On the "Orthodoxy" of the Phonologies and the Vocabularies of the Present Chinese Varieties

Chen Yuanhao

16 December 2018

Abstract The traditional Chinese literati had always sought to retain "orthodoxy" of pronunciations in order to study the ancient rhymes and to create new works of phonological harmony since the late Old Chinese era. This pursuit expedited the study of phonology in China and initiated the creation of numerous rime dictionaries, the most notable of which was *Guangyun*. This essay, using the reconstruction of *Guangyun* phonology as the standard, investigates into the "orthodoxy" of 13 major present Chinese varieties under the criterion of traditional literati, including contrastivities of four tones, 38 initials and 95 finals, which this essay proposes a "contrastive phoneme test" to analyse. After classing the phonetic material under the literati's criteria, the result of the contrastive phoneme test can be converted to be expressed mathematically as the Pearson's correlation coefficient of two vectors related to the phonemes of *Guangyun* and the Chinese variety studied. The conclusion reached, consistent with public impressions, is that southern Chinese dialects preserve Middle Chinese phonology more adequately. The results not only can be used to cluster Chinese varieties but also strongly support a credible hypothesis that there exists a genetic relationship between Gan and Hakka peoples. In the end, this essay reviews the literature on the lexical constitution of Chinese, reaching to the conclusion that Chinese words originate from numerous sources such as Kam-Tai substratum of southern regionalects, contact with Turkic languages and possible co-origination with Indo-European languages. Besides reflecting the great amount of non-Sinic vocabulary particularly in southern regionalects, the evidence also broadens our apprehension, inspiring us to reappraise the ontogenesis of the Chinese lexicon.

Keywords prosodic analysis, contrastive phonemes, lexical composition, quantitative linguistic analysis

1 Introduction

In the acclaimed *Yan Family Mottoes* 顏氏家訓, Yan Zhitui 顏之推 rebuked the common pronunciation of his time:

> 南染吳越，北雜夷虜。[31]
>
> [The people's accent] is tainted by that of the southern Wu and Yue aboriginals and intermingled with that of the northern Yi and Lu barbarians.

Biased as the stricture was, it epitomised the emphasis on the "orthodoxy" of pronunciation in Middle Chinese literature, which derived from Old Chinese literature dating back to the 13th century BC, deeply imprinted in the literati of past dynasties. In linguistic terms, they were seeking to shorten the linguistic distance of their own tongues to that of former literati.

The Book of Songs 詩 $[*\text{hljɯ}]_{\text{OC}}$, compiled in the 6th century BC, covering 305 poems from the 11th to 7th centuries BC[3], was the earliest collection of rhyming Chinese literary works, considerably influencing the Chinese literati up to this day. For modern phonologists, this marked the conclusion of the Old Chinese phonology and the transition to the Middle Chinese *Qieyun* 切韻 (a rime table believed to have recorded the authentic Early Middle Chinese phonology[22]) system.

In the *Qieyun* era, phonology occurred as two phenomena transpired in the literary arena:

1. Literati began to study ancient texts such as *The Book of Songs*, which led to an emergent branch of learning, 訓詁 (critical interpretation of ancient texts), which involved the reconstruction of Old Chinese phonology.
2. In the Yongming 永明 era (483–93), a poet, widely believed to be Shen Yue 沈約 (441–513), created the poetic prosody of *si sheng ba bing* 四聲八病 (the four tones and eight defects), which became a principle of using tones and rhymes in poetry, as opposed to the unrestrained rhythm of poems in *The Book of Songs* [25].

As a direct result, the literati created the first rime table, *Qieyun*, inspired by the Sanskrit syllable charts [4], and began to stress the importance of the "orthodoxy" of their own pronunciations; otherwise, the classic works would not rhyme, and the *si sheng ba bing* would be inapplicable in their own sound system.

Besides the aforementioned *Yan Family Mottoes*, a contemporary case of the phonological bias would be the use of accents in some of the *quyi* 曲藝 (folk vocal art forms). For example, in Suzhou Pinghua 評話, there are three accents for different characters: Mandarin used by snobs, colloquial Suzhounese used by commoners and literary Suzhounese used by the Emperor, ministers and scholars. Similar discriminations between accents are seen in Comedic Opera 滑稽戲, Tanci 彈詞, etc.

But Ming Dynasty scholar Chen Di 陳第 wrote:

> 蓋時有古今,地有南北,字有更革,音有轉移,亦勢所必至。
>
> There is the past and the present; there is the north and the south. It is only inevitable that characters evolve, and sounds change. [3]

Following the steps of Chen Di, without adhering to any prejudices in the ways by which people created hierarchies for Chinese varieties, we try to evaluate the "orthodoxy" of **present Chinese varieties**, which is the combination of Modern Chinese varieties and Modern Min varieties, with the Middle Chinese *Guangyun* 廣韻 (supplement version of *Qieyun*, literally: *Broad Rimes*) system reconstructed by Zhengzhang [27] as the standard, with conventional criteria in Chinese phonology, but with modern methodologies regarding prosodical analysis, contrastivity and quantitive linguistics. Further, it is necessary to discuss the "orthodoxy" of vocabularies of these varieties, though there are no comprehensive ways to do so for individual varieties. Thus, the second part of this essay would be a comprehensive review of the literature about the constitution of Chinese vocabulary.

The subjects of investigation are shown in blue, and the phonological standard, Middle Chinese, is in red in figure 1.

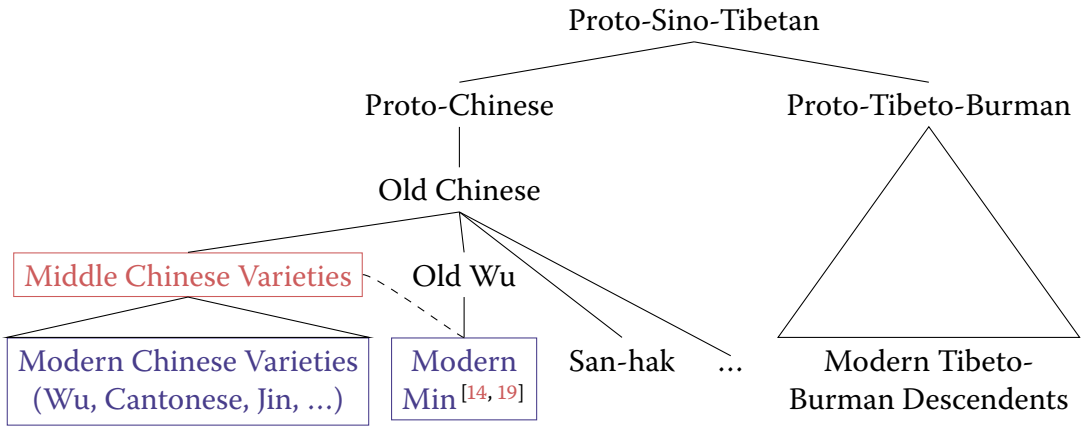

Figure 1: Language Family Tree of Chinese Varieties

2 Prosodical Analysis

2.1 Methodology

The criteria should contain all prosodic features of Middle Chinese:

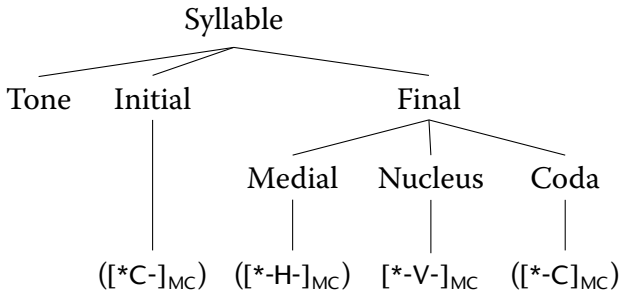

Figure 2: Middle Chinese Syllable Structure [18, 3]

Then, we can decide how crucial each element is. Because the purpose of pursuing "orthodoxy" in phonology was clear, – to maintain the rhymes in the classics, and to produce new literary works adhering to the traditional rhyming schemes – the criterions would have to include all the significant phonemes in Middle Chinese rhyming schemes, including tones and final rimes, whose interrelationship should be given specific consideration.

Besides, preservation of the initials also affects a variety's "orthodoxy." Not only are initials themselves the inheritances from older prosodies, but in several Chinese varieties, they also determine the yin and yang tones of characters. For example, one phonological rule in Modern Shanghainese is

$$[\mathrm{V}^{\text{平 level}}]_{\mathrm{MC}} \rightarrow \begin{cases} [\mathrm{V}^{\text{陰平 yin level}}]_{\mathrm{SH}} \ / \ \text{Unvoiced}\ [\mathrm{C}]_{\mathrm{MC}\,\rightarrow\,\mathrm{SH}}\ \_\_ \\ [\mathrm{V}^{\text{陽平 yang level}}]_{\mathrm{SH}} \ / \ \text{Voiced}\ [\mathrm{C}]_{\mathrm{MC}\,\rightarrow\,\mathrm{SH}}\ \_\_ \end{cases},$$

Phonological Rule 1: Middle Chinese and Shanghainese level 平 tones

where MC and SH are sampled from *Guangyun* and Middle Modern Shanghainese respectively. As shown in rule 1, all level tone characters in Middle Chinese become yin level tone when the initial is unvoiced, and yang level tone otherwise. Therefore, there is a necessity to consider the interrelationship between tones and initials.

2.1.1 Tones

Guangyun (finished in 1008), the most influential Middle Chinese rime dictionary, recorded in it the pronunciations of 26,194 characters, which were categorised into four tones (level 平, rising 上, departing 去 and entering 入), and these are the only significant tonal features in the *Guangyun* system.

However, different tonal shifts cause violations of the traditional rhyming schemes of fairly different consequences.

In Old Chinese, there were no tones, yet the codas and post-codas, sometimes consonant clusters, corresponded with Middle Chinese tones [15, 3]. For instance, in the Fifth Chapter of the July Volume of *the Book of Songs* 詩·七月·五章, the rhymes all had the same codas, the glottal stop $[\text{*-ʔ}]_{\mathrm{OC}}$,

as shown in table 1 (using Zhengzhang's OC reconstruction [26]), all of which fall into the MC rising tone category.

Table 1: Evolvement of $[*\text{-}ʔ]_{OC}$-Type Characters

Rhymes in 詩·七月·五章								
股	羽	野	宇	戶	下	鼠	子	處
$[*klaːʔ]_{OC}$	$[*ɢ^{w}aːʔ]_{OC}$	$[*laːʔ]_{OC}$	$[*ɢ^{w}aːʔ]_{OC}$	$[*ɢ^{w}aːʔ]_{OC}$	$[*graːʔ]_{OC}$	$[*hljaːʔ]_{OC}$	$[*ʔslɯʔ]_{OC}$	$[*kʰljaːʔ]_{OC}$
				↓				
				$[*CV^{上\ rising}]_{MC}$				

Abstracted into one phonological rule, this is

$$[*CVʔ]_{OC} \rightarrow [*CV^{上\ rising}]_{MC}.$$

Phonological Rule 2: Correspondence of MC Rising Tone and OC Glottal Stop Coda

Therefore, an exemplary Chinese variety should conserve the correspondence to the four MC tones to be **inflexionally harmonic** 平仄相調 (harmonising the level and the oblique).

Few Chinese varieties preserve such a correspondence, yet amongst the inconsistencies, two would debase the phonology more: the absence of the entering tone and the jumbling of the level tone with the other three or vice versa:

The MC entering tone is the only one which contains coda stops $[*p, k, t]_{MC}$, coming directly from $[*b, g, d]_{OC}$. The short and forceful auditory impression that the entering tone had was an efficacious literary device that the *Qieyun* era poets used.

Song dynasty lyricist Li Qingzhao 李清照 wrote at the beginning of her song 聲聲慢 (in about 1127):

> 尋尋覓覓,冷冷清清,淒淒慘慘戚戚。[2]
>
> I look for and I look for; it's desolate and it's desolate; it's bleak, bleak, wretched, wretched and lonesome, lonesome.

She used the repetition of entering rhymes 覓覓 $[*mek\ mek]_{MC}$ and 戚戚 $[*\widehat{ts}^{h}ek\ \widehat{ts}^{h}ek]_{MC}$ to let out her desolate sigh for her unfortunate divorce and disintegrating homeland. The rest of the song continued to rhyme with the entering tone:

Table 2: Example of Entering Rhymes in Literature

Rhymes in 聲聲慢									
覓	戚	息	急	識	積	摘	黑	滴	得
$[*mek]_{MC}$	$[*\widehat{ts}^{h}ek]_{MC}$	$[*sɨk]_{MC}$	$[*kɣɨɪp]_{MC}$	$[ɕɨk]_{MC}$	$[*\widehat{ts}ɨɛk]_{MC}$	$[*t^{h}ek]_{MC}$	$[*hək]_{MC}$	$[*tek]_{MC}$	$[*tək]_{MC}$

And because the MC level tone has a poised auditory impression, it can bear distinctive literary meanings in a way similar to the entering tone. When writing couplets, for example, the last character of the first line could not be a level-tone character, and that of the second line must be a level-tone character, in order to close the couplet in inflexional harmony. Therefore, any shift from the entering and the level tones to others is a dissonant alteration of the phonology, while, for instance, an interchange of rising and departing tones would not be considered as such.

Tone sandhi are not taken into consideration for they are not phonemes when reading texts.

2.1.2 Initials

Guangyun recorded 36–38 initials (it is debatable whether the 常 initial [*d͡ʑ-]$_{MC}$ and the 以 or 云 initial [*Ø-]$_{MC}$ were separate from 俟 [*ʒ-]$_{MC}$ and 匣 [*ɦ-]$_{MC}$), which were put into four categories 全清, 次清, 次浊, 全浊[29], which perfectly correspond with unvoiced unaspirated, unvoiced aspirated, sonorant, and voiced consonants[9].

The most unambiguous criterion of initials is thus that initials of different categories should not be confused. For instance, the four palatal alveolar retroflex stops 舌上 in table 3 should remain contrastive under the four-category initials system.

Table 3: Contrastivity of Palatal Alveolar Retroflex Stops 舌上對立[29]

	全清	次清	次浊	全浊
舌上	知 [*ʈ-]$_{MC}$	徹 [*ʈʰ-]$_{MC}$	澄 [*ɖ-]$_{MC}$	娘 [*ɳ-]$_{MC}$

Some Middle Chinese literati even believed that manoeuvring the contrastive relation of the unvoiced and the voiced is more essential to the beauty of poems and the conformity of sounds than the contrastive tones:

> 王元長創其首，謝朓、沈約揚其波。……余謂文制，本須諷讀，不可蹇礙。但令清濁通流，口吻調利，斯爲足矣。至如平上去入，則余病未能；蜂腰、鶴膝，閭裏已具。[30]

> Wang Rong initiated; Xie Tiao and Shen Yue whipped up its waves. … I think that poems by nature need to be read out, so they should not be difficult and hindering to pronounce. It is enough if the "clear" [voiced 清] and "turgid" [unvoiced 濁] sounds go fluently together and the lines flow smoothly. Talking about [using] the "level," "rising," "departing," and "entering" tones [to govern rhymes], I am afraid it is an impossible thing. As for "wasp's waists" and "crane's knees," they are common in poems composed in villages.[25]

Therefore, even though the stop consonants of the unvoiced unaspirated and the voiced categories have similar auditory impressions, and in fact in all present Chinese varieties except Wu and Old Xiang the contrast has been lost[19], it is a vital characteristic of MC initials that indicates "orthodoxy."

Contrastive Phoneme Test Aside from checking the four-category contrast, a contrastive phoneme test is proposed in order to test phonological "orthodoxy" of phonemes. Figure 3 illustrates an example where the contrastivity between two phonemes P_1 and P_2 are preserved in the contrastive independent new phonemes Q_1 and Q_2, which also function the same as P_1 and P_2 when constructing morphemes. Thereby, the phonology is unaltered by phonological processes $P_1 \rightarrow Q_1 \,/\, C$ and $P_2 \rightarrow Q_2 \,/\, C$. If a set of phonological changes alters any independence of phonemes so that either Q_1 or Q_2 equals to another phoneme in the language, then it does not pass the contrastive phoneme test. The test is applicable to all phonetic features, including tones.

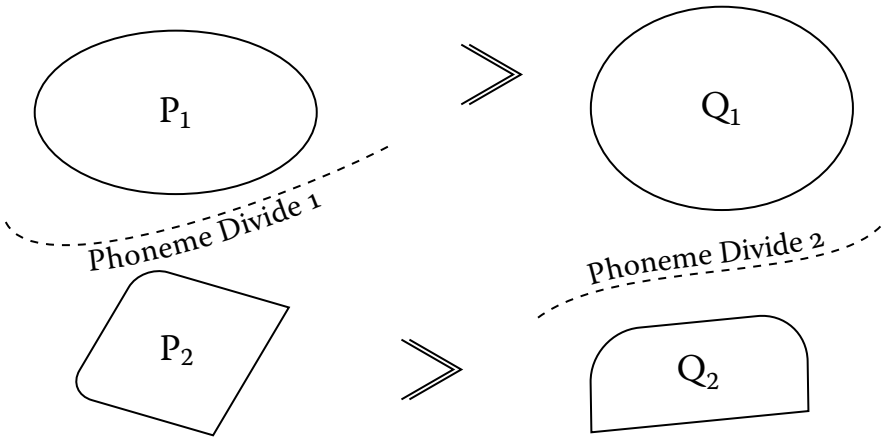

Figure 3: Preserved Contrastivity of Phonemes

Under this definition, the conversion of MC 疑 and 影 initials, $[*ŋ, ø\text{-}]_{MC}$, in Gan Chinese 贛語 does not pass the test, disrupting the MC phonology, as shown in phonological rule 3.

$$[*ŋ\text{-}]_{MC} \rightarrow \begin{cases} [n, ɳ\text{-}]_{GC} / \_[\text{-i}, y]_{GC} \\ [Ø\text{-}]_{GC} / \text{Otherwise} \end{cases}$$

$$[*ʔ\text{-}]_{MC} \rightarrow \begin{cases} [ŋ\text{-}]_{GC} / \_[a, ɔ, e\text{-}]_{GC} \\ [Ø\text{-}]_{GC} / \text{Otherwise} \end{cases}$$

Phonological Rule 3: GC Change of MC 疑 and 影 Initials[20]

2.1.3 Finals

Guangyun recorded 206 rimes; leaving out tones, there are 95 finals in the system. Except for Standard Yue Chinese (having 94 finals, 92 of which are inherited from Middle Chinese) and Southern Min (preserving 41–86 initials depending on region), present Chinese varieties generally have lost about half of the initials: Suzhounese has 48; Beijing Mandarin has 40; Shanghainese has 32; Wenzhounese has only 14. — Though the "old-fashioned" 老派 tongues (spoken when dialect sampling was widely conducted in China in the early 20th century) of these varieties generally preserve much more finals (such as old-fashioned Shanghainese with 51 initials, spoken by locals born around 1920), it is indubitable that the irreversible breakdown of phonologies took place all of a sudden in mere decades when social reforms and cultural interactions thereon increased. In consequence, this analysis would only look at the most competitive tongue, though the methodology applies to all other tongues.

Besides the contrastive phoneme test, which is also applicable to finals, we should recognise that the entering tone codas are a special case since they are always unvoiced and only serve as stops. Thus, confusing entering tone finals $[*\text{-Vp}, \text{-Vk}, \text{-Vt}]_{MC}$ or replacing them with glottal stops [ʔ] does not thoroughly affect auditory impression or inflexional harmony.

Unlike some studies which create exceptions for the confusion of medials and quotas, or for some finals with nasal codas when investigating intelligibility[8, 23], because this analysis discusses "orthodoxy" which traditional scholars proposed trying to retain old phonologies, and not to make old phonologies intelligible, so all other mutations of finals reduces a variety's "orthodoxy."

2.2 Implementation

2.2.1 Material

The sources of the tonal structures are researches by Norman and Margaret Mian [18, 24], which further divided the major present varieties of Chinese into regionalects that represent the tongues of 86.7% of the Chinese population [1].

Table 4: Tonal Categories and Pitch Contours in Colloquial Layers of Regionalects

		Guangyun Tones and Contrastive Initial Types											
		level			rising			departing			entering		
		vl.	n.	vd.	vl.	n.	vd.	vl.	n.	vd.	vl.	n.	vd.
Jin	Taiyuan	1 1			3 13			5 1			7 11		8 13
Mandarin	Xi'an	1 11	2 11		3 11			5 1			1		2
	Beijing	1 1	2 11		3 113			5 51			1, 2, 3, 5	5	2
	Chengdu	1 1	2 11		3 13			5 11			2		
	Yangzhou	1 11	2 11		3 11			5 1			7 1		
Xiang	Changsha	1 1	2 11		3 11		6	5 1	6 11		7 11		
	Shuangfeng	1 1	2 11		3 11		6	5 11	6 1		2, 5		
Gan	Nanchang	1 11	2 11		3 113		6	5 11	6 11		7 1		8 11
Wu	Suzhou	1 1	2 11		3 11		6	5 51	6 11		7 1		8 11
	Wenzhou	1 1	2 11		3 11	4 11		5 11	6 1		7 11		8 11
Min	Xiamen	1 1	2 11		3 51		6	5 1	6 1		7 11		8 1
Hakka	Meixian	1 1	2 11		3 11	1, 3	1	5 11			7 11		8 1
Yue	Guangzhou	1 13	2 11		3 11	4 11 (6)		5 1	6 1		7a 1	7b 1	8 1

And the data of initials and finals of these same regionalects are from a study conducted by Peking University [6]; some examples are shown in tables 5 and 6.

Table 5: Diversions of 3 Sample MC Initials in Regionalects

		Jin	Mandarin				Xiang		Gan	Wu		Min	Hakka	Yue
		TY	XA	BJ	CD	YZ	CS	SF	NC	SZ	WZ	XM	MX	GZ
幫 $[*p]_{MC}$	p	89	88	88	87	89	88	85	86	90	89	85	85	87
	p^h	2	3	3	5	3	4	6	6	0	0	7	7	5
	b	0	0	0	0	0	0	1	0	2	1	0	0	0
	m	1	1	1	0	0	0	0	0	0	0	0	0	0
滂 $[*p^h]_{MC}$	p^h	40	38	40	39	40	38	38	40	39	38	37	38	39
	p	1	3	1	1	1	3	2	1	2	3	4	2	2
	f	0	0	0	1	0	0	0	0	0	0	0	0	0
	x	0	0	0	0	0	0	1	0	0	0	0	0	0
並 $[*b]_{MC}$	p	38	32	37	37	38	72	4	5	0	0	57	10	33
	p^h	38	44	39	39	38	5	9	72	0	0	20	67	43
	b	0	0	0	0	0	0	64	0	77	77	0	0	0
	f	1	1	0	1	1	0	0	0	0	0	0	0	1

Table 6: Diversions of 1 Sample MC Final in Regionalects

		Jin	Mandarin				Xiang		Gan	Wu		Min	Hakka	Yue
		TY	XA	BJ	CD	YZ	CS	SF	NC	SZ	WZ	XM	MX	GZ
東董送合口一等 [*uŋ]$_{MC}$	uŋ	45	0	42	0	0	0	0	48	0	0	0	48	48
	oŋ	0	43	0	48	0	48	0	0	48	48	0	0	0
	ɔuŋ	0	0	0	0	48	0	0	0	0	0	0	0	0
	aŋ	0	0	1	0	0	0	47	0	0	0	10	0	0
	ɔŋ	0	0	0	0	0	0	0	0	0	0	36	0	0
	əŋ	3	3	3	0	0	0	1	0	0	0	0	0	0
	uəŋ	0	0	2	0	0	0	0	0	0	0	0	0	0
	uoŋ	0	2	0	0	0	0	0	0	0	0	0	0	0
	iã	0	0	0	0	0	0	0	0	0	0	1	0	0
	ɪŋ	0	0	0	0	0	0	0	0	0	0	1	0	0

2.2.2 Quantitive Analysis

To conduct a quantitive analysis with the material, we need to convert the methodologies in section 2.1 into computable mathematical language.

Going back to the introduction in section 1 and the contrastive phoneme test, we recall that Chinese literati pursued "orthodoxy" shorten the linguistic distance between their phonology and that of former scholars; hence, when a group of phonemes P undergo P → Q, it should be checked whether there are many-Q-to-one-P$_i$ or many-P-to-one-Q$_j$ correspondences. The sparser the correspondences, the long the linguistic distance.

In practice, to quantify the extent to which linguistic changes fail the test, we match up phonemes (could be either tones, initials or finals) in *Guangyun* and those in each regionalect to form dichotomies according to contrastivity discussed in section 2.1: MC level and entering tones are strictly isolated and not considered exchangeable with other tones; initials from different four-category groups are considered independent; all entering-tone finals are considered exchangeable. Then, label the phonemes from p_1 to p_n (these are different notations from what were used for explaining the contrastivity test; n could be different if we consider different types of phonemes, e.g., $n = 4$ when we consider only tones). Let N_i^{CV} denote the number of usages of p_i in a Chinese variety CV. There would be unmatchable phonemes, for example, when processing the Beijing regionalect, the MC entering tone cannot be matched to any BJ tones, so by definition, since the entering tone is the fourth phoneme, $N_4^{\mathsf{BJ}} = 0$. For convenience, let $\overrightarrow{\mathsf{CV}} = [N_i^{\mathsf{CV}}]$.

Hence, we have created a nominal-dichotomous database for initials in pairs formed by *Guangyun* and each regionalect. In this way, we can quantify the score of the contrastive phoneme test mathematically as the Pearson correlation coefficient of the vectors $\overrightarrow{\mathsf{RL}}$ (RL = regionalect) and $\overrightarrow{\mathsf{MC}}$; therefore,

$$\begin{aligned} s_{\mathsf{RL}} &= r_{\overrightarrow{\mathsf{RL}},\overrightarrow{\mathsf{MC}}} \\ &= \frac{n\sum N_i^{\mathsf{RL}}N_i^{\mathsf{MC}} - \sum N_i^{\mathsf{RL}}\sum N_i^{\mathsf{MC}}}{\sqrt{n\sum\left(N_i^{\mathsf{RL}}\right)^2 - \left(\sum N_i^{\mathsf{RL}}\right)^2}\sqrt{n\sum\left(N_i^{\mathsf{MC}}\right)^2 - \left(\sum N_i^{\mathsf{MC}}\right)^2}}, \end{aligned} \tag{1}$$

where s_{RL} is the score of RL under the contrastive phoneme test with *Guangyun* as the standard.

Further, let s_{RL}^{mat} denote s_{RL} measured with data derived from the material *mat*. "Orthodoxy" can

be defined as the geometric mean of $s_{\mathrm{RL}}^{\mathrm{tone}}$, $s_{\mathrm{RL}}^{\mathrm{initial}}$ and $s_{\mathrm{RL}}^{\mathrm{final}}$:

$$\textsc{orth}_{\mathrm{RL}} = \left(s_{\mathrm{RL}}^{\mathrm{tone}}\, s_{\mathrm{RL}}^{\mathrm{initial}}\, s_{\mathrm{RL}}^{\mathrm{final}}\right)^{\frac{1}{3}}. \tag{2}$$

2.2.3 Example

To give a brief example, we discuss s_{XA}, investigating the tones only. From table 4 we obtain the relations as shown in figure 4.

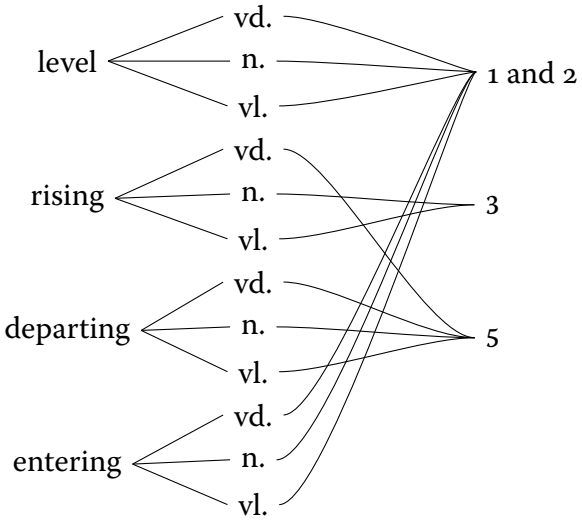

Figure 4: Genetic Relation of XA and MC Tones

The phonemes sorted out are hence $p_{1,2,3,4} = [\text{*-V}^{\text{level, rising, departing, entering}}]_{\mathrm{MC}}$, corresponding to $[\text{-V}^{(1\text{ and }2),3,5,\text{absent}}]_{\mathrm{XA}}$. We can then calculate N_i^{MC} and N_i^{XA} simply from the numbers of entries of each combination of tone and initial type in *Guangyun*; the results are

$$N_1^{\mathrm{XA}} = 3788 + 1806 + 2032 + 1949 + 772 + 783 = 11130 \tag{3a}$$
$$N_2^{\mathrm{XA}} = 872 + 1868 = 2740 \tag{3b}$$
$$N_3^{\mathrm{XA}} = 780 + 871 + 783 + 2041 = 4475 \tag{3c}$$
$$N_4^{\mathrm{XA}} = 0; \tag{3d}$$

$$N_1^{\mathrm{MC}} = 7626 \tag{4a}$$
$$N_2^{\mathrm{MC}} = 3520 \tag{4b}$$
$$N_3^{\mathrm{MC}} = 3695 \tag{4c}$$
$$N_4^{\mathrm{MC}} = 3504. \tag{4d}$$

Substitute equations 3 and 4 into equation 1. Hence

$$s_{\mathrm{XA}}^{\mathrm{tone}} = 0.934. \tag{5}$$

when examining its tones.

2.2.4 Results

Repeating the example given above for other regionalects and for initials and finals, we can obtain ORTH_{RL} for all regionalects.

Table 7: "Orthodoxy" of Each Variety

	Jin	Mandarin				Xiang		Gan	Wu		Min	Hakka	Yue	*Guangyun*
	TY	XA	BJ	CD	YZ	CS	SF	NC	SZ	WZ	XM	MX	GZ	MC
ORTH_{CV}	.242	.608	.611	.592	.598	.679	.702	.651	.743	.747	.759	.650	.741	1.00

With the *Mathematica*® [1] command `ClusteringTree` we can further investigate the data and cluster the regionalects based on ORTH_{CV}.

```
ClusteringTree@{0.6106` -> "BJ", 0.7412` -> "GZ",
  0.6079` -> "XA", 0.24198` -> "TY",
  0.59215` -> "CD", 0.59835` -> "YZ",
  0.74332` -> "SZ", 0.74735` -> "WZ",
  0.678683` -> "CS", 0.702416` -> "SF",
  0.65061` -> "NC", 0.65025` -> "MX",
  0.75908` -> "XM", 1 -> "MC"}
```

The resultant clustering tree is shown in figure 5, and Xiamen Min is hence the most "orthodox" regionalect. Surprisingly, though merely drawn based on linguistic distances to the *Guangyun* system, not on the affinities amongst the regionalects themselves, the tree corresponds to the classification and genetic relations of these varieties well. The two main branches under MC divide present Chinese varieties into the northern and the southern ones, of which the latter preserves MC phonology well, especially the entering tone. Also, this graph could verify the hypothesis that the Hakka 客 people are descendants of Gan-speaking people since their tongues are phonetically proximate (see NC and MX).

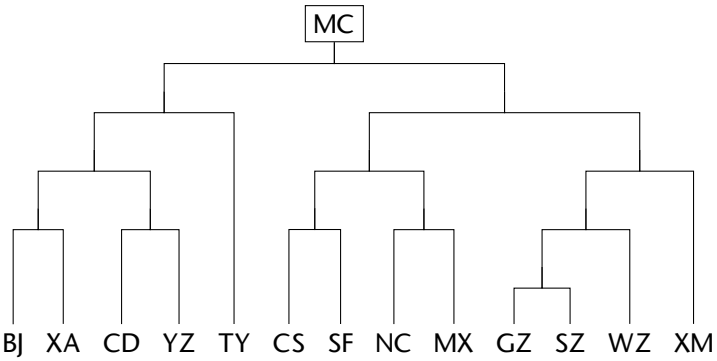

Figure 5: Clustering Tree Based on ORTH_{CV}

3 Review of the Literature on Lexical Composition

3.1 Kam-Tai Substratum

Yan was right when he concluded in his family mottoes that the southern people's accents had been mixed with non-"orthodox" elements [31], yet not only the accents but also some of the vocabulary were clearly of alien origin. This was due to the Kam-Tai substratum preserved in all southern tongues after sinicisation [14, 19], as shown in figure 6.

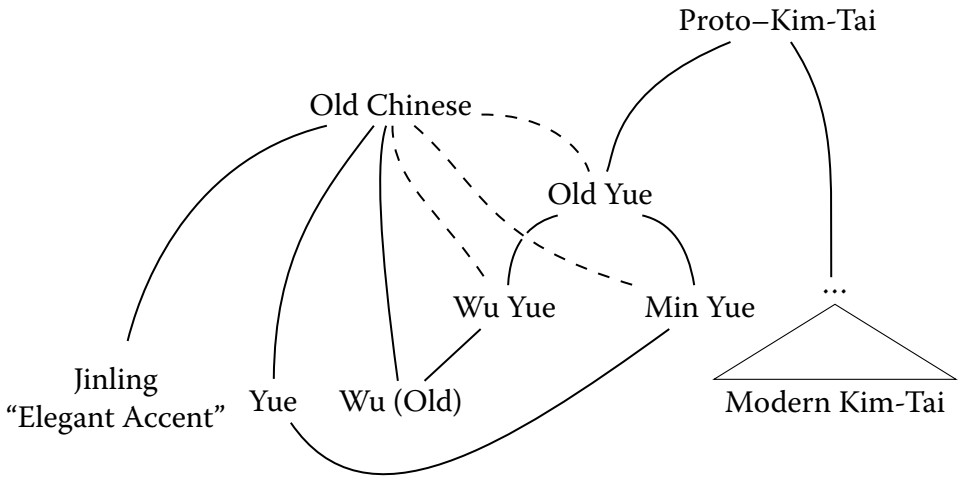

Figure 6: Influence of Kam-Tai Substratum

Of the regionalects investigated in section 2.2, Min, Yue and Wu are all preserving this substratum. Here are some examples [12, 16]:

1. /dai^{213}/$_{ngpin}$ (bird pecking), /yi^{213}/$_{ngpin}$ (also), /dzeng^{31}/$_{ngpin}$ (pot for spirits) mean the same in Wu, Zhuang and Dong.
2. /ba^{31}/$_{ngpin}$ (raft), /lai^{35}/$_{ngpin}$ (caress) and /gang^{42}/$_{ngpin}$ mean the same in Wu and Thai.
3. [ni]$_{yue}$ (this) functions the same way in Yue Chinese and all modern Kim-Tai and Hmong varieties.
4. [hɐn]$_{yue}$ (itchy) is similar to [hon]$_{Lianshan}$, [hun]$_{Wuming}$, [xum]$_{tdd}$ and [kum]$_{onb}$ which all have the same meaning.
5. "Collapsing" has the exact same pronunciation [lɐm] in Yue Chinese, North Zhuang, Maonan and Thai.

These words, besides frequently used colloquially in south China, are clearly of Kam-Tai origin, so it is unsurprising that more than 20% of them do not have corresponding Chinese characters yet [12]. Ironically, the southern Chinese varieties, though preserving the MC phonology significantly better than northern ones, are of Bai Yue alien origins.

3.2 Turkic Influence

Besides Yue influence, Yan also abhorred "barbaric" influence from the north. One notable example are the Chinese words [*kɑ]$_{MC}$ 哥, 阿哥 (elder brother), which is of Altaic origin. Before Tang Dynasty, the only word for elder brother was [*hʷɣiæŋ]$_{MC}$ 兄. Then the word 哥 emerged, initially meaning either "father" or "brother," later only meaning "elder brother." Chen Yanke 陳寅恪 (1943) made a comment on the phenomenon:

> 若以女系母統言之，唐代創業及初期君主，如高祖之母爲獨孤氏，太宗之母爲竇氏，即紇豆陵氏，高宗之母爲長孫氏，皆是胡種，而非漢族。
>
> If we look at the maternal bloodline of the first Tang emperor and several after him, Gao Great Emperor's mother's family name is Dugu; Tai Emperor's mother's family name is Dou, or Qi-Douling; Gao Emperor's mother's family name is Zhangsun; all of them are of barbaric blood, not Han blood.

Mei Tzu-Lin[14] also proved this by examining the alternative form of 哥, 阿干 $[*a\ kɑn]_{MC}$, which had one extra suffix [-n], corresponding to the Altaic plural inflexion [*-n].

3.3 Co-origination with Indo-European Languages

While only the boldest linguists are trying hard to explore ways to connect Old Chinese with ancient Sumerian, Na-Dene, Caucasian languages and Basque[17], researching the cognates in Old Chinese and Indo-European languages is not a newly emergent study.

Early systematic analyses of the cognate phenomenon include Tsung-tung Chang's study of cognate Proto–Indo-European roots and Old Chinese words[5]. His long list of cognate words also exhibited patterns between PIE finals and OC finals, later MC tones, as he made separate lists and analysed the correspondences between the finals. For example, $[*\text{-CVk}(^{w})]_{PIE}$ and $[*\text{-CVg}(^{w})]_{PIE}$ are clearly congruous with $[*\text{CV}^{ping}]_{OC}$ and $[*\text{CVg}^{ping}]_{OC}$; all $[*\text{kV(l)m}]_{PIE}$ have the same meanings with all $[*\text{ɦVm}^{ping}]_{OC}$ in Chang's lists. His knowledge of both Chinese and Indo-European languages guided his discernment to find more than 300 credible cognates, which reflected systematic, not idiosyncratic resemblance.

Zhou Ji-xu[28] took a similar approach, analysing cognate words in Proto–Sino-Tibetan and Proto–Indo-European. Notably, he discussed the emergent time of the cognate words listed, proving that they were not loanwords, and dating exactly the time nodes of when the languages separated from each other. He proposed the following relations:

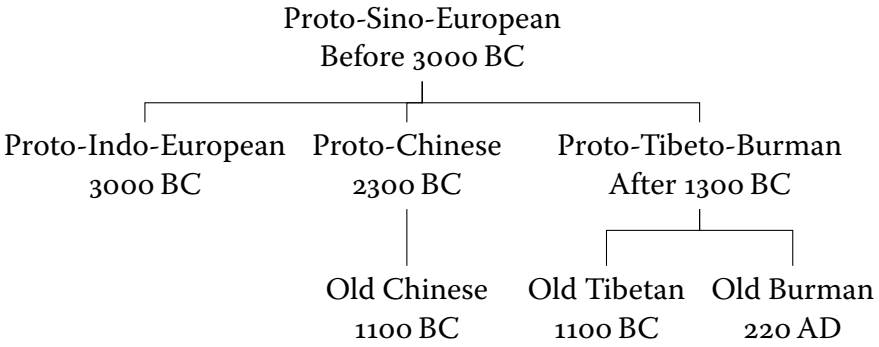

Figure 7: Zhou's Proposed Relation of PST and PIE

Later studies were more specific, since Proto–Indo-European is a broad concept. Mainly, there are two directions: studying the existence of Finno-Sinic family, and the genetic relations between Old Chinese, Tocharian and other Centum languages.

Gao[11] used SWADESH-100 as the range of vocabulary studied. He admitted that using OC reconstruction, whose validity and corresponding time period were unclear, to propose the family was relatively unsound in logic, stating that his study was partly to 抛砖引玉 (make some simple introductory remarks to set the ball rolling) and to attract more scholars to research in this new field.

Nonetheless, his study was rigorous, especially when proposing phonological rules between Old Estonian and Old Chinese. His material included Estonian, Finnish, Hungarian, Chinese and Tibetan, depicting a comprehensive system.

He also discussed "exact ratio" (in which only the most frequently used roots are examined) and "extended ratio" (examining cognates in all possible terms), and concluded that the exact ratio of Chinese and Hungarian was about 20%, and the extended ratio of Tibetan and Estonian was about 60%.

The other direction is a more researched field. As Li Yan and Li Baojia [13] stated in their paper, the Tocharian languages discovered in the late 19th century in Xinjiang was a great bond between Chinese and Centum languages – Tocharian itself was a mysterious exclave of the Centum family. The researches on Sino-Centum relations triggered by its discovery soon revealed possibilities and evidence of co-origination.

From the hypothesis of Rudbeck [21] about Chinese and Germanic languages, to the studies of Edkins [7] in 1871, to Zhou's cognate list [10], comparative studies strongly supported the co-origination of at least part (Centum and Germanic) of Indo-European and Sinic languages.

The Chinese literati certainly would not have imagined that their language contained vocabulary from so many sources.

4 Conclusion

The "orthodoxy" of phonology had always been a great concern of traditional Chinese literati, as they studied the ancient rhyming texts and created their own works. Without taking in the biases originating from the pursuit of "orthodoxy," this essay tried to reconstruct the literati's criterion using modern linguistic concepts such as contrastivity, as well as mathematical tools, and finally reached to a conclusion of which present Chinese variety would be considered orthodox by the literati. Following that, this essay also reviewed the literature about the lexical composition of Chinese, discussing lexical "orthodoxy" from a higher viewpoint.

In the first part, by proposing the contrastive phoneme test and using Pearson's correlation coefficient, we calculated "orthodoxy" for 13 regionalects, comparing their tones, initials and finals. The results accorded with people's general impression that southern varieties preserve Middle Chinese phonologies, and the hypothesis of the genetic relationship between Hakka and Gan. The sound results showed that this was indeed a comprehensive analysis of the phonologies of Chinese varieties.

From the literature reviews, we see that the lexical composition of Chinese, to the contrary of literati's will, is in fact complicated and non-unitary. The Bai Yue descendants Wu, Yue and Min are, after all, of Kam-Tai origin, and Old Chinese was merely their superstratum. A number of substratal vocabularies are still retained to this day. Also, since Tang dynasty, the Turkic influence on Chinese vocabulary became more powerful because the emperors all had Altaic bloodlines; a notable example is the word 哥 (brother). Lastly, the review of 4 articles and books revealed the credible co-origination of Chinese and some Indo-European languages; 3 other researches are also mentioned in the end. This is a field that goes beyond the discussion of "orthodoxy," yet it provides a broader vision for us to discuss the origin of Chinese and its vocabulary.

References

[1] Wolfram Research, Inc., Wolfram|Alpha Knowledgebase, Champaign, IL (2019).

[2] Li Qingzhao 李清照 (1084-1155). 李清照集 *Collection of Li Qingzhao*. 商务印书馆 The Commercial Press, 2007. ISBN: 9787807291206.

[3] William H. Baxter. *A Handbook of Old Chinese Phonology*. Berlin: Mouton de Gruyter, 1992. ISBN: 9783110123241.

[4] David P. Branner. "The rime-table system of formal Chinese phonology". In: *Auro/ua (hgg.): Geschichte der Sprachwissenschaften. Ein internationales Handbuch zur Entwicklung der Sprachforschung von den Anfängen bis zur Gegenwart. Bd* 1 (2000), pp. 46–55.

[5] Tsung tung Chang and Victor H. Mair. "Indo-European Vocabulary in Old Chinese: A New Thesis on the Emergence of Chinese Language and Civilization in the Late Neolithic Age". In: *Sino-Platonic Papers* (1988).

[6] Teaching and Research Office of Chinese Language and Linguistics of Peking University. 汉语方音字汇 *Hanyu Fangyin Zihui*. 北京：文字改革出版社 Beijing: Language Reform Press, 2003. ISBN: 9787801840349.

[7] Joseph Edkins. *China's Place in Philology: an Attempt to show that the Languages of Europe and Asia have a Common Origin*. Luzac, 1871.

[8] Xiao-shan Huang. "The Division II Medial and the Amount of Vowel in Middle Chinese". In: 浙江大學學報（人文社會科學版）*Journal of Zhejiang University (Humanities and Social Sciences)* 32.1 (2002).

[9] Guillaume Jacques. "Traditional Chinese Phonology". In: *Encyclopedia of Chinese Language and Linguistics* (2015). Ed. by Rint Sybesma.

[10] Zhou Ji-xu. 汉语印欧语词汇比较/汉语史研究丛书 *Hanyu Yin'ou Cihui Bijiao*. 四川民族出版社 Sichuan People's Publishing House, 2002.

[11] Gao Jingyi. *Comparison of Swadesh 100 words in Finnic, Hungarian, Sinic and Tibetan: Introduction to Finno- Sinic languages*. Talin: Hugarian Fund, 2005.

[12] Jingzhong Li. "粤语中的百越语成分问题 On the Bai Yue Elements in Yue Chinese". In: 学术论坛 *Xueshu Luntan* 88 (1991).

[13] Yan Li and Baojia Li. "The Role of Tocharian in Establishing the Relationship between Chinese and Indo-European Languages". In: 语言科学 *Linguistic Sciences* (2011).

[14] Tzu-Lin Mei. "The 'Wu Dialect' of Southern Dynasties and the Origin of Modern Min". In: *Language and Linguistics* (2015).

[15] Tzu-Lin Mei. "Tones and Prosody in Middle Chinese and The Origin of The Rising Tone". In: *Harvard Journal of Asiatic Studies* 30 (1970), pp. 86–100.

[16] Yuanyao Meng. "A Research to the Homogenous Vocabulary of Ancient Luoyue Language and Chinese Language". In: *Guangxi Ethnic Researches* 136 (2017).

[17] Aleksandar Mikić. "Origin of the words denoting some of the most ancient Old World pulse crops and their diversity in modern European languages". In: *PLoS One* 7.9 (2012), e44512.

[18] Jerry Norman. *Chinese*. Cambridge: Cambridge Language Surveys, Cambridge University Press, 1988. ISBN: 9780521296533.

[19] Wuyun Pan. "吴语形成的历史背景 Historical Background of the Origin of Wu Group". In: 方言 *Dialects* (2009), pp. 193–203.

[20] Hsin-yi Peng. “汉语方言 ŋ-声母的脫落与新生 Addition and Loss of Initial ŋ- in Chinese Dialects”. In: 语言学论丛 *Series in Linguistics* 52 (2015).

[21] Olof Rudbeck. *Specimen usus linguæ gothicæ in emendis atque illustrandis obscurissimis quibusdam Sacræ Scripturæ locis...: addita analogia linguæ gothicæ cum sinica nec non finnonicæ cum ungarica*. impressum â Joh. Henr. Werner, 1717.

[22] Li Wang. 中国語言學史 *History of Chinese Languages*. 香港: 中國圖書刊行社 Hong Kong: China Book Press, 1984. ISBN: 9787309047219.

[23] Fei Wu. “An Experimental Study on the Perceptual Process of Mandarin Nasals by L1 and L2 Learners”. In: *Linguistic Sciences* (2016), pp. 268–279.

[24] Margaret Mian Yan. *Introduction to Chinese Dialectology*. Munich: LINCOM Europa, 2006. ISBN: 9783895866296.

[25] Hongming Zhang. “On the Origin of Chinese Tonal Prosody: Argumentation from a Case Study of Shen Yue’s Poems”. In: *Journal of Chinese Literature and Culture* 2.2 (2015), pp. 347–379.

[26] Shangfang Zhengzhang. 上古音系 *The Old Chinese Phonology*. Shanghai: Shanghai Educational Publishing House, 2003. ISBN: 9787544446754.

[27] Shangfang Zhengzhang. 廣韻擬音 *Reconstruction of Guangyun*. Ed. by Jerry You. 古今文字集成 The Complete Collection of Ancient and Modern Characters, 2019.

[28] Ji-xu Zhou. “The Correspondent Words between Chinese, Tibetan, Burman and Indo-European Languages and their Position in Chinese Language History”. In: *Studies in Language and Linguistics* 30.4 (2010), pp. 23–27.

[29] 廣韻 *Guangyun*. 2007. URL: http://www.pkucn.com/viewthread.php?tid=175767&extra=page%3D1&page=2 (visited on 01/19/2019).

[30] Shen Yue 沈約 (441-513). 宋書 *History of Song*. Beijing: Zhonghua Shuju, 1974.

[31] 顏氏家訓 *(The Yan Family Mottoes)*. Vol. 音辭 Phonology. 420–581.