Graph model of Old Chinese phonological system and computing

Jiajia Hu
School of Computer Science and Engineering, Beihang University, China

Ning Wang
School of Chinese Language and Literature, Beijing Normal University, China

Abstract

Old Chinese phonology is considered one of the most difficult branches of learning in Chinese philology. The traditional methodology has been empirical and its conclusions are very hard to understand and remember for modern people who do not use Old Chinese any more. This article, based on the theories and findings in Old Chinese phonology studies, uses the graph model in mathematics as the method to describe the initial-rhyme relationships of Old Chinese, and the Euclidean distance on two-dimensional plane to represent the similarity of the sounds in Chinese words (or characters) in the phonological harmony. It provides a more intuitive method of quantitative analysis, which makes the qualitative phonological relationships of Old Chinese easier to be measured and compared. Although the model is still in its infancy and there is room for refinement, it may bring insights to Chinese study, such as the research of Xiesheng series in Chinese characters, or Chinese etymology.

Correspondence:
Jiajia Hu, School of Computer Science and Engineering, Beihang University, China.
E-mail:
hjj81@126.com

1 Introduction

Old Chinese, or shanggu hanyu in Chinese, refers to the Chinese language in the pre-Han period (or the Zhou and Qin dynasties). The research of Old Chinese phonology was once an important part of traditional Chinese philology. It was developed from the original practical purpose of interpreting early Chinese classics written during the early and mid Zhou Dynasty. From those texts, traditional philologists found that the speech sounds of Chinese characters had changed with times. They began to analyze those changes and tried to explain their reasons and find the laws that gradually built up the science of Old Chinese phonology.

Old Chinese phonology focuses on syllable studies (Baxter, 1992). In general, a syllable is divided into three parts (the various dialects and historical stages of Chinese are similar enough in syllable structure; Fig. 1): the initial (shengmu), the final (yunmu), and the tone (shengdiao). The final can be subdivided into the media (jieyin or yuntou), the main vowel (yuanfu) and the coda (yunwei). Since the medial is normally assumed to have no effect on rhyming, the main vowel and the coda are sometimes grouped together as the rhyme.

There are two aspects of Old Chinese phonology research (Wang, 2007). One is the reconstruction of the phonological system, or the division of its initial, rhyme and tone categories, and the other is to assign

Literary and Linguistic Computing, Vol. 27, No. 2, 2012.

doi:10.1093/llc/fqs001 Advance Access published on 27 February 2012

phonetic values to these categories, or simulate the pronunciation of these categories. The simulation, of course, is based on the categories, and should reflect the relationships among different categories. Once we know all the categories of Old Chinese phonology, and the exact categories each word belongs to, we can figure out the sound relationships among words and thus the relationships of their meanings that can finally help us understand those classical words written in Old Chinese and see the original purpose of phonology research.

2 Old Chinese Phonology Study

Most Chinese characters consist of two parts: a sense token and a phonetic token. The phonetic token is a character originally similar in sound to the word represented by the compound character. The sense token is a character used for its semantic value to distinguish one compound from another that has the same phonetic token. For example, the character 梅 (mei) consists of the phonetic token 每 (mei) and the sense token 木 (mu). The phonetic token 每 (mei) was chosen for its sound similarity to 梅 (mei), while the sense token 木 (mu) suggests the meaning of tree, and hence distinguishes the character for tree from other characters also with 每 (mei) as the phonetic token, such as 海 (hai). The set of characters with the same phonetic token is called a Xiesheng series, namely, one of the main sources of information about Old Chinese sounds, since many Xiesheng characters were created during the Old Chinese period. In this example, 梅 (mei) and 每 (mei) are still similar in modern sound; but in other cases, such as 海 (hai)and 每 (mei), words in the same Xiesheng series may, for sound changes, show little sound resemblance in modern pronunciation.

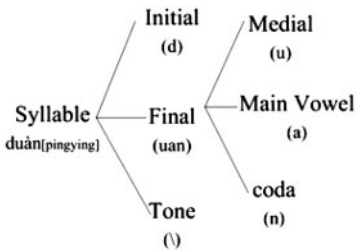

Fig. 1 Syllable structure in Chinese

Tones do not play much role in the relationship of sounds, so once we have the categories of initials and rhymes of characters, we can judge whether the sounds of two different words (or characters) are the same or similar. Therefore, if two words (or characters) have the same initial and the same rhyme, then we can say that they have the same sound. However, objective criteria for sound similarity among words (or characters) are hard to determine. That is, the similarity between sounds is relative, and how to judge the similarity level, different scholars have different approaches. Traditionally, the same or similar relationships between initials are called 'Shuangsheng', and those between rhymes 'Dieyun'. Traditional phonologists often use tables to illustrate the similarity levels, the closer their positions, the more similar their sounds.

As Table 1 shows, (1) the initials from the same points of articulation (in the same row) are called Pangniu, like 帮, 滂, 並, and 明. (2) the initials of throat sounds and the initials of the back-tooth sounds are called Houya-Shuangheng, like 影 and 见, 匣 and 疑. (3) the initials of tongue sounds and the initials of the tooth-head sounds are called Shechi-Shuangsheng, like 端 and 精, 定 and 清.

Table 1 Nineteen categories of Old Chinese initials (Huang, 1934)

Pronouncing Positions	Pronouncing Methods						
Laryngeals (throat sounds)	影	晓	匣				
Velars (back-tooth sounds)	见	溪			疑		
Linguals (tongue sounds)	端	透	定	泥	来		
Dentals (tooth-head sounds)	精	清	从				心
Labias (lip sounds)	帮	滂	並	明			

Table 2 Twenty-eight categories of Old Chinese rhymes (Huang, 1934)

Main vowel / coda's points of ariticulation	e	a	O	u	Ô	ə	Coda
Zero/back codas	齐 e	模 a	侯 O	萧 u	豪 Ô	咍 ə	
	锡 ek	铎 ak	屋 Ok		沃 Ôk	德 ək	-k
	青 eŋ	唐 aŋ	东 Oŋ	冬 Ôŋ	冬 Ôŋ	登 əŋ	-ŋ
Lingual codas		歌 ai	灰 Oi				-i
	屑 et	曷 at	没 Ot				-t
	先 en	寒 an	痕 On				-n
Labial codas							
		帖 ap	合 Op				-p
		添 am	覃 Om				-m

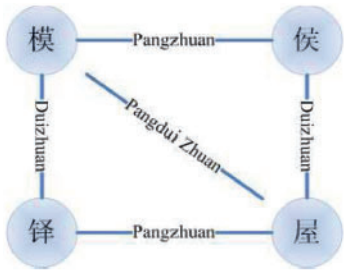

Fig. 2 Example of Pangdui-Zhuan

As Table 2 shows, (1) the rhymes having the same main vowel (in the same column) with codas from the same points of articulation are called Duizhuan, like 齐, 锡, and 青; or 歌, 曷, and 寒. (2) the rhymes having the same main vowel (in the same column) with codas from different points of articulation are called Ci-Duizhuan, like 青 and 先, 模 and 歌. (3) the rhymes having very similar main vowels with the same coda (next to each other in the same row) are called Pangzhuan, like 侯 and 萧, 齐 and 咍. (4) the rhymes having similar main vowels with the same coda (not next to each other in the same row) are called Ci-Pangzhuan, like 齐 and 侯, 模 and 豪. (5) the rhymes having similar main vowels with codas from the same points of articulation are called Pangdui-Zhuan, like 齐 and 铎, 歌 and 没, and it may look like the original rhyme that first its main vowels changed, and then its codas became the target rhyme, like 模 and 屋 in Fig. 2.

3 Graph Model of Old Chinese Phonological System

If one views every initial or rhyme category as an entity, and the relationships between them as their links, then the system of Old Chinese initials or rhymes could be viewed as a graph with the initial or rhyme categories as its vertices connected by edges that represent their relationships. As the relationships between sounds are symmetric, the edges are undirected. Different relationships reflect different similarity levels between sounds, so, if we use proper values to represent the relative distances between sounds, the system of Old Chinese initials or rhymes would be an undirected weighted graph.

Given an undirected weighted graph, we then know its weighted adjacency matrix. With the calculation of matrix (like the Dijkstra algorithm), we also know whether there exists a path between two vertices and the length of the shortest path between two vertices. As a result, by constructing the weighted adjacency matrix of initial or rhyme system, we obtain the shortest distance between the two sounds. Generally, the less the distance is, the closer the relationships would be.

3.1 Graph model of Old Chinese initial system

The system of Old Chinese initials is an undirected weighted graph $G_S = (V_S, E_S, W_S)$ (Fig. 3).

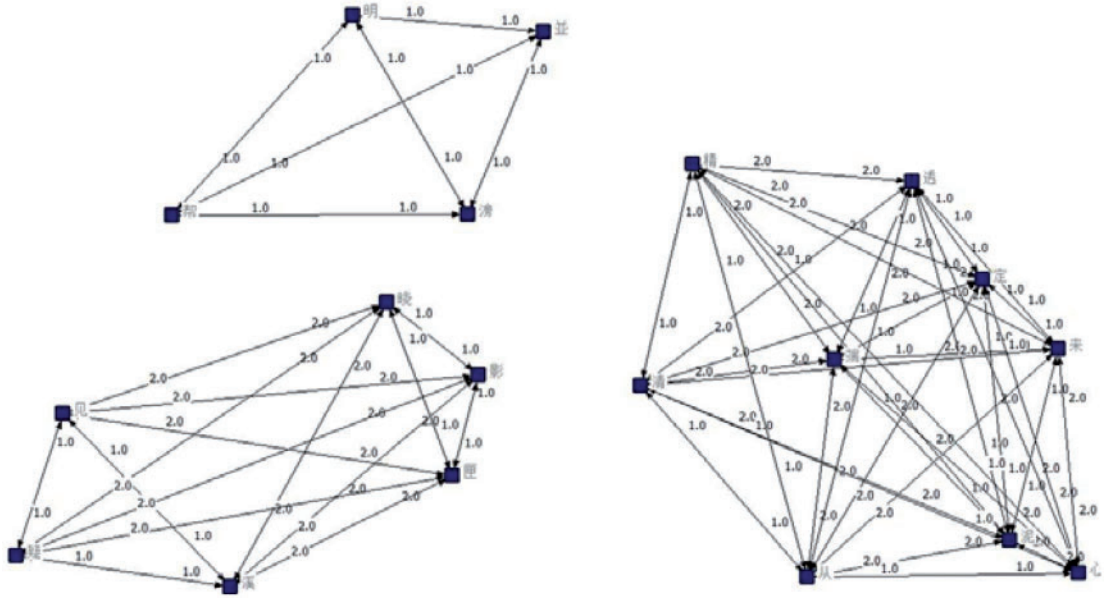

Fig. 3 Undirected weighted graph of Old Chinese initials.

(1) V_S is the set of vertices: the elements are the Old Chinese initial categories. According to Table 1, V_S = {影, 晓, 匣, 见, 溪, 疑, 端, 透, 定, 泥, 来, 精, 清, 从, 心, 帮, 滂, 並, 明};
(2) E_S is the set of edges: the element $e = (v_i, v_j)$ represents the relationship between initials v_i and v_j (Table 3, where A represents Shuangsheng, B Pangniu, C Houya-Shuangsheng, and D Shechi-Shuangsheng) and
(3) W_S is the weighted adjacency matrix (Table 4).

$$W_S = \begin{cases} 0: & e(v_i, v_j) = A \\ 1: & e(v_i, v_j) = B \\ 2: & e(v_i, v_j) = C/D \end{cases}$$

3.2 Graph model of Old Chinese rhyme system

The system of Old Chinese vowels is also an undirected weighted graph $G_Y = (V_Y, E_Y, W_Y)$ (Fig. 4).

(1) V_Y is the set of vertices: the elements are the Old Chinese rhyme categories. According to Table 2, V_Y = {歌, 曷, 寒, 灰, 没, 痕, 屑, 先, 齐, 锡, 青, 模, 铎, 唐, 侯, 屋, 东, 萧, 豪, 沃, 冬, 咍, 德, 登, 合, 覃, 帖, 添};
(2) E_Y is the set of edges: the element, $e = (v_i, v_j)$ represents the relationship between rhymes v_i and v_j (Table 5, where A represents Dieyun, B Duizhuan, C Pangzhuan, D Ci-Pangzhuan, and E Ci-Duizhuan) and
(3) W_Y is the weighted adjacency matrix (Table 6).

$$W_Y = \begin{cases} 0: & e(v_i, v_j) = A \\ 1: & e(v_i, v_j) = B \\ 2: & e(v_i, v_j) = C \\ 3: & e(v_i, v_j) = D \\ 4: & e(v_i, v_j) = E \end{cases}$$

3.3 Sound distance between syllables

In Old Chinese, each syllable is analyzed as an initial plus a rhyme with a tone. So the distance between

Table 3 Relationship adjacency matrix of Old Chinese initials

	影	晓	匣	见	溪	疑	端	透	定	泥	来	精	清	从	心	帮	滂	並	明
影	A	B	B	C	C	C													
晓	B	A	B	C	C	C													
匣	B	B	A	C	C	C													
见	C	C	C	A	B	B													
溪	C	C	C	B	A	B													
疑	C	C	C	B	B	A													
端							A	B	B	B	B	D	D	D	D				
透							B	A	B	B	B	D	D	D	D				
定							B	B	A	B	B	D	D	D	D				
泥							B	B	B	A	B	D	D	D	D				
来							B	B	B	B	A	D	D	D	D				
精							D	D	D	D	D	A	B	B	B				
清							D	D	D	D	D	B	A	B	B				
从							D	D	D	D	D	B	B	A	B				
心							D	D	D	D	D	B	B	B	A				
帮																A	B	B	B
滂																B	A	B	B
並																B	B	A	B
明																B	B	B	A

two syllables should be decided jointly by their initials and rhymes. We can see the distances between syllables' initials and rhymes from their weighted adjacency matrix, respectively as discussed above (Tables 4 and 6), which are denoted by x and y. On that basis, we use the Euclidean distance on two-dimensional plane to represent the sound distance of syllables (Fig. 5).

$$D = \sqrt{x^2 + y^2}$$

With this formula, we obtain the quantitative criteria to measure the similarity level of different monosyllabic words or characters (as each character was created to record a monosyllabic word).

4 The Applications in Chinese Study

The purpose of introducing the graph model into Old Chinese phonology is to make this long-established empirical study more quantitative and intuitive, and be more easily accepted by modern people. There are some very practical uses in Chinese study.

4.1 The study of Xiesheng series in Chinese characters

Chinese writing is a representative of ideographic system. It does not mean that, however, Chinese

Table 4 Weighted adjacency matrix of Old Chinese initials

	影	晓	匣	见	溪	疑	端	透	定	泥	来	精	清	从	心	帮	滂	並	明
影	0	1	1	2	2	2													
晓	1	0	1	2	2	2													
匣	1	1	0	2	2	2													
见	2	2	2	0	1	1													
溪	2	2	2	1	0	1													
疑	2	2	2	1	1	0													
端							0	1	1	1	1	2	2	2	2				
透							1	0	1	1	1	2	2	2	2				
定							1	1	0	1	1	2	2	2	2				
泥							1	1	1	0	1	2	2	2	2				
来							1	1	1	1	0	2	2	2	2				
精							2	2	2	2	2	0	1	1	1				
清							2	2	2	2	2	1	0	1	1				
从							2	2	2	2	2	1	1	0	1				
心							2	2	2	2	2	1	1	1	0				
帮																0	1	1	1
滂																1	0	1	1
並																1	1	0	1
明																1	1	1	0

characters have no information about their sounds. Most Chinese characters, in particular those constructed in Old Chinese period (about >87%), have a phonetic token, which we call picto-phonetic characters. Such characters with the same phonetic token are called Xiesheng series, as mentioned above. To study these sound relationships in Xiesheng series can help us learn more knowledge about the formation mechanism and the laws of Chinese characters.

Figure 6 shows a screen shot of the software (Hu, 2010) to measure and visualize the distribution of characters based on the sound distance in a Xiesheng series (like characters with 母 as a direct or indirect phonetic token). Based on the graph models of Old Chinese initials and rhymes, the software can calculate the minimal distance, the maximal distance, the average distance, the distance variance between the initials or rhymes in a Xiesheng series, and their joint syllable distance by the formula cited above. Given a syllable (an initial plus a rhyme) as the origin, it can also draw their distribution of sound distance on a two-dimensional plane that would help people understand intuitively the sound similarity or difference in a Xiesheng series.

4.2 The study of cognate words in Chinese etymology

Chinese etymology aims at the inner relationships between the sounds and meanings of words.

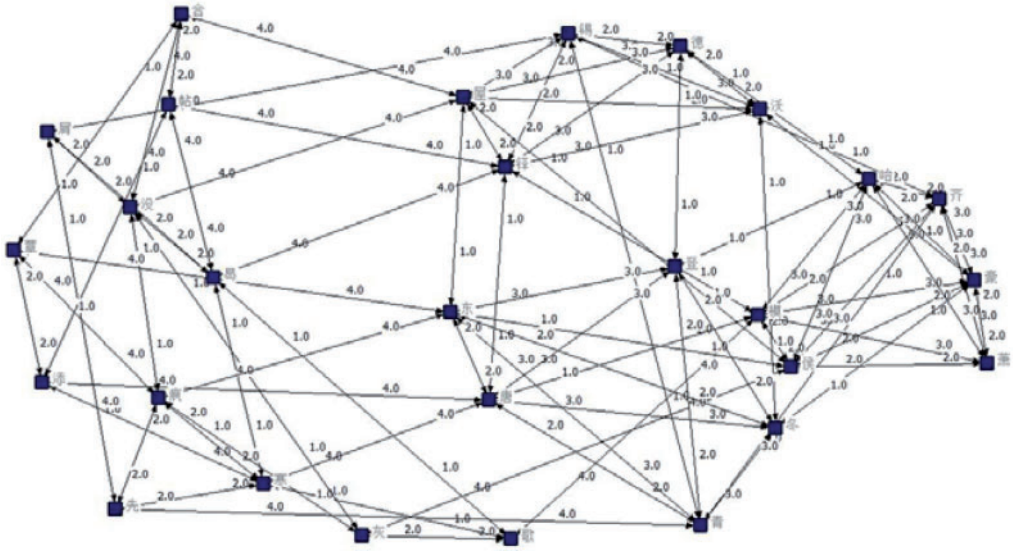

Fig. 4 Undirected weighted graph of Old Chinese rhymes

Table 5 Relationship adjacency matrix of Old Chinese rhymes

	歌	曷	寒	灰	没	痕	屑	先	齐	锡	青	模	铎	唐	侯	屋	东	萧	豪	沃	冬	哈	德	登	合	覃	帖	添
歌	A	B	B	C								E																
曷	B	A	B		C		C						E														E	
寒	B	B	A			C		C						E														E
灰	C			A	B	B									E													
没		C		B	A	B	C									E									E			
痕			C	B	B	A		C									E									E		
屑		C		C			A	B		E																		
先			C		C		B	A			E																	
齐									A	B	B	C			D			D	D			C						
锡							E		B	A	B		C			D				D				C				
青								E	B	B	A			C			D				D				C			
模	E								C			A	B	B	C			D	D			D						
铎		E								C		B	A	B		C				D			D				E	
唐			E								C	B	B	A			C				D			D				E
侯				E					D			C			A	B	B	C	C			D						
屋					E					D			C		B	A	B			C			D		E			
东						E					D			C	B	B	A				C			D		E		
萧									D			D			C			A	C			D						
豪									D			D			C			C	A	B	B	C						
沃										D			D			C			B	A	B		C					
冬											D			D			C		B	B	A			C				
哈									C			D			D			D	C			A	B	B				
德										C			D			D				C		B	A	B				
登											C			D			D				C	B	B	A				
合					E											E									A	B	C	
覃						E											E								B	A		C
帖		E											E												C		A	B
添			E											E												C	B	A

Table 6 Weighted adjacency matrix of Old Chinese rhymes

	歌	曷	寒	灰	没	痕	屑	先	齐	锡	青	模	铎	唐	侯	屋	东	萧	豪	沃	冬	哈	德	登	合	覃	帖	添
歌	0	1	1	2								4																
曷	1	0	1		2		2						4														4	
寒	1	1	0			2		2						4														4
灰	2			0	1	1									4													
没		2		1	0	1	2									4									4			
痕			2	1	1	0		2									4									4		
屑		2			2		0	1		4																		
先			2			2	1	0			4																	
齐									0	1	1	2			3			3	3			2						
锡							4		1	0	1		2			3				3			2					
青								4	1	1	0			2			3				3			2				
模	4								2			0	1	1	2			3	3			3						
铎		4								2		1	0	1		2				3			3				4	
唐			4								2	1	1	0			2				3			3				4
侯				4					3			2			0	1	1	2	2			3						
屋					4					3			2		1	0	1			2			3		4			
东						4					3			2	1	1	0				2			3		4		
萧									3			3			2			0	2			3						
豪									3			3			2			2	0	1	1	2						
沃										3			3			2			1	0	1		2					
冬											3			3			2		1	1	0			2				
哈									2			3			3			3	2			0	1	1				
德										2			3			3				2		1	0	1				
登											2			3			3				2	1	1	0				
合						4										4									0	1	2	
覃							4										4								1	0		2
帖		4											4												2		0	1
添			4											4												2	1	0

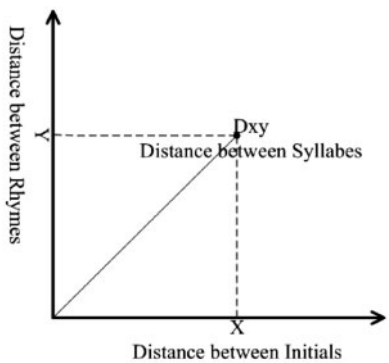

Fig. 5 Euclidean distance between syllables

The main task of Chinese etymology is to infer a given word's source token, correlate its cognate words, and put all of them into a family that have the same etymon. To judge whether two words are cognates must examine whether their sounds are similar and their meanings have the same features. To compare two words' meanings, we can use various materials of philological exegesis in Chinese classics; whereas to compare two words' sounds, we can use the Old Chinese phonological system.

Figures 7 and 8 show screen shots of the software (Hu, 2010) to measure and visualize the sound distance distribution of words with similar meanings. Based on the graph models of Old Chinese initials and rhymes, it can calculate, again, the minimal distance, the maximal distance, the average distance, the distance variance between the initials or rhymes, and their joint syllable distance by the formula cited above. Given a word's syllable (an initial plus a rhyme) as the origin, it can draw the distance distribution between sounds on

Fig. 6 Sound distance distribution in a Xiesheng series

Fig. 7 Sound distance distribution of four words with similar meanings

Fig. 8 Sound distance distribution of three cognate words

a two-dimensional plane that would help people understand intuitively the similarity or difference between sounds.

In this example, we want to find out the cognates of Chinese word 天 (tian), and list all the words that have the same explanation as 天 in Shuowen Jiezi (a wordbook of Old Chinese compiled in Han Dynasty). As seen in Fig. 7, there is a word's position far from others in the sound distance distribution plane, which may not be the cognate of 天. So we exclude this word, and work out that the sound distance of other three words are very close in Fig. 8, and they are thus real cognates.

5 Conclusions

As a Pan-time and Pan-domain social phenomenon, the sound relationships in Old Chinese are complicated and the research methods about them are empirical. The graph model of Old Chinese phonological system proposed in this article is still in its infancy, and there are numerous aspects to be developed, such as the arrangement of weights for different relationships between initials or rhymes, and the calculation of joint distance in between. However, this article put forward a mathematic model to measure the similarity level of the sounds of words (or characters) in Old Chinese for the first time and provided the long-established qualitative study of traditional Old Chinese phonology with some quantitative references and tests that may bring insights to Chinese phonology studies, such as the research of Xiesheng series in Chinese characters, or Chinese Etymology like what this article has discussed above.

References

Baxter, H. (1992). *A Handbook of Old Chinese Phonology*. Berlin, Germany: Mouton de Gruyter.

Hu, J. (2010). *The Construction of the Inner Systems Model of "ShuoWen"* Ph.D. thesis, Beijing Normal University, China.

Huang, K. (1934). *Phonology Tables*. Shanghai, China: The Commercial Press.

Wang, N. (2007). The concepts and methods of the traditional study of Old Chinese phonology. *The Academic Journal of Sino-Tibetan*, **2**: 48–64.