Cite as: Hill, N. W. and List, J.-M. (forthcoming): Using Chinese character formation graphs to test proposals in Chinese historical phonology. Bulletin of Chinese linguistics.

Using Chinese character formation graphs to test proposals in Chinese historical phonology

Nathan W. Hill and Johann-Mattis List

Abstract

This paper proposes the use of network techniques in the exploration of Old Chinese phonology as reflected in the phonophoric determinatives of *xiéshēng* 諧聲 characters. We use the approach to examine five specific proposals in Chinese historical phonology, and whether they can be said to be recoverable on the basis of phonophoric choice. The major finding is that the type A versus type B distinction is in some cases encoded in the choice of phonophoric, while other distinctions are only spuriously if at all reflected in the phonophoric subseries.

Keywords

network approaches, Historical Chinese Phonology, Chinese character formation, reconstruction of Old Chinese phonology

1 Introduction

In contrast to many writing systems (Tangut, ’Phagspa, Hangul, Cherokee), Chinese characters were not created as a one-off, but instead developed over millennia. As a result, the formation of Chinese characters (*zàozìfǎ* 造字法, Qiú ([1988] 2007)) is best viewed as a derivational process with striking similarities to processes of word formation (Kunze 1937, List 2008, 41–44). This derivational process applies specifically to the phonetic characteristics of the writing system, as reflected in the category of *xiéshēng* 諧聲 characters, which consist of one element that hints at the pronunciation of the word encoded by the character (the phonophoric determinative), and one element that hints at the word’s meaning (the semantic determinative) (see Boltz 1994, 71–72). For example, the character 被, which writes the word *bjeX* ‘cover oneself with’ is composed of the phonophoric determinative 皮, which as a character itself represents the word *bje* ‘skin’, and the semantic determinative 衤, a contracted version of 衣 *ʔjɨj* ’clothes’. The Middle Chinese pronunciations *bjeX* and *bje* differ only in tone.[1] The meaning ‘clothes’ is sufficient to distinguish *bjeX* ‘cover oneself with’ from 疲 *bje* ‘weary’, 陂 *pje* ’dam’, etc. At the time of the development

1 We give Middle Chinese in the system of (Baxter 1992) and Old Chinese in the system of (Baxter and Sagart 2014); for the purposes of the present paper this choice is arbitrary and has no affect on the argument.

of the Chinese script, in order for a character typically associated with a particular word to serve as a phonetic indication of a word with an unrelated meaning, the two words must have been sufficiently similar in sound.

Duàn Yùcái 段玉裁 (1735-1815), in his 六書音均表 *lìushū yīnyùn biǎo*, first elaborated the principle, called the '*xiéshēng* hypothesis', that the same phonophoric determinative in the writing of two characters implies that the words expressed by these characters were able to rhyme in the *Book of Odes* (*Shījīng* 詩經, ca. 1050–600 BC).[2] Li Fang-kuei 李方桂 adds the stipulation that each Old Chinese rime category have one vowel (Li 1974, 243, (Baxter 1992, 348, Schuessler 2009, 11). For words that occur as rhyme words in the *Odes* and are represented by characters with the same phonophoric determinative, whether or not the readings of these characters rhyme is a testable hypothesis. There are many such cases, e.g. 袺 *ket* < *kˁit and 襭 *het* < *gˁit rhyme in Ode 8.3 and 脫 *thwajH* < *l̥ˁots and 帨 *sywejH* < *l̥ots in Ode 23.3. For characters that do not occur as rhyme words in the *Odes* the *xiéshēng* hypothesis is necessarily an assumption.

The structure of the Chinese script not only provides a way of confirming the rime category of words that do not occur as rhyme words in the *Odes* it also provides the primary mechanism for exploring Old Chinese initials, about which the *Odes* are silent. The analogous theory used for initials, also called the '*xiéshēng* hypothesis', is the presumption that all morphemes written with characters that share the same phonophoric determinative originally began with sounds of the same place of articulation (Li 1974, 228, Baxter 1992, 348, Schuessler 2009, 11).

A suite of characters sharing the same phonophoric determinative are referred to as a '*xiéshēng* series'. Zhū Jùnshēng 朱駿聲 (1788-1858) arranged characters by their phonophoric determinatives in his 說文通訓定聲 *Shuōwén tōngxùn dìngshēng* (1832). Karlgren (1957) produced a similar work, with the advantage that he provides reference numbers for each series. The *xiéshēng* series built on 皮 *bje*, with the reference number #0025 in Karlgren's system, illustrates both the rhyme and the initial facets of the *xiéshēng* hypothesis.

- 皮 *bje*
- 被 *bjeX*
- 陂 *pje*
- 披 *phje*
- 跛 *paX*
- 簸 *paX*, *paH*
- 破 *phaH*
- 婆 *ba*

2 While the usefulness of *xiéshēng* characters for historical reconstruction was already noted much earlier (Behr 2004, 33–35); Duàn Yùcái first noted that *xiéshēng* characters with the same phonophoric determinative could freely rhyme in the *Odes*.

All Middle Chinese readings in this series have homorganic initials (唇 *chún* bilabials), and it is reasonable to assume that this state of affairs is a continuation of what prevailed in Old Chinese. These readings show two Middle Chinese rimes, *-a* (*gē* 歌 *ka*) and *-je* (*zhī* 支 *tsye*). However, words with these rimes regularly rhyme in the *Odes*; for example, 皮 *bje* rhymes with 紽 *da* in Ode 18.1 and 宜 *ngje* rhymes with 何 *ha* in Ode 47.1. Thus, one may presume that *-a* and *-je* descend from a single rime, namely *-aj, which developed into -a in one phonetic environment (type A syllables) and *-je* in another phonetic environment (type B syllables).[3] The *xiéshēng* hypothesis has equipped historical Chinese phonology with a powerful analytic tool. With this tool at hand, the reconstruction of Old Chinese phonology has steadily improved, culminating in a series of very similar reconstruction systems, proposed independently by different scholars in the early 1990s (Behr 1999) and since then continually improved (Sagart 1999; Schuessler 2007; Baxter and Sagart 2014).

Because the phonophoric characters were not all formed in the same time period, their formation reflects different stages in the complex history of the Chinese language. Similar to compound words such as German *Krankheitsverlauf* 'disease progression', which reflects different stages of derivation and compounding, taking place at different times (see List et al. 2016 and Table 1), a Chinese character like 膀 *bang* 'wing' reflects different layers of formation. The character appears to have 月 *ngjwot* as a semantic element with 旁 *bang* 'side' as a phonophoric element, but historically this results from a visual confusion of 月 *ngjwot* and 肉 *nyuwk* 'flesh', which as a character component takes the form 月. In turn, the character 旁, which writes the word *bang* 'side, broad', has 方 *pjang* 'square' as its phonophoric determinative and 凡 *bjom* 'tray, dish' as its semantic determinative.[4]

Word	Approximate Origin	Meaning	Stage of German
kranc	900	'weak'	Old High German
kranc	1200	'not healthy'	Middle High German
kranc-heit	1200	'weakness'	Middle High German
(h)loufan	700	'run, dance'	Old High German
(h)louf	800	'gait, course'	Old High German
Ver-lauf	1400	'process, development'	Early New High German
Krank-heit-s-ver-lauf	1800	'disease progression'	Modern German

Table 1: Different stages of German reflected in the compound Krankheitsverlauf 'disease progression'. Data taken from the DWDS (Geyken 2010) and Pfeifer (1993).

3 It may seem simpler to propose *a as the origin of *-a* and *-je*, but final *-j is required for two reasons: 1. to account for pronunciations such as Foochow dialect *muai* 2 and Sino-Korean *may* for the character 磨 (MCH *ma*) (Baxter 1992, 293–94), 2. to leave OChi. *-a available as an origin of MChi. *-u* to explain such phenomena as the early transcription of Buddha as 浮屠 (MChi. *bjuw du*, see Pulleyblank 1983: 78) or cognates such as 吾 (MChi. *ngu*) 'I, me' compared to Tib. *ṅa* and Bur. *ṅā* 'id.'. Type A and type B syllables are discussed further below.

4 In fact, the history of 旁 is more complex than presented here, since many different variants for the character exist; further detail is not relevant here.

2 Phonophoric Characters in Old Chinese Reconstruction

As described above, the standard way of employing *xiéshēng* characters in Old Chinese reconstruction is to assemble characters into *xiéshēng* series in which characters that share the same phonophoric are brought together. From the divergent character readings that can be found in Middle Chinese, as originally reflected in the rhyme book *Qièyùn* 切韻 (601 AD), scholars carry out an internal reconstruction of Old Chinese character readings to explain differences in Middle Chinese based on the assumption that readings with the same phonophoric determinative should have homorganic initials, the same nuclear vowel, and the same coda (modulo final -ʔ and -s) (for details, see Baxter 1992). There is no ironclad methodology for deciding upon membership of a *xiéshēng* series and scholars often disagree with one another in their grouping of characters into series. In particular, what a particular researcher sees as the phonophoric determinative in a given character does not necessarily coincide with the classical practice of *liùshū* 六書 analysis as presented Xǔ Shèn's *Shūowén Jiězì* 說文解字 (121 AD). For example, as shown in Table 2, Karlgren assigns the character *bǎng* 膀 'wing' to the *xiéshēng* series built on *fāng* 方 'square' (Karlgren 1957#0740), whereas Xǔ Shèn gives the phonophoric determinative as *páng* 旁 'side' (从肉。旁聲). In other words, whereas Karlgren identifies the same phonophoric 方 for all characters in Table 2, Xǔ Shèn provides two phonophorics, viz. 方 and 旁. We can see that those characters where Xǔ Shèn identifies 方 as the direct phonophoric determinative belong to the third division (*sānděng* 三等) in the rhyme table system of Middle Chinese (as reflected by the medial *-j-* in the Middle Chinese transcription of Baxter (1992)),[5] while the characters that are given 旁 as phonophoric all belong to the first division (*yīděng* 一等). This pattern gives preliminary evidence that the mysterious Old Chinese A/B distinction, which divides Old Chinese syllables in two distinct groups, one surfacing as third division in Middle Chinese, the other as non-third division (Sagart 1999), is – at least to some degree – also reflected in the *xiéshēng* characters.

Character	Middle Chinese					Phonetic (GSR)	Phonetic (SW)
方	p	j	a	ng		-	-
放	p	j	a	ng	H	方	方
昉	p	j	a	ng	X	方	Ø
舫	p	j	a	ng	H	方	方
牥	p	j	a	ng	H	方	Ø
旁	b		a	ng		方	-
謗	p		a	ng	H	方	旁
騬	p		a	ng	H	方	旁
膀	p		a	ng	H	Ø	旁

Table 2: Contrasting Karlgren's and Xǔ Shèn's analysis of Chinese characters containing 方. *Ø indicates that the character is missing in the source, the dash indicates that the character is not analysed phonetically.*

5 For an overview on the role of divisions, see Shen (2017).

That *xiéshēng* series are further divisible is by no means a new discovery. This observation is reflected in later revisions of Karlgren's influential system (Zhèngzhāng 2003; Schuessler 2009), and in recent specialist literature, where scholars have illustrated that subseries of a given *xiéshēng* series may reflect not only the A/B distinction (Sagart and Mǎ 2017), but also further distinctions, like uvulars versus velar stops Hill (2015, 53), or vowel quality (Schuessler 2009, 246f). Nonetheless, the discipline still lacks a consistent way to model and analyze which *xiéshēng* series can be subdivided and how these subdivisions can be used to substantiate character formation hypotheses.

3 Modeling Chinese Character Formation in Directed Networks

Recalling the aforementioned analogy between character and word formation, we find inspiration in the handling of word formation in linguistics, where the usage of networks to model word formation is common morphology (Hippisley 1998; Ševčíková and Žabokrtský 2014). A crucial aspect of word formation (but also of character formation) is the hierarchical process by which words are derived from each other at different times. If we have a compound word, like German *Krankheitsverlauf* 'disease progression' we can recursively split the word into its respective components which usually were coined at different moments in history. This analysis, which is illustrated in Figure 1, reflects a directed network of word formation processes in which the different subdivisions of the compound correspond to the nodes in the network and the directed edges reflect the compounding processes Given the similarity between word and character formation, it seems straightforward to also use networks to model processes of character formation in the history of the Chinese writing system.

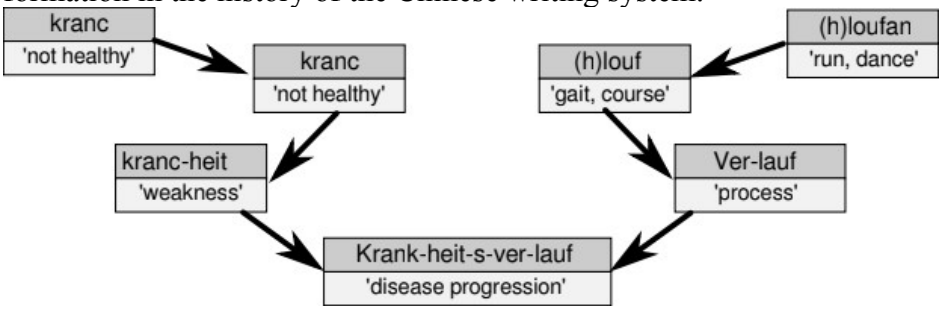

Figure 1: Word formation network for German Krankheitsverlauf, based on information drawn from the DWDS (Geyken 2010) and Pfeifer (1993).

We prepared a model in which Chinese characters sharing phonetic elements are represented in a directed network. The principle by which the network is constructed are straightforward. Nodes in this network are represented by characters, and phonetic relations between characters are represented by directed edges. Whenever a character *A* can be directly divided into a semantic and a phonophoric determinative and the phonophoric determinative *B* constitutes a character in its own right, an edge is drawn from the *B* to *A*, reflecting that *A* has *B* as its phonophoric determinative. For small series such a network can be easily drawn manually, but for larger

amounts of data, it is useful to automate this process, especially also, because thanks to recent digitization efforts, a large amount of data on character structure and *xiéshēng* analyses by different authors are now freely available.

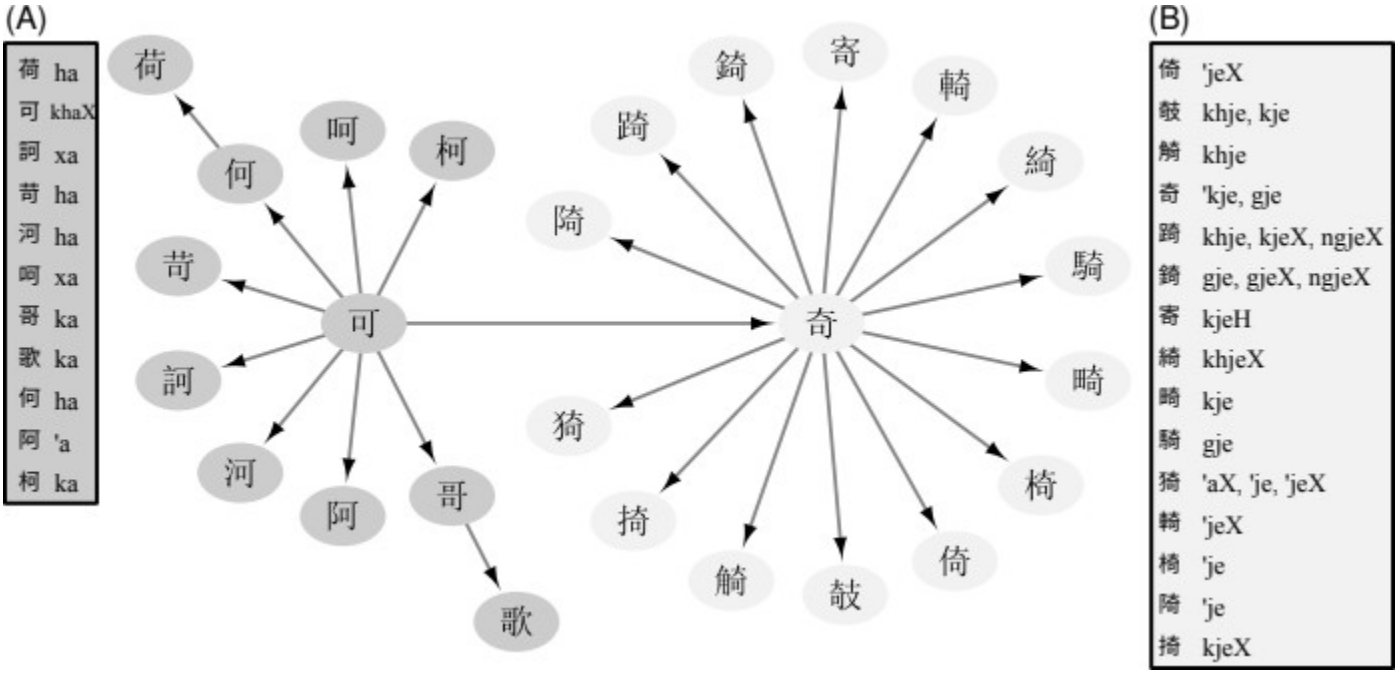

Figure 2: Character formation network of Karlgren's series #0001. Comparing the division into two clusters with the Middle Chinese character readings, one can clearly see that the A/B distinction is also reflected in this series.

Since this is a first experiment on character formation networks, we restricted the analysis to the characters and *xiéshēng* series of Karlgren (1957) which have recently been made available online in the form of a WikiBook (Wikipedia Users 2017). Since Karlgren's analysis only contains information on *xiéshēng* series in general, but not on their further structure, we supplemented Karlgren with a dataset on Chinese character structures provided by Kawabata (2014).[6] This dataset provides information on the semantic and the phonetic structure for as many as 18,348 Chinese characters. Information on the phonophoric determinative of a given character follows Xǔ Shèn's practice, by which, for example, the character 房 *bjang* is glossed as 方聲, indicating that 房 *bjang* has 方 *bang* as its phonophoric determinative.[7] Of the 5,142

6 More recent analyses of *xiéshēng* series (such as Schuessler 2009, and anecdotally in Baxter and Sagart 2014), tend to lump together series that Karlgren had divided, largely on the basis of subsequent research on earlier forms of the Chinese script. In principle such 'lumping' approaches risk loosing information, but this danger is mitigated by a network theory approach; a network preserves the finer structure even when series are joined together. For the purpose of the current publication there is merit in maintaining the more traditional analysis of Karlgren. More recent *xiéshēng* analyses can be explored network-theoretically in future research.

7 Unfortunately, no further information on the criteria by which phonetic components were identified for each character could be found. Manually checking parts of the data shows that the judgments largely follow those of Xǔ Shèn. Of course, Xǔ Shèn's explanations are not always correct (e.g. see Sagart and Baxter 2012).

characters in the digital version of the GSR, Kawabata's data provides a phonophoric analysis for 3,783, corresponding to 625 *xiéshēng* series in Karlgren's analysis. From these characters, we constructed the *xiéshēng* network by adding directed edges for all characters inside a given *xiéshēng* series.

An example of the network is given in Figure 2, where the series built on 可 *khaX* #0001 is plotted along with additional character readings in Middle Chinese This network clearly reflects the A/B distinction, with characters directly derived from 可 corresponding to type A and characters derived from 奇 *gje* reflecting type B. This graphic immediately suggests that the coiner of the character 寄 self consciously chose 奇 and not 可 itself as a phonophore in order to indicate that the word *kjeH* 'send' he was transcribing was of syllable type B. The network in Figure 2 also suggests that a more systematic search for *xiéshēng* subseries that distinguish type A and type B syllables would be profitable and indeed that the same graph theoretic approach would be useful for exploring further hypotheses about Old Chinese phonology.

4 Studying Old Chinese Phonology with Character Formation Graphs

The framework for inspecting character formation networks is based on two types of evidence, (1) a character formation network (which could follow the analysis of any scholar), and (2) a list of Middle Chinese readings for all characters in the network. The character formation network along with Middle Chinese readings is provided in the form of a text file in tabular form (as it is known from spreadsheet editors, like Excel or LibreOffice), with the first line serving as a header and the consecutive lines providing the content. Our current data starts with Chinese characters (`CHARACTER`), which are assigned their respective *xiéshēng* series by Karlgren (1954) (`KARLGREN`), and are further *defined* by their phonophoric character (`XIESHENG`).[8] In addition, each character is given all Middle Chinese (`MCH`) readings in the system by Baxter (1992) available to us from the data.

Our approach for querying this system is based on a regularized form of queries regarding the Middle Chinese readings contrasted with the general *xiéshēng* series (as represented by Karlgren's analysis) and the subseries (as taken from Kawabata 2014). To keep the queries simple, our approach allows one to query up to three characteristics per Middle Chinese reading, of which the first two need to be ask in the form of logical questions with either *true* or *false* as the only acceptable answer. The basic idea of the queries is to check for the "purity" of subseries

8 In an earlier study, which this paper proceeds from, List (2018) illustrated the usefulness of character formation graphs by searching automatically for *xiéshēng* networks that exhibit an A/B contrast, using the rough criterion by which each subseries with more than 70% of third division readings was classified as reflecting an entire B-type series. Table [tab:xiesheng] shows eight of the 16 relevant series identified with this procedure.

In such an approach the usage of different thresholds will yield different results; as is typical in philological disciplines – we are confronted with sparse rather than big data in our research. Consequently, we decided to design an alternative, exhaustive, experiment with more clear-cut rules for selecting *xiéshēng* series which would not only allow experts interested in Chinese character formation and Chinese historical phonology to sift through the relevant data themselves, both in order to test existing hypotheses and in the hope of generating new hypotheses.

in a given *xiéshēng* series, by placing characters into different groups. With respect to the A/B distinction mentioned above, for example, we check two logical conditions per reading: if a reading, such as *fāng* 方 'square' MCH *pjang* contains the vowel *-i-* or the medial *-j-*, we assign it to group 1, the group for type B readings, if it contains neither, as in the case of *bǎng* 膀 'wing' MCH *pang* we assign it to group 2, the group for type A readings. To paraphrase in more general terms, if a character's reading fulfils the first condition, we place the character in the first group, and in the second group, if it fulfils the second condition. We reserve a third group for those characters that cannot be assigned to either of the two basic groups.[9] Once this has been done, we can automatically determine the *purity* of character assignments per subseries. If a subseries consists only of characters conforming to the first query, it is pure with respect to the first condition. The same holds for subseries conforming only to the second query. If a series contains characters assigned to the first group and assigned to the second group, then the series is labelled "mixed".

In order to restrict the number of *xiéshēng* series which are returned to the user, we further allow one to define a criterion to keep or ignore a given series, based on the presence or absence of characters per series conforming to the two base queries and a third condition, which can be more complex than simple yes/no questions. The basic idea here is that the two logical queries can be combined by checking their common truth value, which could require both to be true, one to be true and the other to be false, both to be false, etc. The third condition offers a bit more flexibility, as it can be used to query concrete aspects of a character reading (like the initial). In fact, this increased flexibility we only use in one of our case studies (Section 5.2).

The supplementary material accompanying this paper contains the original character formation graph that we used in this study in the form of a tab-separated-value (tsv) file along with code in Python and Shell to replicate all analyses reported here. The analysis makes use of the well-established LingPy software library (List et al. 2018, http://lingpy.org) for quantitative tasks in historical linguistics to handle phonetic sequences in the data. In addition, it uses the recently published SinoPy library (List 2018, https://github.com/lingpy/sinopy) for processing of Chinese characters and Middle Chinese readings. Given that the code is highly modular and easy to modify, users with a little experience in Python programming can easily design their own queries of the character formation graph data.

5 Case Studies on Character Formation Graphs

In the following, we present five case studies, in which we present and discuss the findings of our new framework to study character formation for the purpose of Old Chinese reconstruction. For each cases study, we provide some examples in the text, and all data in the supplementary material. For each case study, we list the two queries along with the conditions by which *xiéshēng* series are included in the report.

Here we look for evidence that subseries of Karlgren's *xiéshēng* series systematically distinguish type A and type B syllables [Hypothesis *Sagart*] (Section 5.1), evidence that some subseries

9 It is clear that for the A/B distinction, there is no third condition, as our two initial conditions mutually exclude each other.

distinguish uvulars versus velars initials [Hypothesis *Pan* (Pān Wúyùn)] (Section 5.2), evidence for final *-r as distinct from *-n and *-j [Hypothesis *Starostin*] (Section 5.3), *-ps versus *-ts [Hypothesis *Haudricourt*] (§5.4), the *kr- versus *kər- [Hypothesis *Gabelenz*] (Section 5.5).

5.1 The A/B distinction in Old Chinese [Hypothesis Sagart]

While the example of the series built on 可 *khaX* #0001 itself is already interesting in the context of *xiéshēng* series and Old Chinese reconstruction, the real advantage of using an automatic approach for the construction of *xiéshēng* networks is that we can now search the network automatically for similar series in which type A and type B are contrasted. In order to do so, we defined the search criteria for our automated approach as specified in Table 3. The first condition defines type B syllables in Middle Chinese readings as containing either a *j* or an *i*. Type A syllables are defined negatively. The query procedure now lists all *xiéshēng* series which have syllables of both types.

Condition 1	$B_{type} := \texttt{"j"} \in \texttt{MCH} \cup \texttt{"i"} \in \texttt{MCH}$
Condition 2	$A_{type} := \neg B_{type}$
Show Series	$B_{type} \in \texttt{XS} \cap A_{type} \in \texttt{XS}$

Table 3: Conditions for A/B distinction.

The results of this search comprise 293 different series, which are all given in the file `sagart.pdf` in the supplement. Not all of these examples are as interesting as the case of 可 *khaX* #0001, shown above, since we find quite a few cases of *mixed* output, in which syllables of type A and B can be found in one of the subseries. Investigating all these cases would amount to a book-length study of its own. We can, however, identify quite a few interesting cases in which the A/B distinction is quite clearly reflected.

Among the most interesting cases we identified by eyeballing the data, are (besides the case of #0001):

- #0002, which can be split into 我 *nga*, encoding almost exclusively type A syllables and 義 *ngjeH*, encoding exclusively type B syllables,
- #0094, which can be split into 女 *nrjoH* (type B),

  如 *nyo* (type B), and 奴 *nu* (type A), which was already discussed by Sagart and Mǎ Kūn 馬坤 (2017),
- #0256, with 袁 *gjwieng* pointing to type B, and 爰 *hwaen* pointing to type A, and
- #0740, with 方 *pjang* (type B) opposed to 旁 *paeng* (type A), which was discussed above.

5.2 Uvulars in Old Chinese [Hypothesis Pan]

Some *xiéshēng* series mix readings with Middle Chinese velars (見 *k*-, 群 *g*-, 匣 *h*-, 曉 x-) and readings with initial glottal stop (影 ʼ-) or initial 以 *y*-. Following a proposal of Pān (2000), Baxter and Sagart assume that these originate from uvular initials (Baxter and Sagart 2014, 28,

3–46, 168–70), Sagart and Baxter 2009). Hill (2015, 53) points out that uvular initial series built on 或 #0929 has a velar initial subseries built on 國 *kwok*.

Condition 1	$I_{velar} := \texttt{MCH}_{initial} \in$ `["k", "g", "kh", "x", "h", "ng"]`
Condition 2	$I_{glottal} := \texttt{MCH}_{initial} \in$ `["'", "y"]`
Show Series	$I_{velar} \in \texttt{XS} \cap I_{glottal} \in \texttt{XS} \cap \texttt{MCH}_{initial}\neg \in$ `["m", "th", "d"]`

Table 4: Conditions for uvular distinction.

Our search for similar examples is based on the conditions given in Table 4. As our first condition, we define the set of velar initials. As our second condition, we define the set of glottal initials. We force the search procedure to list only those cases in which we find a velare initial, a glottal initial, and the whole series does not contain dental plosives or the nasal *m*.

The results of this search comprise 30 different series, which are all given in the file `pan.pdf` in the supplement. Unfortunately, the results show that our filter procedures were not successful in identifying the relevant cases, as they could not even detect the test cases upon which the hypothesis was originally built. All we can conclude for now is that the uvular distinction is not as strongly reflected in the Chinese character formation process, as the A/B distinction. More research and more experiments to improve the search procedure will be needed before any definite conclusion can be made with respect to Pān's uvular hypothesis.

5.3 Contact between Middle Chinese -n and -j [Hypothesis Starostin]

Starostin motivated final *-r in Old Chinese, primarily on the basis of phonetic series, that mix reading that have both final *-an* and (*aj >) *-a* in Middle Chinese (Starostin 1989, 399–401). The series built on 釆 #0195 is a case in point, as it includes readings such as 蟠 *ban* < *bˁar and 蟠 *ba* < *bˁaj < *bˁar. Rhyme evidence seems to further support this hypothesis (List 2017). Baxter and Sagart provide extensive evidence to argue that *-r > *-n is the mainstream development whereas *-r > *-j is typical of an eastern dialect (Baxter and Sagart 2014, 242–68). Although the overall thrust of Baxter and Sagart's discussion strongly suggests *-r as part of the Old Chinese syllable canon, because they approach the evidence unsystematically it is difficult to be confident of *-r in any one of their reconstructions (Hill 2017).

To search for cases that could further support Starostin's hypothesis, we applied the criteria outlined in Table 5. Our first criterion searches for Middle Chinese finals ending in *-n*. Our second criterion searches for finals *not* ending in *-ng* and *-t* (as these show also frequent contacts that are not relevant for the hypothesis at hand). We list all series in which we can find character readings fulfilling the first and the second condition.

Condition 1	$F_n := \texttt{MCH}_{final} \in$ `["n", "nH", "nX"]`
Condition 2	$F_{not-n-ng-t} := \texttt{MCH}_{final}\neg \in$ `["ng", "t"]`
Show Series	$F_n \in \texttt{XS} \cap F_{not-n} \in \texttt{XS}$

Table 5: Conditions for final -r.

The results of this search comprise 17 different series, which are all given in the file `starostin.pdf` in the supplement. Eyeballing these results shows, however, that mixed series can be found, but that the hypothesis itself is not directly reflected in any subseries of the data. More thorough analyses will be needed in order to check to which degree (if at all) character structure also reflects Starostin's hypothesis regarding final *-r in Old Chinese.

5.4 Distinction of Middle-Chinese -p, -t, and -jH [Hypothesis Haudricourt]

Building on his explanation for the origin of the Vietnamese *hỏi-ngã* and *sắc-nặng* tones as arising from laryngeal finals, Haudricourt (1954) proposed that the departing tone (*qùshēng*) arose from a final *-s. The explanation for departing tone readings appearing in phonetic series with or rhyming with entering tone readings is the simplification of final clusters. In the case of velar finals the change in question is *-ks > *-s > *-H*, e.g. 各 *kak* < *k^{ʕ}ak, 賂 *luH* < *r^{ʕ}aks.[10] In the case of dental finals, the final does not disappear entirely, but leaves a trace into Middle Chinese as final -j. The change in question is *-ts > *-js > -jH, as shown in the phonetic series built on 毕 #0302, viz. 括 *khwat* < *k$^{w ʕ}$at, 活 *hwat*< *g$^{w ʕ}$at, 話 *hwaejH* < *g^{w}r^{ʕ}ats.

In the case of labial finals the early change *-ps > *-ts lead to a merger of original *-ps and *-ts. After this merger *-ts developed as immediately described (i.e. *-ts > *-js > *-jH*). The early date of the change *-ps and *-ts means that readings with final *-ts were available within a *xiéshēng* series when characters were still being coined. As a result phonetic series with final *-p sometimes include readings with final *-t, as is the case with the series built on 入 #0695: 入 *nyip* < *nup, 内 *nwojH* < *n^{ʕ}uts < *n^{ʕ}ups, 訥 *nwot* < *n^{ʕ}ut.

We wondered whether series such as this might distinguish subseries that consistently pointed either to final *-p or to final *-t. One could imagine for example that 入 used directly as a phonophoric indicated always *-p, but that 内 as a phonophoric indicated always *-t. Our search criteria were thus defined, as shown in Table 6. Our first condition were readings ending in *-p*, and the second condition were readings ending in *-t*. The filtering procedure would then select only those readings in which a *xiéshēng* series would show readings supporting both conditions.

Condition 1	$F_p :=$ `"p"`$in\texttt{MCH}_{final}$
Condition 2	$F_t :=$ `"t"`$in\texttt{MCH}_{final}$
Show Series	$F_p \in \texttt{XS} \cap F_t \in \texttt{XS}$

Table 6: Conditions for departing tone.

The results of this search comprise only 1 distinct series: the aforementioned case of 内 *nop*, which is given in the file `haudricourt.pdf` in the supplement. While it is possible that our search procedure is lacking, we have to conclude for the time being, that the development of the

10 A poem from the *Odes* (Ode 209) also demonstrates the power of this hypothesis to improve the interpretation of rhyming in early poetry. This poem has the rhyme words 踖 *tshjek* < *ts^{h}ak, 碩 *dzyek* < *dak, 炙 *tsyaeH* < *tak-s, 莫 *meak* < *mr^{ʕ}ak, 庶 *syoH* < *stak-s, 客 *khaek* < *k^{h}r^{ʕ}ak, 錯 *tshak* < *ts^{hʕ}ak, 度 *duH* < *l^{ʕ}ak-s, 獲 *hweak* < *mq^{wʕ}rak, 格 *kaek* < *kr^{ʕ}ak, 福 *pjuwk* < *pək, and 酢 *dzak* < *dz^{ʕ}ak.

departing tone is not directly reflected in the subseries of the Chinese characters.

5.5 Clusters of velars and laterals [Hypothesis Gabelentz]

At least since van der Gabelentz ([1881] 1953, 99) it has been noticed that some *xiéshēng* series include a mixture of readings with Middle Chinese initial velars (見 *k-*, 溪 *kh-*, 群 *g-*, or 匣 *h-*) and initial 來 *l-*. These cases are explained by reconstructing two syllable types in Old Chinese, one with a syllabic prefix **kə-* and an **r-* initial (corresponding to MCH 來 *l-*), and one with **Kr-* (corresponding to MCH *K-*). We wanted to know whether the subseries to Karlgren's *xiéshēng* series reflect this distinction. The conditions for our search are provided in Table 7. Our first condition checks for velar initials, and our second condition checks for the initial *l-*. The condition for filtering lists only those series in which both conditions occur.

The results of this search comprise 13 different series, which are all given in the file `gabelentz.pdf` in the supplement. While our results show that there are occasional *pure* subseries encoding only characters starting with *l-*, such as in series #0766, where 路 *luH*, as opposed to the mixed primary series 各 *kak*, these cases are very spurious. While we cannot exclude that velar-lateral clusters find their reflection in subseries of the *xiéshēng* series, it seems that this does not hold for the majority of cases.

Condition 1	$I_{velar} := \texttt{MCH}_{initial} \in$ `["k", "kh", "g", "x", "h"]`
Condition 2	$I_{lateral} := \texttt{MCH}_{initial} =$ `"l"`
Show Series	$I_{velar} \in \texttt{XS} \cap I_{lateral} \in \texttt{XS}$

Table 7: Conditions for velar-lateral clusters.

6 Discussion

Tatsuo Nishida proposes the useful notion of a *sonus grammae* (字音質／示音質), the phonology that is implied by a script system itself as analytically distinct from the phonology of a language that uses the particular script system at a certain time and place (Yabu 2014, 181, 187). The current paper may be seen as an attempt to identify further points on which the *sonus grammae* of *kǎishū* 楷書 can be identified with reconstructed Old Chinese. If *kǎishū* were well-designed for the purpose of writing Old Chinese (which it was not) then the *sonus grammae* of a given *kǎishū* character would closely predict the Old Chinese pronunciation of the linguistic form that this character writes. It is clear that the *sonus grammae* of *kǎishū* better matches the pronunciation of Chinese at the time of the coinage of these character forms than it does the pronunciations they find in spoken Mandarin today. Nevertheless, it would be useful to have an explicit metric of the distance between the *sonus grammae* and the sound of the language in a given period.

Scholars have so far emphasized in Old Chinese reconstruction, that they allow certain divergences in the pronunciation: the rhyme is usually assumed to be strictly reflected in the *xiéshēng* series, but the initials need only have the same place of articulation. However, given that the characters were coined at different times, in different places, and within different textual

communities, it is far from obvious that at the time they were created, the speakers who coined new characters had some principle of loose phonetic similarity in mind. We would rather assume a principle by which people tried to be exact as possible, unless pragmatic factors (the elegance of a given solution, or a certain association that not all speakers would follow) prevailed over the principle of choosing an exact phonophoric determinative for a new character. Reasons for pragmatic overrides are manifold, and they are historical, so they could in principle be stratified, i.e., one could find out, when a character was accepted as a valid *xiéshēng* for a given word, if one had a full account on their creation history.

While scholars think of divergence in absolute terms: every series can have voicing differences, for example, this does not need to be true in general, and it is more likely that we find high regions of regularity in certain series, and less regularity in less frequently used series. Finding out which distinctions are encoded in the selection of phonophoric determinative is crucial for a better understanding of Chinese character formation and for a better reconstruction of Old Chinese phonology.

7 Outlook

It is obvious that these case studies have only touched on the potential of network analyses of phonophoric relationships between characters to better understand Old Chinese phonology. One could use the approach presented here to search for additional distinctions which are not explicitly noted as such when relying on a loose grouping of *xiéshēng* characters into series. The analyses could also be used to enhance network approaches to rhyme analysis, following up on the work presented in List et al. (2017). Even more important, however, is the potential to impact scholarly discussion. If experts working on *xiéshēng* analyses began to provide their data in network form, listing direct phonophoric determinatives which reflect the derivational character of the Chinese writing system, it would be much easier to compare different analyses and to build on the research of our colleagues.

To achieve a better understanding and make better use of *xiéshēng* characters in the study of Chinese historical phonology, it will be necessary to have datasets that annotate the time (and where possible the place) of the character's first attestation and including a full set of character types (Oracle Bones, Bronze Inscriptions, Chǔ, etc.). For these data sets to appear, philologists and paleographers must add network methods into the repetoire of tools and techniques deployed in their craft.

Supplementary Material

The supplementary material accompanying this paper contains the source code and the data as well as additional information regarding the software required to run the analyses. The data is curated at Github (https://github.com/lingpy/character-formation-paper), and can be downloaded from Zenodo (https://zenodo.org/record/3246393).

Acknowledgments

This research would not have been possible without the LFK Young Scholars Symposium (Academia Sinica, Taipei, 2017), generously hosted by the Li Fang-Kuei Society for Chinese Linguistics (http://lfksociety.org/), during which both authors developed their approach. We would like to acknowledge the generous support of the European Research Council for supporting this research under the auspices of 'Beyond Boundaries: Religion, Region, Language and the State' (ERC Synergy Pro- ject 609823 ASIA, NWH) and 'Computer-Assisted Language Comparison' (715618, ERC Starting Grant, JML). We also thank Jonathan Smith for helpful comments on a previous version of this paper, as well as the two anonymous reviewers.

References

Baxter, William H. 1992. *A Handbook of Old Chinese Phonology*. Berlin: de Gruyter.

Baxter, William H., and Laurent Sagart. 2014. *Old Chinese. A New Reconstruction*. Oxford: Oxford University Press. http://ocbaxtersagart.lsait.lsa.umich.edu/BaxterSagartOC2015-10-13.xlsx.

Behr, Wolfgang. 1999. "Odds on the Odes." 1999.

Behr, Wolfgang. 2004. "Language Change in Premodern China: Notes on Its Perception and Impact on the Idea of a "Constant Way"." In *Historical Truth, Historical Criticism and Ideology: Chinese Historiography and Historical Culture from a New Comparative Perspective*, edited by Helwig Schmidt-Glintzer, Achim Mittag, and Jörn Rüsen, 13–51. Leiden; Boston: Brill.

Boltz, William G. 1994. *The Origin and Early Development of the Chinese Writing System*. New Haven: American Oriental Society.

Gabelentz, Georg v. d. (1881) 1953. *Chinesische Grammatik: Mit Ausschluss Des Niederen Stiles Und Der Heutigen Umgangssprache*. Reprint. Berlin: Deutscher Verlag der Wissenschaften.

Geyken, Alexander, ed. 2010. "Digitales Wörterbuch Der Deutschen Sprache DWDS. Das Wortauskunftssystem Zur Deutschen Sprache in Geschichte Und Gegenwart." Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. 2010. http://dwds.de.

Haudricourt, André-Georges. 1954. "Comment Reconstruire Le Chinois Archaïque." *Word* 10 (2-3): 351–64.

Hill, Nathan W. 2015. "Proposal for a Transcription of Chinese Characters in the Study of Early Chinese Language and Literature." *Bulletin of Chinese Linguistics* 8: 48–60.

Hill, Nathan W.. 2017. "The Evidence for Chinese *-R." *Bulletin of Chinese Linguistics* 9: 190–204.

Hippisley, Andrew. 1998. "Indexed Stems and Russian Word Formation: A Network Morphology Account of Russian Personal Nouns." *Linguistics Faculty Publications* 36 (6): 1093–1124. http://uknowledge.uky.edu/lin_facpub/43.

Karlgren, Bernhard. 1954. "Compendium of Phonetics in Ancient and Archaic Chinese." *Bulletin of the Museum of Far Eastern Antiquities* 26: 211–367.

Karlgren, Bernhard. 1957. "Grammata Serica Recensa." *Bulletin of the Museum of Far Eastern Antiquities* 29: 1–332.

Kawabata, Taichi. 2014. “Ideographic Description Sequence Data.” 2014. https://doi.org/https://doi.org/10.5281/zenodo.1181440.

Kunze, R. 1937. *Bau Und Anordnung Der Chinesischen Zeichen. Oder: Wie Lernen Wir Leichter Zeichen Lesen?* Tokyo: Deutsche Gesellschaft für Natur und Völkerkunde Ostasiens.

Li, Fang-kuei 李方桂. 1974. “Studies on Archaic Chinese.” Translated by Gilbert L. Mattos. *Monumenta Serica* 31: 219–87.

List, Johann-Mattis. 2008. “Rekonstruktion Der Aussprache Des Mittel- Und Altchinesischen: Vergleich Der Rekonstruktionsmethoden Der Indogermanischen Und Der Chinesischen Sprachwissenschaft.” Magister thesis, Berlin: Freie Universität Berlin.

List, Johann-Mattis. 2017. “Using Network Models to Analyze Old Chinese Rhyme Data.” *Bulletin of Chinese Linguistics* 9 (2): 218–41.

List, Johann-Mattis. 2018. “More on Network Approaches in Historical Chinese Phonology (音韵学).” In *The 2nd Li Fang-Kuei Society Young Scholars Symposium*, 157–74. Taipei: Li Fang-Kuei Society for Chinese Linguistics.

List, Johann-Mattis. 2018. SinoPy. A Python library for quantitative tasks in Chinese historical linguistics. Jena: Max Planck Institute for the Science of Human History. URL: https://github.com/lingpy/sinopy/. DOI: https://doi.org/10.5281/zenodo.1403028.

List, Johann-Mattis; Greenhill, Simon; Tresoldi, Tiago; and Forkel, Robert (2018): *LingPy. A Python library for historical linguistics*. Version 2.6.4. URL: http://lingpy.org, DOI: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy.

List, Johann-Mattis, Jananan Sylvestre Pathmanathan, Nathan W. Hill, Eric Bapteste, and Philippe Lopez. 2017. “Vowel Purity and Rhyme Evidence in Old Chinese Reconstruction.” *Lingua Sinica* 3 (1): 1–17.

List, Johann-Mattis, Jananan Sylvestre Pathmanathan, Philippe Lopez, and Eric Bapteste. 2016. “Unity and Disunity in Evolutionary Sciences: Process-Based Analogies Open Common Research Avenues for Biology and Linguistics.” *Biology Direct* 11 (39): 1–17.

Pān, Wùyún 潘悟云. 2000. *Hànyǔ Lìshǐ Yīnyùnxué*. Shànghǎi 上海: Shànghǎi Jiàoyù 上海教育.

Pfeifer, Wolfgang, ed. 1993. *Etymologisches Wörterbuch Des Deutschen*. 2nd ed. 2 vols. Berlin: Akademie.

Qiú, Xīguī 裘錫圭. (1988) 2007. *Wénzìxué Gàiyào*. Běijīng: Shāngwù 商务.

Sagart, Laurent. 1999. *The Roots of Old Chinese*. Amsterdam: John Benjamins.

Sagart, Laurent, and William Baxter. 2012. “Reconstructing the *S- Prefix in Old Chiense.” *Language and Linguistics*, nos. 13, 1: 29–59.

Sagart, Laurent, and William H. Baxter. 2009. “Reconstructing Old Chinese Uvulars in the Baxter-Sagart System (Version 0.99).” *Cahiers de Linguistique – Asie Orientale* 38 (2): 221–44.

Sagart, Laurent, and Mǎ Kūn 馬坤. 2017. “Xiānqín Shíqī Xiéshēng Shēngfú de Xuǎnzé Wèntí.” University of Macau; talkatm. 2017. http://www.academia.edu/35852895/.

Schuessler, Axel, ed. 2007. *ABC Etymological Dictionary of Old Chinese*. Honolulu: University of Hawai’i Press.

Schuessler, Axel. 2009. *Minimal Old Chinese and Later Han Chinese. A Companion to Grammata Serica*. Honolulu: University of Hawai’i Press.

Ševčíková, Magda, and Zdeněk Žabokrtský. 2014. “Word-Formation Network for Czech.” Edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, and Joseph Mariani. Reykjavík: European Language Resources Association. 2014.

Shen, Ruiqing. 2017. “Děng 等 (Division and Rank).” In *Encyclopedia of Chinese Language and Linguistics*, edited by Rint Sybesma, 2:13–20. Leiden; Boston: Brill.

Starostin, Sergej Anatol’evič. 1989. *Rekonstrukcija Drevnekitajskoj Fonologičeskoj Sistemy (Reconstruction of the Phonological System of Old Chinese)*. Moscow: Nauka.

Wikipedia Users. 2017. “Character List for Karlgren’s GSR.” 2017. https://en.wikibooks.org/w/index.php?title=Character_List_for_Karlgren%27s_GSR&oldid=3229525.

Yabu, Shirō 藪 司郎. 2014. “Professor Nishida, Tatsuo and the Study of Tibeto-Burman Languages.” *Memoirs of the Research Department of the Toyo Bunko* 72: 179–205.

Zhèngzhāng, Shàngfāng 郑张尚芳. 2003. *Shànggǔ Yīnxì*. Shànghǎi 上海: Shànghǎi Jiàoyù 上海教育.