play an important role in Chinese Historical Phonology as well. Up to now, their full potential has barely been exhausted, and it is very likely that we will see new and interesting approaches in the nearer future.

3.3. Phonetic Alignment and Cognate Annotation

Methods for the automated and computer-assisted detection of cognates and phonetic alignments have greatly improved during the past decade (List, Greenhill, and Gray 2017). In the last couple of years, further improvements have also been made with respect to the flexible annotation of complex cognacy relations (Schweikhard and List 2020) and detailed workflows for the computer-assisted investigation of etymologies derived from lexical data have been developed (Wu et al. 2020). While these approaches have not yet been applied to address traditional problems in Chinese Historical Phonology, I see great potential for their application, specifically when it comes to the handling of ancient borrowings from early varieties of Chinese into neighboring language families, and to the exploration of transliterations of foreign words in Chinese language history.

4. Challenges for Quantitative Approaches in Chinese Historical Phonology

With respect to the chances offered by quantitative approaches in the field of Chinese Historical Phonology, I feel rather optimistic, and I expect that the next two decades will yield a wealth of promising new approaches. This should, however, not gloss away from the particular challenges that the field faces at the moment. Here, the experience from neighboring disciplines seems particularly important to me.

In the field of historical linguistics, for example, we have on the one hand seen a surge in computational approaches and also a raise in approaches that make active use of larger datasets. At the same time, however, there are still far too many approaches in which scholars do not supply neither data nor code along with their studies. The reasons for missing data and code are manifold, ranging from pure ignorance with respect to the importance of data sharing up to the deliberate wish to not share data and code in order to avoid that colleagues would replicate or use them for additional analyses.
Even if data and code are shared, it is often difficult to reuse them, since scholars often do not provide the metadata or explanations one would need in order to interpret their data correctly, or since scholars code their data in a form that would facilitate their reuse. This so far largely unsatisfying situation in the field of historical linguistics has lead to some recent initiatives that try to overcome the current problems of data quality and data sharing by providing concrete guidelines on data citation (Berez-Kroeker et al. 2018) and new frameworks for data curation and sharing (Forkel et al. 2018).

What one can see from these recent initiatives is that every change needs time, and although scholars’ awareness has been increasing lately, with more and more journals insisting that data and code be submitted to scientific repositories and that reviewers are given access to data and code already during the stage of peer review, the large amount of data and code produced in linguistic research still lacks sustainability and is not well integrated. As a result, it is still very difficult to make active use of the wealth of data that has been produced of late.

For the field of Chinese Historical Phonology with its complex types of evidence and the long tradition of individual – as opposed to collaborative – data management, the major challenge for quantitative approaches can therefore be expected when it comes to the development of initial standards and examples for best practices. Since these efforts can only be addressed by the community as a whole, or at least by a critical amount of scholars who would be ready to devote time and resources to the development of common standards, it is difficult to say if and when this will happen in the field of Chinese Historical Phonology. In order to foster the full potential of quantitative applications, however, it will be crucial that questions of standardization and integration of data will be addressed rather sooner than later.

5. Outlook

In this study I have summarized some thoughts regarding the potential of quantitative approaches in the field of Chinese Historical Phonology. Starting from a discussion of the numerous types of evidence we find in the field and how they could be best modeled and represented, I have tried to give some concrete examples on the potential of future quantitative studies. While there are a lot of chances for computational and quantitative approaches in Chinese Historical Phonology, I also identified the standardization and integration of the data we use
in the field as the major challenge for future work. All in all, there is a lot of reason to be optimistic with respect to the potential of computational approaches, and it is quite likely that we will soon see the rise of new methods that could further advance the whole field.

References


List

Chances and Challenges 2021


Schweikhard, Nathanael E., and Johann-Mattis List. 2020. Developing an annotation framework


Acknowledgments

This research would not have been possible without the Ifk Young Scholars Symposium (Academia Sinica, Taipei, 2017), generously hosted by the Li Fang-Kuei Society for Chinese Linguistics (http://ifksociety.org/), during which the author presented parts of the approaches mentioned here. I would like to acknowledge the generous support of the European Research Council for supporting this research under the auspices ‘Computer-Assisted Language Comparison’ (715618, ERC Starting Grant). I would also like to thank Nathan W. Hill and Christopher Foster for inspiring discussions on the topic of standardization and modeling of data in Chinese Historical Phonology.