Chinese Computational Linguistics

Johann-Mattis List (University of Passau)

1 Background

There are many myths regarding the Chinese language or the Chinese language**s**, and without having had proper insights into both the language structure and the writing system that is used to write the language, it may be difficult to assert whether something belongs to the realm of myths or the realm of known facts. There is not enough time to discuss all myths and facts in one session, but we will try to quickly look at the grammatical structure of Chinese and at the role that dialectal variation play in the context of the language.

What myths about Chinese do you know?

Grammatical Structure of Chinese

As a language, the major characteristics of Chinese are its *isolating* structure, reflected in the quote below by the Sergey Yakhontov:

> In Chinese, grammatical relations among words in a sentence are expressed by word order or by the use of specific function words, for example, prepositions, but not by modifying the word forms themselves. (Jachontov 1965: 12 [1])

What is meant by this structure can be easily seen when inspecting examples in interlinear-glossed text.

(1)

我	爸爸	不	在
wǒ	bàba	bú	zài
I	father	not	be present

„My father is not here."

(2)

我	会	告诉	他	的
wǒ	hùi	gàosu	tā	de
I	function verb	tell	he	*particle*

„I shall tell him."

(3)

我	以前	在	柏林	学习
wǒ	yǐqián	zài	Bólín	xuéxí
I	earlier times	be present	Berlin	study

„I used to study in Berlin."

What is remarkable with respect to parts of speech in the second example sentence above?

[1] My translation, original text: «В китайском языке грамматические отношения между словами в предложении выражаются порядком их расположения, а также специальными служебными словами, например предлогами, но не изменением формы слов.»

Linguistic Variation

According to the official definition of *pǔtōnghuà* 普通话, the variety often called *Mandarin Chinese* takes the variety of Běijīng as its phonetic basis (*yǐ Běijīng yǔyīn wéi biāozhǔn* 以北京语音为标准), while following the classical texts composed in *Báihuà* 白话 (*yǐ diǎnfàn de báihuàwén zhùzuò wéi yǔfǎ guīfàn* 以典范的白话文著作为语法规范) grammatically (Huáng and Liào 2002: 4). In practice, however, the Standard Chinese is spoken with many flavors by the multitude of people who speak different dialects of Chinese as their first language. In the last decades, we can see a rise of proficiency in Standard Chinese among younger people, accompanied by a drastic loss of dialectal variation. Nevertheless, it is problematic to speak of Chinese as one single language without knowing about the specific sociolinguistic situation in which this "language" is realized. In practice, however, linguists still tend to talk about the Chinese language as a certain kind of unity. The reason that justify to treat Chinese with all its Sinitic varieties as one unit is the sociolinguistic situation in which many distinct varieties share a common history, a common writing system, and a common "roof language" that is used to communicate across the individual dialectal varieties.

What differences and similarities can we find in the sociolinguistic context of language in China with the sociolinguistic context of language in Europe?

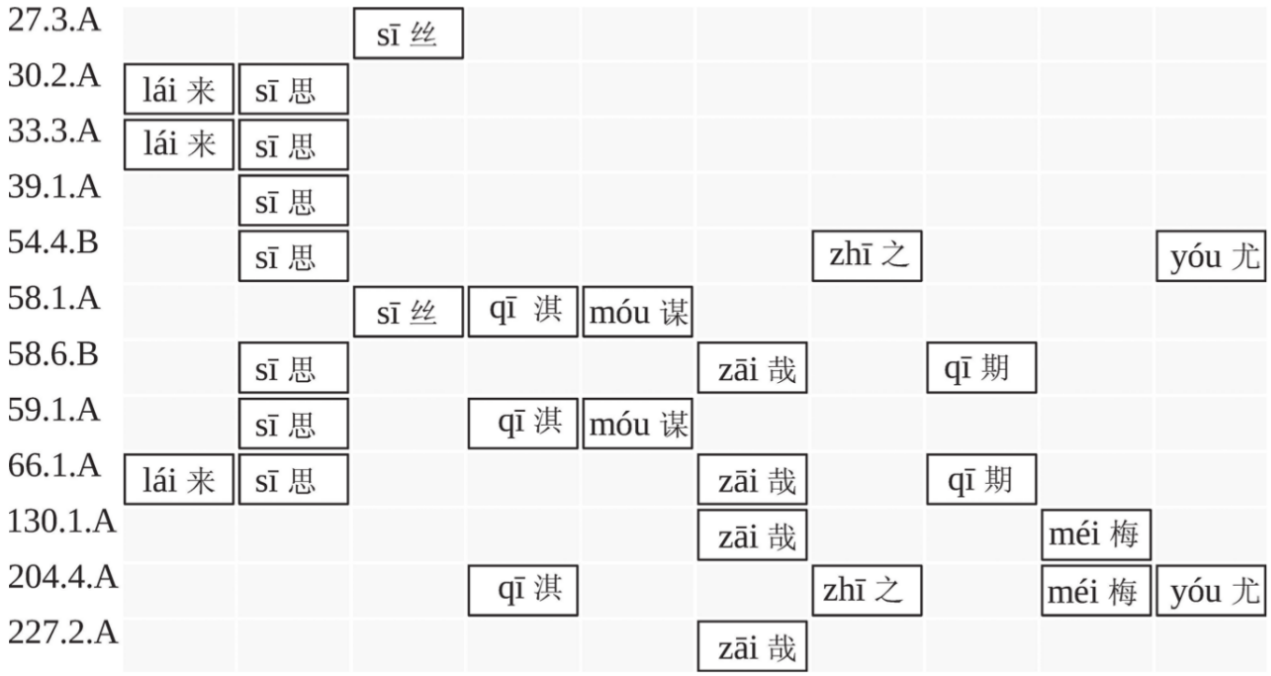
2 Rhyme Analysis

The analysis of rhyme patterns is one of the core methods for the reconstruction of Old Chinese phonology. It emerged when scholars of the Suí 隋 (581–618) and Táng 唐 (618–907) dynasties realized that old poems, especially those in the Book of Odes (Shījīng 詩經 ca. 1050–600 BCE), were full of inconsistencies regarding the rhyming of words. While the first reaction was to attribute inconsistencies to a different, less strict attitude towards rhyming practiced by the ancestors (as advocated by Lù Démíng 陸德明, 550–630), or to a habit of the elders to switch the pronunciation in certain words in order to make them rhyme (a practice called *xiéyīn* 諧音 'sound harmonization', Baxter 1992: 153). Later scholars from the Míng 明 (1368–1644) and Qīng 清 dynasties (1644–1911) realized that the inconsistencies in the rhyme patterns reflect the effects of language change (ibid.: 153-157). This is illustrated in Table 1.

Chinese Text	Translation	RW	Patterns	MCH	OCBS-Rhyme
燕燕於飛	The swallows go flying	*fēi* 飛	A	*pjɨj	*-ər
下上其音	falling and rising are their voices;	*yīn* 音	B	*ʔim	*-əm
之子於歸	This young lady goes to her new home,	*guī* 歸	A	*kjwɨj	*-əj
遠送於南	far I accompany her to the south.	*nán* 南	B	*nom	*-əm
瞻望弗及	I gaze after her, can no longer see her,	[*jí* 及]	–	[**gip*]	[**-əp*]
實勞我心	truly it grieves my heart	*xīn* 心	B	*sim	*-əm

Assuming that rhyming was originally rather consistent, with rhyme words being mostly identical in the pronunciation of nucleus and coda, the analysis of rhyme words makes it not only possible to estab- lish rhyme categories but also to interpret them further phonetically or phonologically. The classical approach for rhyme analysis, which is called *sīguàn shéngqiān fǎ* 絲貫繩牽法 'link-and-bind method' (Gēng 2004), or *yùnjiǎo xìlián fǎ* 韻腳系聯法 'rhyme linking method' (Lǚ 2009), consists of roughly two steps: In a first step, groups of Old Chinese words, mostly represented by one Chinese

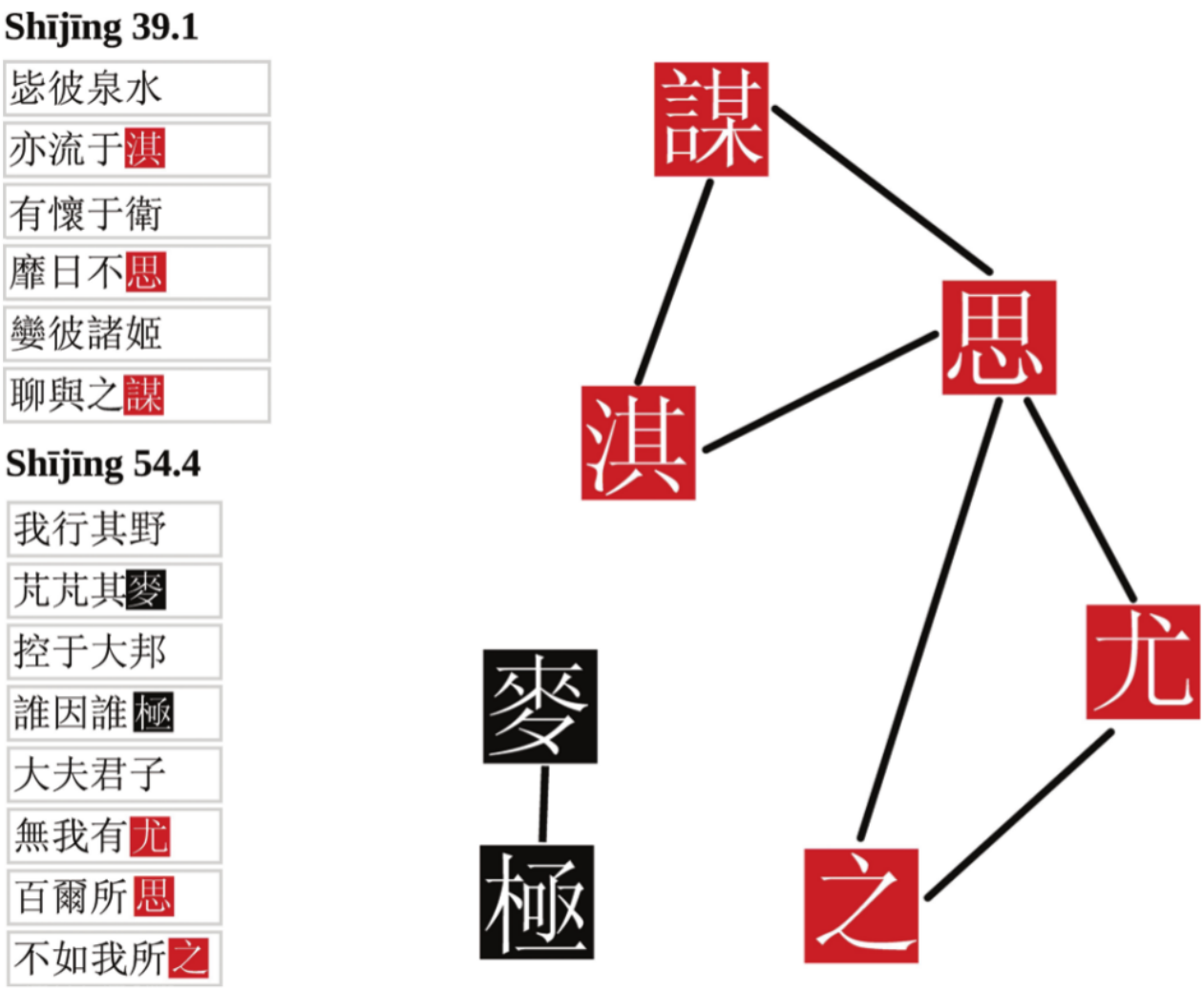
character and identified to rhyme with each other in a given text are collected. In a further step, these groups are com- pared with each other. If identical words are found in different groups, those groups can be combined to form larger groups. This procedure is then repeated until categories of rhymes can be identified that ide- ally do not show any more transitions among each other. This approach is essentially similar to the 'linking method' *xìlián fǎ* 系聯法 see Liú 2006: 56-67), first proposed in Chén Lǐ's 陳禮 (1818–1882) Qièyùnkǎo 切韻考 (1848), by which characters used in *fǎnqiè* 反切 readings in rhyme books are clustered into groups of supposedly common pronunciations for initials and rhymes. In both approaches, similarities in pronunciation are indirectly inferred by spinning a web of direct links between characters.



The figure above illustrates the linking method for the zhī 之 group in the Book of Odes. What is the obvious drawback of this method?

Network Approach to Rhyme Analysis

The crucial idea of our computer-assisted approach to rhyme analysis is to construct a *network of rhyme patterns* in which nodes represent rhyme words and connections between nodes represent how often those rhymes co-occur in the Book of Odes. The following graphic illustrates this procedure for two stanzas of the Shījīng:



The major advantage of this representation is that we can apply various methods for network analysis to data which was assembled in this form. As a result, we can investigate the rhyme network and test to which degree different reconstruction systems offer a consistent view on Old Chinese rhyming. As a very simple test, we can check whether a given reconstruction system conforms to the principle of *vowel purity* (Ho 2016) which expects words with similar vowels to rhyme more often than words with different vowels. Our test, which is reported in List et al. (2017) could show that most of the Old Chinese reconstruction systems which postulate 6 vowels correspond more closely to vowel purity than other reconstruction systems with more or less vowels. Even by eyeballing the figure above, in which vowel quality is reflected with help of colors following the OC reconstruction system by Baxter and Sagart (2014), one can see that words rhyming with each other tend to have the same vowel.

If six-vowel reconstruction systems perform better on vowel purity, does this automatically mean that they are better in general?

The Shījīng Rhyme Browser

In order to make it more convenient for the readers to investigate the data underlying this paper in full detail, an interactive web-based application was created. This freely available Shījīng Browser (`http://digling.org/shijing/`) lists all potential rhyme words in tabular form along with additional information including the *pīnyīn* transliteration, the Middle Chinese reading, the reconstruction by

Baxter and Sagart (ibid.), the reading by Pān (2000), the GSR index (Karlgren 1957), and the number of poem, stanza, and section. With help of interactive search fields, the data can quickly be filtered, enabling the users to search for specific poems, for specific characters, or for specific readings. When clicking on the "Poem" field in the application, a window pops up and shows the whole poem, in which all rhyme words are highlighted. In certain cases, where potential alternative rhymes were identified, this is marked in an additional column. In a recently modified version, we contrast rhyme annotations by Wáng (1980 [2006]) with those given in Baxter (1992) (`http://digling.org/shijing/wangli/`, List 2017). The table below gives an example on the organization of the interface.

Text	Stanza	MCH	Pān Wúyùn	OCBS	Wáng Lì	Starostin	Rhyme	Group
遵彼汝墳,伐其條枚	1.AB	mwoj	mɯul	mˁəj	muəi	məj	A	微
未見君子,怒如調飢	1.CD	tsiX	kril	Cə.kə[j]	kiei	krəj	A	脂
遵彼汝墳,伐其條肄	2.AB	sijH	ph-ljɯds	s-ləp-s	jiet	slhəps	B	质
既見君子,不我遐棄	2.CD	khjijH	khids	[kʰ]i[t]-s	khiet	khijs	B	质
魴魚赬尾	3.A	mj+jX	mɯlʔ	[m]əjʔ	miuəi	məjʔ	C	微
王室如燬	3.B	xjweX	qhʷralʔ	[m̥](r)ajʔ	xiuəi	hʷejʔ	C	微
雖則如燬	3.C	xjweX	qhʷralʔ	[m̥](r)ajʔ	xiuəi	hʷejʔ	C	微
父母孔邇	3.D	nyeX	mljelʔ	n[ə][r]ʔ	njiei	n(h)ejʔ	C	脂

What could be the problem of comparing rhymes in books other than the Book of Odes?

3 Character Analysis

The Chinese writing system, as we know it today, is famous for its structural properties reflected by a complicated interaction of phonetic and semantic elements.

Chinese characters can be divided into elements carrying phonetic function and in elements carrying semantic functions. As a result, scholars tend to call it a "semanto-phonetic writing system" (*yìyī wénzì* 意音文字, cf. Zhōu 1998: 60. But this characterization exaggerates the potential of character elements to *predict* a certain pronunciation or meaning.

> Most of the "phonetic" characteristics of the [Chinese writing system] are relics of the processes of character formation which, as they took place asynchronously, were always characterized by a complex interaction between the Chinese language spoken at different times of its history, the socio-cultural background of those people who created the characters, and general patterns of reasoning and conceptualization. (List et al. 2016: 49)

Thus, although people often *say* that the Chinese characters have phonetic and semantic characteristics, from the perspective of their potential, these aspects are very limited, since the predictive force is extremely limited. The reason lies in the fact that the Chinese writing system has been derived at different stages, similar to the way in which the lexicon of human languages shows different layers of transparency and opacity with respect to the motivation of individual lexemes. What can be said for sure is that – in the majority of all cases – one Chinese character expresses one morpheme of the Chinese language and that the morpheme is usually one syllable in length.

Below is a quote that defines *motivation* in linguistics. What are the reasons that motivation is lost in the lexicon of human languages, and what consequences would this have for the Chinese writing system?

> **Motivation**: Extent, to which the [complex word] can be understood as the result of its parts and their composition. (Glück 2000: s.v. "Motivation")[2]

Phonetic Elements in Chinese Writing

Chinese characters were developed over millennia and their formation (*zàozìfǎ* 造字法, Qiú 1988 [2007]) is best seen as a derivational process with striking similarities to word formation processes (Kunze 1937, List 2008).

> 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) [...]. 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 *'jɨj* 'clothes'. (Hill and List 2019: 186)

Chinese characters thus show some degree of *recursion*: phonetic elements can themselves consist of complex characters and contain formerly transparent semantic and phonetic elements.

Is recursion the correct term to describe the derived character of phonetic elements in the Chinese writing system?

Network Analysis of Phonetic Elements

The parallel between word formation and Chinese writing can be used as a source of inspiration for the modeling of Chinese character formation processes.

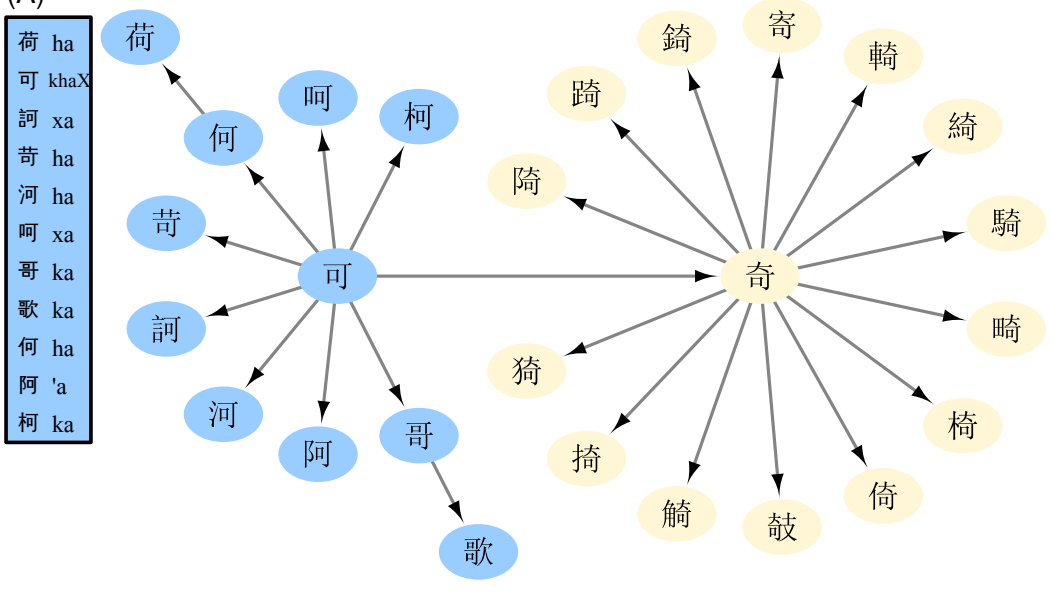
> A crucial aspect of word formation (and also of character formation) is the hierarchical process by which words are derived from each other at different times. If we have a compound word, like German Krankheitsverlauf 'disease progression', we can recursively split the word into its respective components which usually were coined at different moments in history. (ibid.: 190)

These components can be modeled with the help of directed networks. The benefits of this approach is that the organization of phonetic elements in the Chinese writing system can be made much more transparent than it has been done in previous work, where scholars would assign individual characters to monolithic clusters (Karlgren 1957). In our pilot study (Hill and List 2019), we show the benefits of this approach by testing concrete hypotheses on the pronunciation of specific characters in Old Chinese.

How can the different phonetic realizations of the yellow characters in the table (B) be explained linguistically?

[2] My translation, original text: "**Motivation**: Ausmaß, in dem [das komplexe Wort] sich als Summe seiner Teile und der Weise ihrer Zusammenfügung verstehen lässt".

(A)



(B)

倚	'jeX
攲	khje, kje
觭	khje
奇	'kje, gje
跨	khje, kjeX, ngjeX
錡	gje, gjeX, ngjeX
寄	kjeH
綺	khjeX
畸	kje
騎	gje
猗	'aX, 'je, 'jeX
輢	'jeX
椅	'je
倚	'je
掎	kjeX

4 Further Topics

The field of Chinese Historical Phonology is full of interesting topics that have not yet been sufficiently analyzed with the help of computational methods. While scholars have always applied some quantitative approaches, counting occurrences of examples in texts, or making quantitative analyses on sheets of paper, discussing them in prose, a rigorously formal treatment of many topics is still missing. In a forthcoming study (List forthcoming), I discuss three topics that could further advance the field of Computational Chinese Historical Phonology in the future, namely *corpus studies*, specifically studies that inspect the description of character pronunciations in Chinese texts, *additional network approaches*, specifically those that target the description of meanings in tags, as it can be frequently found in Chinese literature, and finally *alignment analyses* that could help to make the transcription of foreign names with the help of Chinese characters more transparent.

The figure below shows an example for the alignment of transcriptions of Buddhist terms in Chinese. So far, these transcriptions have only been analyzed in a manual fashion. Which chances and which challenges do alignment analyses bear for the investigation of transcriptions and transliterations and what alternative methods could be used?

ID	DOCULECT	CONCEPT	FORM	CHARACTERS	COGIDS	ALIGNMENT
3	An Shigao	āgama	āgama		2 3	- a: + g a m a
4	An Shigao	āgama	ʔa gəm	阿 含	2 3	? a + g ə m -
5	An Shigao	ānanda	ānanda		2 4	- a: + n a n d a
6	An Shigao	ānanda	ʔa nan	阿 難	2 4	? a + n a n - -
1	An Shigao	Buddha	buddha		1	b u d dh a
2	An Shigao	Buddha	bjət	佛	1	bʲ ə t - -

References

Baxter, W. H. (1992). *A handbook of Old Chinese phonology*. Berlin: de Gruyter.

Baxter, W. H. and L. Sagart (2014). *Old Chinese. A new reconstruction*. Oxford: Oxford University Press.

Glück, H., ed. (2000). *Metzler-Lexikon Sprache*. 2nd ed. Stuttgart: Metzler.

Gěng, Z. 耿振生. (2004). *20 shìjì Hànyǔ yǔyīnxué fāngfǎ lùn* 20世纪汉语音韵学方法论 [20th centuryś methods in traditional Chinese phonology]. Běijīng 北京: Běijīng Dàxué 北京大學.

Hill, N. W. and J.-M. List (2019). "Using Chinese character formation graphs to test proposals in Chinese historical phonology." *Bulletin of Chinese Linguistics* 12.2, 186–200.

Ho, D.-a. (2016). "Such errors could have been avoided. Review of "Old Chinese: A new reconstruction". by William H. Baxter and Laurent Sagart." *Journal of Chinese Linguistics* 44.1, 175–230.

Huáng, B. and X. Liào (2002). *Xiàndài Hànyǔ* 现代汉语 [Modern Chinese]. 3rd ed. Vol. 1. 2 vols. Běijīng: Gāoděng Jiàoyù.

Jachontov, S. E. (1965). *Drevnekitajskij jazyk* [Old Chinese]. Moscow: Nauka.

Karlgren, B. (1957). "Grammata serica recensa." *Bulletin of the Museum of Far Eastern Antiquities* 29, 1–332.

Kunze, R. (1937). *Bau und Anordnung der chinesischen Zeichen. Oder: Wie lernen wir leichter Zeichen lesen?* [Structure and assembly of Chinese characters. Or: How can we learn to read characters more easily?] Tokyo: Deutsche Gesellschaft für Natur und Völkerkunde Ostasiens.

List, J.-M. (2008). "Rekonstruktion der Aussprache des Mittel- und Altchinesischen. Vergleich der Rekonstruktionsmethoden der indogermanischen und der chinesischen Sprachwissenschaft [Reconstruction of the pronunciation of Middle and Old Chinese. Comparison of reconstruction methods in Indo-European and Chinese linguistics]." Magister thesis. Berlin: Freie Universität Berlin. PDF: `http://hprints.org/docs/00/74/25/52/PDF/list-2008-magisterarbeit.pdf`.

— (2017). *Vertikale und laterale Aspekte der chinesischen Dialektgeschichte* [Vertical and lateral aspects of Chinese dialect history]. Jena: Max Planck Institute for the Science of Human History.

— (forthcoming). "Chances and challenges for quantitative approaches in Chinese Historical Phonology." *Bulletin of Chinese Linguistics* 0.0, 1–19.

List, J.-M., J. S. Pathmanathan, N. W. Hill, E. Bapteste, and P. Lopez (2017). "Vowel purity and rhyme evidence in Old Chinese reconstruction." *Lingua Sinica* 3.1, 1–17.

List, J.-M., A. Terhalle, and D. Schulzek (2016). "Traces of embodiment in Chinese character formation. A frame approach to the interaction of writing, speaking, and meaning." In: *Sensory-motor concepts. At the crossroad between language & cognition*. Ed. by L. Ströbel. Düsseldorf: Düsseldorf University Press, 45–62.

Liú, X. 刘晓南. (2006). *Hànyǔ yīnyùn yánjiū jiàochéng* 汉语音韵研究教程 [Reader in traditional Chinese phonology]. Běijīng 北京: Běijīng Dàxué 北京大學.

Lǚ, S. 吕胜男. (2009). "A brief study of the methodology of the study of ancient rhyme. And Concurrently on the study of the rhyme of "Jinwen Shangshu" 古韵研究方法论发微. 兼论今文《尚书》用韵研究 [History of ancient Chinese linguistics]." *Nányáng Shīfàn Dàxué Bào (Shèhuì Kēxué Bǎn)* 南阳师范学院学报(社会科学版) *[Journal of Nanyang Normal University (Social Sciences)]* 8.2, 57–61.

Pān, W. 潘悟云. (2000). *Hànyǔ lìshǐ yīnyùnxué* 汉语历史音韵学 [Chinese historical phonology]. Shànghǎi 上海: Shànghǎi Jiàoyù 上海教育.

Qiú, X. 裘錫圭. (1988 [2007]). *Wénzìxué gàiyào* 文字學概要 [Foundations of graphemics]. Běijīng: Shāngwù 商务.

Wáng, L. 王力. (1980 [2006]). *Hànyǔ shǐgǎo* 漢語史稿 [History of the Chinese language]. Repr. Běijīng 北京: Zhōnghuá Shūjú 中华书局.

Zhōu, Y. 周有光. (1998). *Bǐjiào Wénzìxué Chūtàn* 比较文字学初探 [Introductory investigations on the comparison of writing systems]. Yǔwén 語文.