OCP Avoidance in Classical Chinese: Implications for Tonogenesis

Jack Isaac Rabinovitch
Harvard University

1 Introduction

Historical reconstructions of Chinese are often divided into two periods: Middle Chinese, which maintains tonal contrasts, an aspect of all modern Chinese dialects, and Old Chinese, which is reconstructed as a toneless language with so-called 'post-codas'. Post-codas were segments that made up the coda of a Chinese syllable, which caused tonal contours and were subsequently lost as a result of tonogenesis between Old Chinese and Middle Chinese (Baxter 1992; Baxter & Sagart 2014). While changes in grammar and adoption of vocabulary are relatively easy to detect and date, the actual time of tonogenesis (where tones developed and post-coda segments were lost) is difficult to pinpoint in a language like Chinese, where the logographic script gives no hints of sound change.

While the Chinese script provides little to work with in terms of phonological chronology, the texts themselves may help to hint at the phonology of their authors. Recent work in synchronic phonology has found that variation in word order and word choice is often influenced by the phonology of the language. As noted by Breiss & Hayes (2020), bigrams which cause violations of certain constraints at word boundaries tend to be avoided; this includes constraints against certain stress assignment (such as *Iambic Clash), constraints against disagreement in features such as *NC, and constraints derivative of the Obligatory Contour Principle (OCP) such as OCP[Sibilant]. This phenomenon is seen throughout morphosyntactic alternation across languages, including the genitive construction in English (Shih et al. 2015), noun-adjective order in Tagalog (Shih & Zuraw 2017), and noun-modifier order in Rigvedic Sanskrit (Gunkel & Ryan 2011, 2015). If Classical Chinese is no exception to this, then the proportion of bigrams which violate certain phonological constraints should be rarer than expected by chance. For Chinese at a pre-tonogenesis stage, post-codas, which are located at the end of syllables, would cause or block certain constraints to and from being violated at cross-word boundaries, and thus would affect the syntax and diction of the texts they make up. The lack of post-codas in a post-tonogenesis stage of Chinese would likewise influence the syntax and diction of a given text.

In this paper, I provide an account of OCP avoidance in Classical Chinese. I analyze a corpus of five pre-Qin (before 221 BCE) texts, and find that for bigrams for which the consonant cluster of the word boundary would not be affected by tonogenesis, pseudogeminate clusters, which would violate OCP constraints on features of place and manner of articulation, are significantly ($p < 0.00001$) less common than expected through randomization of characters. I then analyze all bigrams, including those which would be affected by tonogenesis. This paper considers three hypotheses regarding the state of tonogenesis for the analyzed works: 1) the Atonal Hypothesis, where post-codas were present during the time that these works were written and should thus influence OCP restrictions, 2) the Tonal Hypothesis, where post-codas were not present and so would not impact OCP constraint violations, and 3) the Non-Glottal Hypothesis, where post-codas were present but where glottal stop post-codas do not affect OCP constraint violations. I find that across all texts, under any of the three hypotheses, OCP violations are significantly avoided in both poetry and prose; however, only under the Atonal and Non-Glottal Hypotheses are OCP violations avoided among bigrams which would have historically had an intervening post-coda segment. This suggests that at the time of the writing of the texts in the corpus, post-coda segments were still present in Chinese. In addition, this paper finds that poetic

[*] I would like to thank Zhongwei Shen for his wisdom in historical Chinese phonology and for providing inspiration in taking on this project, Kevin Ryan for his encouragement and invaluable guidance in both phonology and statistics, and the organizers and attendees of the 2020 Annual Meetings on Phonology at University of California, Santa Cruz for their helpful comments and insight during the presentation of these findings. All mistakes are my own.

© 2021 Jack Isaac Rabinovitch
Proceedings of *AMP 2020*

texts avoid OCP violations more than prose texts, suggesting poetic sensitivity to OCP violations.

1.1 *The Corpus* Five works make up the corpus of this study. The Book of Odes (詩經 *Shījīng*, hereafter abbreviated SJ) is a compilation of 305 poems dating from the Shang (1600 – 1046 BCE) and Western Zhou (1046 – 771 BCE) from various states in the Kingdom of Zhou. Songs of the South (楚辭 *Chǔcí*, hereafter abbreviated CC) is another anthology of Chinese poetry dating from the Warring States period (475 – 221 BCE), though according to Tz'ù & Hawkes (1959) parts of it may date from the Han Dynasty (202 BCE – 220 CE); Songs of the South has its origin in the State of Chu (楚 *Chǔ*). The Commentary of Zuo (左傳 *Zuǒzhuàn*, hereafter abbreviated ZZ) is a narrative history dated from at latest 300 BCE (Durrant et al. 2016), during the the Warring States period (475 – 221 BCE). Classic of Mountains and Seas (山海經 *Shānhǎi Jīng*, hereafter abbreviated SHJ) is a bestiary and compilation of mythic geography. The oldest surviving version is dated to the Han Dynasty (202 BCE – 220 CE), though Lust (1996) notes that it most likely originated during the Warring States period (475 – 221 BCE). The Art of War (孫子兵法 *Sūnzi Bīngfǎ*, hereafter abbreviated AoW) is a military and philosophical treatise dating from the 5th century BCE, during the Spring and Autumn period (776 – 471 BCE). All of these texts were provided by Sturgeon (2019)'s Chinese Text Project.

(1)

Abbr.	English	Chinese	Pinyin	Date	Genre
SJ	The Book of Odes	詩經	*shījīng*	Oldest	Poetry
AoW	Art of War	孫子兵法	*sūnzi bīngfǎ*	Intermediate	Prose
ZZ	The Commentary of Zuo	左傳	*zuǒzhuàn*	Newest	Prose
CC	Songs of the South	楚辭	*chǔcí*	Newest	Poetry
SHJ	Classic of Mountains and Seas	山海經	*shānhǎi jīng*	Newest	Prose

The corpus ranges in time, genre, and location of origin. For our purposes, we will only look at the difference between poems and prose, as well as the relative times that the works were produced. This is shown in (1), alongside the abbreviations used in the rest of the paper.

Reconstructions in this paper will follow Baxter & Sagart (2014) (abbreviated B&S) for Old Chinese and Zhèngzhāng (2003) (abbreviated ZZSF) for Middle Chinese. In Zhèngzhāng (2003)'s reconstruction, Middle Chinese tones are denoted with superscript letters, a superscript X represents a *shǎng* 上 'rising' tone and a superscript H represents a *qù* 去 'departing' tone. Middle Chinese reconstructions without a superscript letter are *rù* 入 'entering' tone if they end in -p, -t, or -k; otherwise they are *píng* 平 'level' tone. Following Shen (2020)'s standard, Zhèngzhāng (2003)'s reconstructions are considered 'half reconstructions', as they are phonetic reflections of known categorial differences, and are not written with an asterisk.

2 Reconstruction and Tonogenesis

Kingston (2011) discusses tonogenesis through a process of 'Exaggeration and Transfer'; Exaggeration is a process whereby segments in a language phonetically influence F0 contour and are exaggerated to the point of being part of the phonology of the language, albeit predictable by the segments they co-occur with, and Transfer is the process whereby the segments which cause tonal contour undergo elision or merge with other segments, leaving the contours to be completely contrastive, and making the language tonal. This is shown in (2). The process thus has three stages, Stages I, II, and III, as defined in (3).

(2)

Stage I	Exaggeration	Stage II	Transfer	Stage III
ABC	→	ABC + Contour	→	AB + Tone 1
AB		AB + No Contour		AB + Tone 2

Modern examples of Stage II can be seen in Yabem (Ross 1993) and Korean (Cho 2017), where tone contour is completely regular and predictable but distinct enough that it is most likely part of the phonological system, rather than incidental from the phonetics. Western and Eastern Cham are two Chamic languages where certain segments in Western Cham have tonal equivalents in Eastern Cham, betraying that Eastern Cham has already undergone 'transfer' and has arrived at the Stage III of tonogenesis (Phu et al. 1997).

(3)
a. Stage I: A language has some set of segments X with no influence on tone.
b. Stage II: The set of segments X cause a predictable tonal contour on their syllables.
c. Stage III: The segments X are no longer contrastive; their resulting tonal contours are contrastive.

The introduction of tones through tonogenesis is one of the major innovations reconstructed in Middle Chinese, as opposed to its toneless ancestor of Old Chinese. The phonology of Middle Chinese is represented by the *Qièyùn* (601 CE) rime dictionary, and its grammar is represented by works produced between the Northern and Southern Dynasties and the Song dynasty (386 – 1279 CE). Tone in Middle Chinese is backed by historical record, where the *Qièyùn* explicitly discusses the four tone system of Middle Chinese. Old Chinese is represented by inscriptions dating from the Late Shang dynasty to the Warring States period (1250 – 221 BCE). Old Chinese phonology has been reconstructed through analysis of modern Chinese dialects, reconstructions of Middle Chinese provided by aforementioned phonological works, and related languages such as Tibetan and Bai. Unlike Middle Chinese, there are no contemporary texts which discuss the phonology of Old Chinese.

2.1 *The Old Chinese Coda* Following Haudricourt (1954)'s analysis of tonogenesis in Vietnamese, Old Chinese reconstructions include so called 'post-codas' as the segments which underwent tonogenesis and loss to become the tones which are present in Middle Chinese, as solidified in Baxter (1992) and continued in the most recent prominent works such as Baxter & Sagart (2014). The Old Chinese coda consisted of the aforementioned optional 'post-coda' as well as another optional segment preceding the post-coda often simply called the 'coda'. To avoid confusion between 'coda' referring to the entire coda (including post-coda) or just the non-post-coda part of the coda, I will refer to the segment preceding the post-coda as the 'pre-coda'. The options for each of these are seen in (4).

(4) a. 'pre-coda': /w, j, m, n, ŋ, p, t, k/
b. 'post-coda': /s, ʔ/

In Old Chinese reconstructions, any post-coda can co-occur with any pre-coda segment except for glottal stops after plosives (p, t, k). While *-ps, *-ts, and *-ks are reconstructed, it is assumed that the stops in these codas underwent deletion and became *qù* 去 'departing' tone syllables with zero codas by Middle Chinese, and thus are reconstructed in Middle Chinese as coda-less syllables.[1]

2.2 *From Old to Middle Chinese* The sound changes which resulted in tonogenesis between Old and Middle Chinese are shown in (5), where each segment represents the last segment in a syllable. For the 'otherwise' case, the final segment (denoted Z) may contain any of the sonorant 'pre-coda' segments (w, j, m, n, ŋ), or in the case of syllables lacking a coda, the nucleus.

(5)

Old Chinese Segment	Stage II		Stage III		Middle Chinese Tone
(Stage I)	segment	tone	segment	tone	
ʔ]$_\sigma$	ʔ	X		X	*shǎng* 上 'rising'
s]$_\sigma$	s	H		H	*qù* 去 'departing'
p]$_\sigma$	p		p		*rù* 入 'entering'
t]$_\sigma$	t		t		
k]$_\sigma$	k		k		
Otherwise Z]$_\sigma$	Z		Z		*píng* 平 'level'

The deletion of final consonants only occurs for *-s and *-ʔ. At first this would seem to mean that the 'transfer' stage is extremely limited. However, because these post-codas can co-occur with all *píng* 平 'level' tone inducing pre-codas (w, j, m, n, ŋ), the result is that *píng* 平 'level', *shǎng* 上 'rising', and *qù* 去 'departing' tones can form minimal triplets distinguishable by tone alone.[2] For example, in (6), three characters are shown, *rén* 壬 'ninth celestial stem', *rěn* 妊 'pregnant', and *rèn* 荏 'perilla frutescens'. In Stage I of tonogenesis (represented by Old Chinese), these characters differed in pronunciation by post-coda alone. In Stage II, the characters with post-codas gain tonal contours, predictable with the given post-coda. In Stage III, represented by Middle Chinese, the post-coda segments have disappeared and the tone is now the only distinguishing feature among the minimal triplet.

[1] See Shen (2020) for a discussion on the reconstruction of these codas and their additional sound changes between Old and Middle Chinese.

[2] *Rù* 入 'entering' tone characters can be distinguished by tone and by endings in Middle Chinese; this tone/ending predictable co-occurance still exists in many dialects of Chinese (though not Standard Mandarin).

	Char.	Stage I (B&S)	Stage II	Stage III (ZZSF)	Middle Chinese Tone	Mandarin (Pinyin)	Definition
(6)	壬	*n[ə]m	*n[ə]m	ȵiɪm	*píng* 平 ‘level’	*rén*	‘ninth celestial stem’
	妊	*n[ə]mʔ	*n[ə]mʔ^X	ȵiɪm^X	*shǎng* 上 ‘rising’	*rěn*	‘pregnant’
	荏	*n[ə]ms	*n[ə]ms^H	ȵiɪm^H	*qù* 去 ‘departing’	*rèn*	‘perilla frutescens’

The exact identity of these tones is not completely known, although given their names and descriptions at the time, it is likely that *píng* 平 ‘level’ tone was a level tone, *shǎng* 上 ‘rising’ tone had a rising contour, *qù* 去 ‘departing’ tone had a falling contour, and *rù* 入 ‘entering’ tone was a short or ‘checked tone’ syllable (Shen 2020). The basis for the assumption that Middle Chinese was already at Stage III of tonogenesis comes from the aforementioned written record of the tonal system in phonological works which date back to the Sui dynasty (581 CE – 618 CE), such as the *Qièyùn* 切韻 (601 CE), which distinguishes the four tone system as used in Zhèngzhāng (2003) and shown in (5). However, because tone may be part of the phonology of a language even when it does not form minimal pairs (Ross 1993; Cho 2017), it is possible that the tones of the *Qièyùn* co-occured with post-coda segments, as in Stage II, and that the point in time when post-coda segments were lost came long after the writing of the *Qièyùn*. Thus, for any given text written before the *Qièyùn*, the language of the text may belong to any of the three stages of tonogenesis.

Due to the unidirectionality of tonogenesis (one cannot regain post-codas after they are lost), we should also expect that if there is variation between the stages of tonogenesis between these works, that the older works will be more conservative in keeping post-coda segments, while newer works will be less likely to have post-codas, though regional variation allows a later text to conserve features that an earlier text from a different area or dialect lost.

3 The OCP and Capturing Sound Change

The Obligatory Contour Principle (OCP) is a principle in which languages avoid or ban adjacent segments which are too similar in certain features (Leben 1973, 1978; McCarthy 1986). Under an optimality theoretic (OT) framework, the OCP can be represented by a series of violable constraints (Prince & Smolensky 1993; Myers 1997). While an underlying representation which violates the OCP may be ‘fixed’ through a mismatch between the underlying representation and the surface form, a high ranking Faithfulness constraint may result in the OCP violation continuing to the surface form, with no modification. Importantly, while a given language may allow OCP violations, there may still be a pressure to correct these violations. Among typical synchronic sound changes such as devoicing and assimilation, recent work (Shih et al. 2015; Shih 2017) has found that choice of vocabulary and change in syntax may be utilized as a way to avoid OCP violations, particularly in poetic form, where word choice and order are more free, and where sensitivity to phonological constraints may be heightened (Shih & Zuraw 2017; Breiss & Hayes 2020).

3.1 *Different Times, Different Violations* Depending on the stage of tonogenesis, a bigram may or may not violate certain OCP constraints. Let us take for example, *PseudoGem, an OCP constraint which is violated by identical adjacent segments across a word boundary (eg. “bus stop”, where two ‘s’ segments are adjacent). We may compare the two bigrams *fàng xīn* 放心 ‘be at ease’ and *fàng yǎn* 放眼 ‘take a broad view’ as in (7). In Old Chinese, *fàng xīn* 放心 (B&S: *paŋ-s səm) forms a pseudogeminate due to the presence of the post-coda *-s in *fàng* 放 (B&S: *paŋ-s) being adjacent to the initial *s in *xīn* 心 (B&S: *səm), while post-coda *-s in *fàng yǎn* 放眼 (B&S: *paŋ-s [ŋ]ˤ<r>ə[n]ʔ) intervenes and prevents a pseudogeminate from forming between the two velar nasals. In Middle Chinese, without the post-coda *-s in *fàng* 放 (ZZSF: pʉɐŋ^H), no pseudogeminate forms in *fàng xīn* 放心 (ZZSF: pʉɐŋ^H siɪm), but does form in *fàng yǎn* 放眼 (ZZSF: pʉɐŋ^H ŋɣɛn^X), where now no segment intervenes between the two velar nasals.

	Pinyin	Char.	*PseudoGem	Old Chinese		*PseudoGem	Middle Chinese
(7)	*fàng xīn*	放心	*	*paŋ-s səm	→		pʉɐŋ^H siɪm
	fàng yǎn	放眼		*paŋ-s [ŋ]ˤ<r>ə[n]ʔ	→	*	pʉɐŋ^H ŋɣɛn^X

As suggested by Gunkel & Ryan (2015), just as OCP violations motivate avoidance of geminates within a word, OCP violations may motivate avoidance of pseudogeminates. Lets take for example the Old Chinese in (8). *NounAdj, which is violated by having adjectives after nouns, represents a syntactic constraint. Faith[Word] represents a lexicon choice constraint, and is violated by changing word choice (say, among a choice of synonyms) between the underlying and surface representations. Here the poet wants to write something with the meaning of "dark star". The underlying form is the grammatically unmarked choice (Adjective-Noun) with the closest lexical choices to create the intended meaning: *àn* 暗 (B&S: *q^{ʕ}ums, ZZSF: ʔʌm^{H}) meaning 'dark' and *xīng* 星 (B&S: *sts^{hʕ}eŋ, ZZSF: seŋ) meaning 'star'. In order to avoid the pseudogeminate generated by the faithful candidate, the poet may change the word order, violating the lower ranked *NounAdj, or the poet may change the word choice, choosing a near-synonym, such as *hēi* 黑 (B&S: *m̥^{ʕ}ək, ZZSF: hək), meaning 'black' as a substitute for 'dark', creating a Faith[Word] violation.

(8) Old Chinese:

暗星 /q^{ʕ}ums sts^{hʕ}eŋ/				*PseudoGem	*NounAdj	Faith[Word]
a.		暗星	[q^{ʕ}ums sts^{hʕ}eŋ]	*!		
b.	☞	星暗	[sts^{hʕ}eŋ q^{ʕ}ums]		*	
c.	☞	黑星	[m̥^{ʕ}ək sts^{hʕ}eŋ]			*

In (8) our problematic character, *àn* 暗 'dark', has an *-s post-coda. This is what results in the pseudogeminate forming, which can be avoided by employing marked word choice or word order. However, as shown in (9), the tonogenesis and loss of post-coda segments in Middle Chinese results in this pseudogeminate never forming, and thus there is no motivation to change word order or word choice.

(9) Middle Chinese:

暗星 /ʔʌm^{H} seŋ/				*PseudoGem	*NounAdj	Faith[Word]
a.	☞	暗星	[ʔʌm^{H} seŋ]			
b.		星暗	[seŋ ʔʌm^{H}]		!*	
c.		黑星	[hək seŋ]			!*

As seen in (7), the reverse may happen as well, where pressure to use marked word choice or word order may be present only in the Middle Chinese pronunciation of a bigram, such as *fàng yǎn* 放眼.

3.2 *Three Hypotheses of OCP Interactions* In theories where the glottal stop is featureless, we might expect it to act transparently with respect to constraints concerned with featural similarity; as a result, Cʔ#C sequences such as that in *shǎn míng* 陕明 (B&S: *hljemʔ mraŋ, ZZSF: ɕiᴇm^{X} mɣiæŋ), would be considered as violating *PseudoGem in both Middle and Old Chinese. So, assuming that pseudogeminates are avoided in Classical Chinese (as will be borne out in Section 5.1), there are three possibilities, or hypotheses for avoidance patterns of pseudogeminates in a given text (10).

(10) a. **Atonal Hypothesis:** Post-codas are present and affect OCP (compatible with Stages I & II). We expect avoidance of bigrams where the post-coda of the first character matches in features with the initial of the second character, as well as avoidance of bigrams where the coda of the first character lacks a post-coda, and the pre-coda matches in features with the initial of the second character.

b. **Non-Glottal Hypothesis:** Post-codas are present and all but glottal stops affect OCP (compatible with Stages I & II). We expect avoidance of bigrams where an *-s post-coda of the first character matches in features with the initial of the second character, as well as avoidance of bigrams where the coda of the first character either lacks a post-coda or has a *-ʔ coda, and the pre-coda matches in features with the initial of the second character.[3]

c. **Tonal Hypothesis:** Post-codas are not present (compatible with Stage III). We expect avoidance of bigrams where the pre-coda of the first character matches in features with the initial of the second character.

Reconstructions of Old Chinese additionally have differences in onsets form Middle Chinese, which may affect which bigrams are considered to have OCP violations. Due to the common lack of clarity concerning these onset segments, however, this analysis will refer to the broad categories of Middle Chinese onsets, rather than rely completely on reconstructed Old Chinese segments. Thus, this paper tests for OCP effects

[3] Another interpretation of the Non-Glottal Hypothesis may be one where *-ʔ post-codas have undergone elision (transfer) while *-s have not.

using Middle Chinese reconstructions with the addition of Old Chinese post-coda segments for variation in hypothesis. An additional survey analyzing OCP effects on any tone that may have existed at the time may be useful for teasing out whether or not tonogenesis had begun by the time that these works were written. However, a test like this would have to make guesswork out of the features assigned to each of the tones, and how they would factor into OCP violations.

4 Finding Pseudogeminates

Each character in the texts are assigned to their Middle Chinese initial and coda-tone combination (rewritten as their equivalent Old Chinese coda), with the phonological data taken from Zhèngzhāng (2003). Zhèngzhāng (2003)'s reconstructions are based on the *Guǎngyùn* (1008 CE), a rime dictionary, like the *Qièyùn*, although written significantly later during the Song dynasty (960 – 1279 CE). Despite this difference in time, the *Guǎngyùn* maintains all of the contrasts found in the *Qièyùn*, making it just as conservative. The *Qièyùn* itself is not completely intact, and so using the *Guǎngyùn* is the closest to getting an accurate phonological system of Middle Chinese. Glide codas, which are difficult to distinguish from vowels, are considered non-codas for the purposes of this study.

If a character has multiple readings, the different initials and codas are recorded. Some characters do not have Zhèngzhāng (2003) reconstructions, and so are left "NA". Every character token is assigned the initial and coda readings from its given entry/entries. For example, (11) shows characters 17–24 of the poem *Guān Jū* 關雎, the first poem of the Book of Odes.

(11)

Character	參	差	荇	菜	左	右	流	之
Initials	ʃ, ts^h, tʃ^h	tʃ^h	NA	ts^h	ts	ɦ	l	tɕ
Codas	m	∅,s	NA	s	s,ʔ	s,ʔ	∅	∅

Pseudo-geminates and other OCP violations at word boundaries concern two-word sequences, or bigrams. Thus, our analysis of the corpus will focus on individual bigrams present throughout each text. Bigrams are only considered by the coda of the first character and the initial of the second character of a bigram, as this is what determines the possible segment-based OCP interactions between them. (12) shows the same slice from *Guān Jū* with bigrams instead of character tokens.

(12)

Character	參差	差荇	荇菜	菜左	左右	右流	流之
First Character Coda	m	∅,s	NA	s	s,ʔ	s,ʔ	∅
Second Character Initial	tʃ^h	NA	ts^h	ts	ɦ	l	tɕ

4.1 *Randomization* A baseline of expected bigram frequency is calculated through randomization of characters within each text. The expected frequency of a given bigram type (Coda # Initial combination) can be found by the frequency of a given coda type multiplied by the frequency of a given initial type. (13) is a simplified example of this, in which only consider initials b, p, m and codas ms, m, p for AoW.

(13)

Char 1 Coda	m			ms			p		
Char 2 Initial	b	m	p	b	m	p	b	m	p
Coda Count	49	49	49	7	7	7	90	90	90
Initial Count	105	191	477	105	191	477	105	191	477
Product Count	5145	9359	23373	735	1337	3339	9450	17190	42930
Expected Freq.	0.046	0.083	0.21	0.0065	0.012	0.03	0.084	0.15	0.38

4.2 *Counting Pseudogeminate Violations* For the purposes of this paper, pseudogeminates need not be exactly identical adjacent segments, but rather may be adjacent segments which match in both place and manner of articulation. A given bigram where the first character's coda matches in these features with the second character's initial will be said to violate *PseudoGem (14).

(14) *PseudoGem: Assign one violation for each pair of adjacent segments which match in both (broad) place features and manner of articulation features.

In order to calculate which bigrams violate *PseudoGem, each segment is assigned a category, such that adjacent segments which belong to the same category can be said to violate *PseudoGem. For initials, segments are assigned their category as in (15), where each traditional Chinese initial (categories derived from the groupings of homophones within the *Guǎngyùn*) is shown next to its ZZSF reconstruction, and OCP category. Labial stops are represented by the label P; coronal stops (including affricates) are represented by the label T; velar stops are represented by the label K; the labial nasal (m) is represented by the label M; coronal nasals (ȵ,n,ɲ) are represented by the label N; the velar nasal (ŋ) by the label Ŋ; sibilants are labelled S; the glottal stop is labelled ʔ, and others (glottal fricatives and non-nasal liquids) are labelled "other".

(15)

Category	P	T					K	M	N	Ŋ	S		ʔ	other	
Initial	幫	端	知	精	章	莊	見	明	泥	疑	心	邪	影	曉	以
	p	t	ʈ	ts	tɕ	tʃ	k	m	n	ŋ	s	z	ʔ	h	j
	滂	透	徹	清	昌	初	溪		日		書	船		云	來
	p^h	t^h	ʈ^h	ts^h	tɕ^h	tʃ^h	k^h		ȵ		ɕ	ʑ		ɦ	l
	並	定	澄	從	常	崇	群		娘		生	俟		匣	
	b	d	ɖ	dz	dʑ	dʒ	g		ɲ		ʃ	ʒ		ɦ	

A given character may correspond to multiple readings with different initials. In these cases, whether or not an OCP violation is present may be difficult to discern, where the intended reading is unclear. If all of a given character's possible initial readings belong to one of these categories, then it is considered to be in that group. Otherwise, if readings vary across groups, then the character is considered a "mixed initial".

Coda segments will have different pseudogeminate categories depending on the hypothesis being tested. For the Atonal Hypothesis, characters with *-s and *-ʔ post-codas will belong to the S and ʔ categories respectively. For the Tonal Hypothesis, only the pre-coda part of the coda section will be considered for category, and for the Non-Glottal Hypothesis, only *-s post-codas will be considered for determining OCP category. I provide the correspondence table in (16). The rare *-ps, *-ts, and *-ks codas pattern with *-s. Because syllables must have onsets in Middle Chinese, the ∅ category is only available for codas.

(16)

Coda	Atonal	Non-Glottal	Tonal	Middle Chinese Tone
∅	∅	∅	∅	*píng* 平 'level'
m	M	M	M	
n	N	N	N	
ŋ	Ŋ	Ŋ	Ŋ	
ʔ	ʔ	∅	∅	*shǎng* 上 'rising'
mʔ	ʔ	M	M	
nʔ	ʔ	N	N	
ŋʔ	ʔ	Ŋ	Ŋ	
s	S	S	∅	*qù* 去 'departing'
ms	S	S	M	
ns	S	S	N	
ŋs	S	S	Ŋ	
p	P	P	P	*rù* 入 'entering'
t	T	T	T	
k	K	K	K	

Just as a given character may correspond to multiple readings with different initials, a character may also correspond to multiple codas. If all of a given character's possible coda readings belong to the same category, then it is considered to be in that category. Otherwise, if readings vary across groups, then the character is considered a "mixed coda". As can be seen in the categories in (16), *píng* 平 'level' and *rù* 入 'entering' tone codas do not change their category depending on the hypothesis.

5 Testing for OCP Violation Avoidance

5.1 *Theory-neutral Avoidance* Before testing out which hypothesis best describes the data of the text, we must first determine whether or not OCP violations are avoided in Classical Chinese at all. Because *píng* 平 'level' and *rù* 入 'entering' tone characters lack post-coda segments in Old Chinese, the effect of their codas on OCP violations should be the same regardless of what stage of tonogenesis a given text is in. Thus, character bigrams where the first character is *píng* 平 'level' or *rù* 入 'entering' tone provide a good theory-neutral testing ground for whether OCP effects are significant at all in Pre-Qin Chinese under any theory.

To see the significance of OCP effects on *píng* 平 'level' and *rù* 入 'entering' tone characters, bigrams for which the first-character has a *shǎng* 上 'rising' or *qù* 去 'departing' tone reading — essentially, those which have post-codas present in at least one of their readings — are ignored. Additionally, bigrams with first character mixed-coda, second character mixed-initials, or with NA segments are removed before calculating percentages. *PseudoGem violations were counted for any bigram token whose first character's coda category matched their second character's initial category. In (17), the number of theory-neutral bigrams which violate *PseudoGem as well as the total number of theory-neutral tokens in each book is recorded. The rate of *PseudoGem violations is checked against two expected rates, the first 'Against Local' is the expected rate of *PseudoGem violations by randomization of characters within the specific text, while the second 'Against Total' is against the expected rate of violations by randomization of characters throughout the entire corpus.

(17) χ^2 Test Results for *Píng* 平 'Level' and *Rù* 入 'Entering' Tones:

Book	Viol	Tokens	Rate	Exp. Rate	Against Local		Against Total	
					χ^2	p	χ^2	p
AoW	138	2140	0.0645	0.0694	0.79	0.37	< 0.05	0.96
SHJ	695	13987	0.0497	0.0630	42	< 0.00001	53	< 0.00001
ZZ	3509	67525	0.0520	0.0706	360	< 0.00001	180	< 0.00001
CC	437	11930	0.0366	0.0404	4.4	0.037	160	< 0.00001
SJ	501	10464	0.0479	0.0561	13	0.00025	49	< 0.00001
Total	5280	206046	0.0498	0.0648			390	< 0.00001

The actual rate of *PseudoGem violations in SHJ, ZZ, and SJ are all significantly ($p < 0.0005$) lower than both their local expected rates and the total expected rate. CC also has significantly fewer ($p < 0.00001$) *PseudoGem violations than the total expected rate. AoW has a lower actual rate than expected, though it tests non-significant against the local expected rate, and borderline ($0.05 < p < 0.01$) against the total expected rate. The entire corpus collectively also has significantly ($p < 0.00001$) fewer *PseudoGem violations than expected. These results strongly suggest that bigrams which resulted in *PseudoGem violations were avoided. As discussed in Section 3, this can be explained as the result of utilizing marked word order or choice.

5.2 *Testing Across Hypotheses* Now that it has been established that OCP effects are significant in all texts other than AoW for theory-neutral contexts, we can test all bigrams against each of our hypotheses. For a given hypothesis, bigrams with first character mixed-coda for that hypothesis, second character mixed-initials, or with NA segments are removed before calculating percentages. For each hypothesis *PseudoGem violations were counted for any bigram token whose second character's initial category matched their first character's coda category in their respective hypothesis (as assigned in (16)). For each hypothesis, a test was done considering only *shǎng* 上 'rising' and *qù* 去 'departing' tone characters (thus not including the already tested *píng* 平 'level' and *rù* 入 'entering' tone characters), and a test was done considering all characters.

5.2.1 *Results for the Atonal Hypothesis* When only considering *shǎng* 上 'rising' and *qù* 去 'departing' tone characters, the Atonal Hypothesis finds the two poetic works, CC and SJ, to have significantly fewer *PseudoGem violations than expected tested against their local randomizations, while AoW and SJ, the two oldest texts, test as having significantly ($p < 0.0005$) fewer *PseudoGem violations than expected tested against the entire corpus (18).

(18) χ^2 Test Results for Atonal Hypothesis, *Shǎng* 上 'Rising' and *Qù* 去 'Departing' Tones:

Book	Viol	Tokens	Rate	Exp. Rate	Against Local		Against Total	
					χ^2	p	χ^2	p
AoW	51	1462	0.0349	0.0455	3.8	0.051	14	0.00021
SHJ	343	5579	0.0615	0.0659	1.8	0.19	1.7	0.2
ZZ	2713	45942	0.0591	0.0567	4.8	0.028	2.1	0.15
CC	277	5177	0.0535	0.0634	8.6	0.0034	1.5	0.22
SJ	272	6818	0.0399	0.0508	17	< 0.00005	39	< 0.00001
Total	3656	64978	0.0563	0.0575			1.8	0.18

ZZ tests as having borderline significantly ($0.05 < p < 0.01$) more *PseudoGem violations than expected tested against their local randomization.

For the Atonal Hypothesis across all tones (19), all texts other than AoW test as having significantly ($p < 0.002$) fewer *PseudoGem violations than expected by both their local randomization and by randomization of the entire corpus. AoW tests as having borderline significantly ($0.05 < p < 0.01$) fewer *PseudoGem violations than expected by randomization of the entire corpus. The entire corpus collectively also has significantly ($p < 0.00001$) fewer *PseudoGem violations than expected.

(19) χ^2 Test Results for Atonal Hypothesis, All Tones:

Book	Viol	Tokens	Rate	Exp.	Against Local		Against Total	
				Rate	χ^2	p	χ^2	p
AoW	189	3602	0.0525	0.0586	2.4	0.12	5.7	0.017
SHJ	1038	19566	0.0531	0.0639	38	< 0.00001	27	< 0.00001
ZZ	6222	113467	0.0548	0.0651	200	< 0.00001	100	< 0.00001
CC	714	17107	0.0417	0.0468	9.8	0.0018	120	< 0.00001
SJ	773	17282	0.0447	0.0541	30	< 0.00001	89	< 0.00001
Total	8936	171024	0.0522	0.0620			280	< 0.00001

5.2.2 *Results for the Non-Glottal Hypothesis* When only considering *shǎng* 上 ‘rising’ and *qù* 去 ‘departing’ tone characters, the Non-Glottal Hypothesis has similar results to the Atonal Hypothesis, except that the SHJ is found to have significantly ($p < 0.0002$) more *PseudoGem violations than expected by randomization of the entire corpus (20).

(20) χ^2 Test Results for Non-Glottal Hypothesis, *Shǎng* 上 ‘Rising’ and *Qù* 去 ‘Departing’ Tones:

Book	Viol	Tokens	Rate	Exp.	Against Local		Against Total	
				Rate	χ^2	p	χ^2	p
AoW	41	1462	0.0280	0.0380	4	0.046	9.8	0.0017
SHJ	310	5579	0.0556	0.0516	1.8	0.18	14	0.00015
ZZ	2102	45942	0.0458	0.0458	< 0.01	0.99	0.55	0.46
CC	201	5177	0.0388	0.0485	10	0.0012	4.6	0.031
SJ	171	6818	0.0251	0.032	11	0.0012	63	< 0.00001
Total	2825	64978	0.0435	0.0450			3.7	0.055

When all tones are tested (21), the results of the Non-Glottal Hypothesis is almost the same as that for the Atonal Hypothesis, except that AoW tests as non-significant against the expected rate of the entire corpus.

(21) χ^2 Test Results for Non-Glottal Hypothesis, All Tones:

Book	Viol	Tokens	Rate	Exp.	Against Local		Against Total	
				Rate	χ^2	p	χ^2	p
AoW	179	3629	0.0493	0.0550	2.3	0.13	3.7	0.056
SHJ	1022	19675	0.0519	0.0594	20	< 0.00001	8.2	0.0042
ZZ	5625	115322	0.0488	0.0601	260	< 0.00001	130	< 0.00001
CC	645	17618	0.0366	0.0417	12	0.00067	130	< 0.00001
SJ	680	17782	0.0382	0.0458	23	< 0.00001	110	< 0.00001
Total	8151	174026	0.0468	0.0567			310	< 0.00001

5.2.3 *Results for the Tonal Hypothesis* With respect to *shǎng* 上 ‘rising’ and *qù* 去 ‘departing’ tone, characters, the Tonal Hypothesis has no significant results in favor of fewer violations, and instead has four significant ($p < 0.001$) results in favor of more violations than expected. ZZ tests as having significantly ($p < 0.001$) more *PseudoGem violations than expected tested against both the local expected rate and the entire corpus expected rate; CC tests as having significantly ($p < 0.00005$) more *PseudoGem violations than expected when tested against the entire corpus expected rate. Additionally, the entirety of the corpus tests as having significantly ($p < 0.00001$) more *PseudoGem violations than expected.

(22) χ^2 Test Results for Tonal Hypothesis, *Shǎng* 上 'Rising' and *Qù* 去 'Departing' Tones:

Book	Viol	Tokens	Rate	Exp. Rate	Against Local χ^2	Against Local p	Against Total χ^2	Against Total p
AoW	18	1462	0.0123	0.0096	1.1	0.29	0.57	0.45
SHJ	47	5579	0.0084	0.0061	5.1	0.024	2	0.16
ZZ	584	45942	0.0127	0.0111	11	0.00080	26	< 0.00001
CC	84	5177	0.0162	0.0130	4.1	0.042	18	< 0.00005
SJ	69	6818	0.0101	0.0077	5.2	0.022	< 0.05	0.87
Total	802	64978	0.0123	0.0103			26	< 0.00001

When including *píng* 平 'level' and *rù* 入 'entering' tone characters, the results look almost identical to that of the Non-Glottal Hypothesis, except that CC additionally tests non-significant against the local predicted rate.

(23) χ^2 Test Results for Tonal Hypothesis, All Tones:

Book	Viol	Tokens	Rate	Exp. Rate	Against Local χ^2	Against Local p	Against Total χ^2	Against Total p
AoW	171	4228	0.0404	0.0393	0.15	0.70	< 0.05	0.87
SHJ	795	22407	0.0355	0.0428	30	< 0.00001	17	< 0.00005
ZZ	4547	135108	0.0337	0.0432	300	< 0.00001	180	< 0.00001
CC	596	20545	0.0290	0.0309	2.4	0.12	75	< 0.00001
SJ	629	20363	0.0309	0.0349	9.7	0.0019	53	< 0.00001
Total	6738	202651	0.0332	0.0410			310	< 0.00001

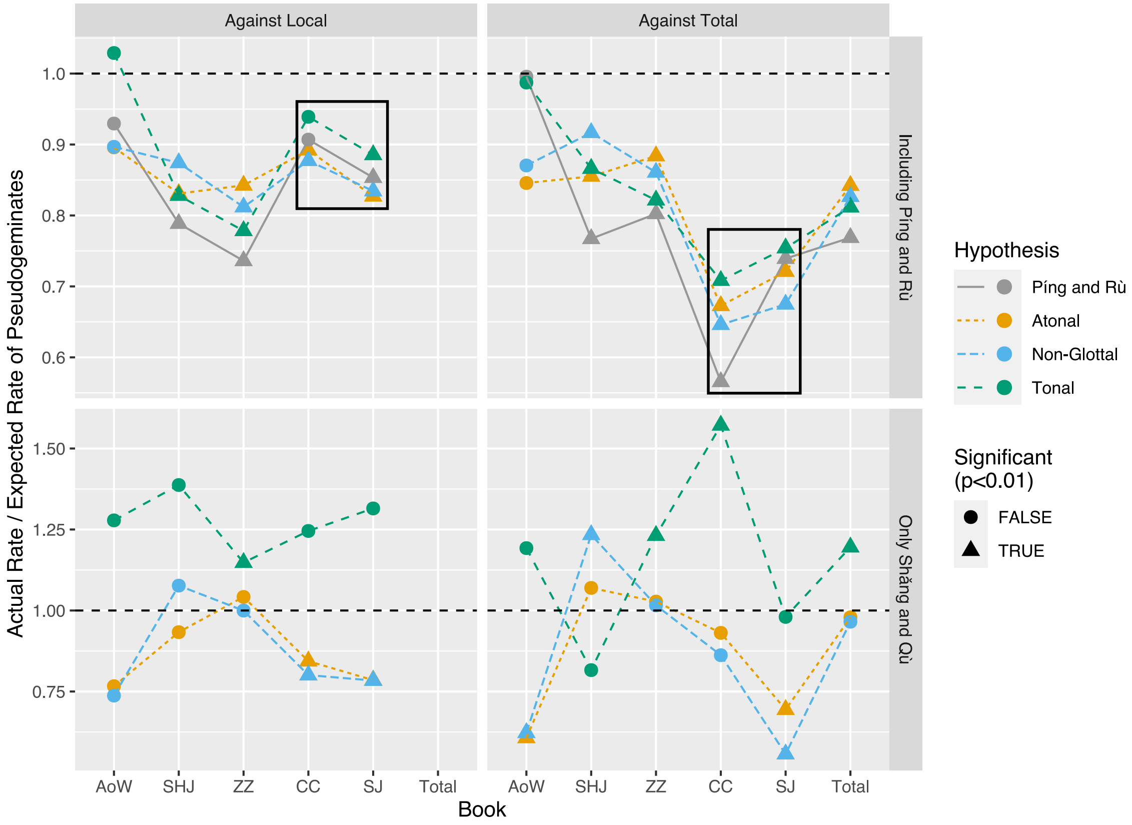

Figure 1: The ratio of the actual and expected rates of *PseudoGem violations within the corpus; rectangles highlight poetic works when tests include *píng* 平 'level' and *rù* 入 'entering' tones.

5.3 *Summary* Figure 1 shows the ratio of actual to expected rates of *PseudoGem violations across the various texts and tests. Significant ($p < 0.01$) results are shown as triangles, while non-significant results are shown as circles. When considering all tones together, all three hypotheses test significant in favor of the presence of OCP avoidance strategies, with all test results being equal except for CC, which tests as having significantly ($p < 0.002$) fewer *PseudoGem violations than the local expected rate in the Atonal and Non-Glottal Hypotheses, but not the Tonal Hypothesis.

When considering *shǎng* 上 'rising' and *qù* 去 'departing' tone characters, the Tonal Hypothesis predicts that in many cases, including the general case testing the entire corpus, OCP violations were sought out. This runs counter to both the previous literature on the effects of OCP violations on language, as well as to our findings among *píng* 平 'level' and *rù* 入 'entering' tone characters, for which OCP violations are avoided. This is highly indicative that the Tonal Hypothesis is most likely not correct for the majority of these texts, and that all of these texts still maintained post-codas. However, each text was written at different points in time, and thus there may be variation across which hypothesis is correct for which text. This is discussed in Section 5.5. Between the Non-Glottal and Atonal Hypotheses, the only large difference in variation is with respect to the SHJ, where the Non-Glottal Hypothesis results in a significant result in favor of an inclination towards OCP violations. Overall both of these hypotheses suggest that OCP violations were avoided.

5.4 *Poetry vs Prose* Of the five texts, SJ and CC are poetry, while the others are prose. As highlighted in Figure 1, the ratio of actual rate of *PseudoGem violations to expected rate of *PseudoGem violations for for poetry is comparable to prose when tested against the local expected rate; however, when tested against the expected rate of the entire corpus, the ratio of actual and expected rates of *PseudoGem is much lower for poetic works in comparison to prose works.

A fixed effects model which predicts the ratio of actual to expected rates of *PseudoGem violations based on the genre (Poetry vs Prose) and the corpus tested against (Local vs Total) finds that the interaction of the Poetry genre against the Total corpus is a significant ($p < 0.00005$) predictor of a lower ratio of rates. This suggests that poets were particularly sensitive to phonological constraints. The neutralization of this effect when tested against the local expected rate may come out of the fact that using marked syntax as a strategy for OCP avoidance, while resulting in fewer *PseudoGem violations, does not change the underlying diction of a work, and thus the predicted rate, based on randomization, should be just as high as it would be if the syntax was unmarked. If poets primarily rely on a more flexible word order as a distinguishing factor from prose, then the local predicted rates of poetic works may be equivalent to that of prose, but when compared to prose works (as in with the total predicted rates), the avoidance of *PseudoGem violations becomes more apparent.

This correspondence between poetry and lower ratios of rate holds across all tests which include *píng* 平 'level' and *rù* 入 'entering' tone characters, but is not true of the Tonal Hypothesis when only accounting for *shǎng* 上 'rising' and *qù* 去 'departing' tone characters. Additionally, even when considering all tones, the results of the Tonal Hypothesis is consistently closer to chance than the Atonal and Non-Glottal Hypotheses. Both of these facts further suggest that post-codas were present in the works of the SJ and CC, and that writers' diction and syntax were much more sensitive to OCP violations in poetry than in prose.

5.5 *The Time Scale of Tonogenesis* As discussed in Section 1.1, SJ and AoW are older than the other works in the corpus of this study, with SJ dating before 771 BCE, and AoW dating before 471 BCE with all other texts appearing after. When testing for only *shǎng* 上 'rising' and *qù* 去 'departing' tone characters, these two texts stand out as consistently having the lowest ratios of actual to predicted rates of *PseudoGem violations when testing across the Non-Glottal and Atonal Hypotheses, particularly against the expected rate of the entire corpus, where they are the only texts which test as having significantly ($p < 0.002$) fewer *PseudoGem violations than expected. The lack of such results in the other tests suggests that if post-coda segments were lost in the time-frame of this study, that the loss of post-coda segments happened near the beginning of the Spring and Autumn period (776 – 471 BCE), between the writing of AoW and the three more recent texts, though more tests across a wider range of texts are required to more confidently pinpoint the time of post-coda loss.

6 Conclusion

This paper finds that violations of the Obligatory Contour Principle (OCP) are avoided in Classical Chinese, and suggests that the OCP may have influenced writers' decisions in using marked word order or choice as a way to avoid forming pseudogeminates across characters. Furthermore, the results show that poetry is significantly more sensitive to this phenomenon, and suggests that poets opted for marked syntax as an avoidance strategy to avoid marked phonology more so than writers of prose text. These findings are consistent throughout multiple hypotheses of the stages of tonogenesis in the language of the texts, though when considering only *shǎng* 上 'rising' and *qù* 去 'departing' tone characters, it is highly suggested that post-codas were present in AoW and SJ, if not all of the texts. Whether or not the presence or absence of OCP violations given certain reconstructions may lead us towards capturing the timeline of sound change in a language is yet to be seen, though the results in this paper are promising and require further investigation.

References

Baxter, William Hubbard (1992). *A Handbook of Old Chinese Phonology*. No. 64 in Trends in linguistics, Mouton de Gruyter, Berlin ; New York.

Baxter, William Hubbard & Laurent Sagart (2014). *Old Chinese: a New Reconstruction*. Oxford University Press, Oxford ; New York.

Breiss, Canaan & Bruce Hayes (2020). Phonological markedness effects in sentence formation. *Language* 96:2, 338–370.

Cho, Sunghye (2017). *Development Of Pitch Contrast And Seoul Korean Intonation*. PhD dissertation, University of Pennsylvania, Philadelphia.

Durrant, Stephen W., Wai-yee Li & David Schaberg (eds.) (2016). *Zuo Tradition/Zuozhuan: Commentary on the "Spring and Autumn Annals."*. Classics of Chinese Thought, University of Washington Press, Seattle.

Gunkel, Dieter & Kevin M Ryan (2011). Hiatus avoidance and metrification in the Rigveda. *Proceedings of the 22nd Annual UCLA Indo-European Conference* 53–68.

Gunkel, Dieter & Kevin M Ryan (2015). Investigating Rigvedic Word Order in Metrically Neutral Contexts.

Haudricourt, André G. (1954). Comment Reconstruire Le Chinois Archaïque. WORD 10:2-3, 351–364.

Kingston, John (2011). Tonogenesis. *The Blackwell companion to phonology* 1–30.

Leben, William Ronald (1973). *Suprasegmental Phonology*. PhD Thesis, Massachusetts Institute of Technology.

Leben, William Ronald (1978). The representation of tone. *Tone*, Elsevier, 177–219.

Lust, John (1996). *Chinese Popular Prints*, vol. 4 of *Handbuch Der Orientalistik. Vierte Abteilung, China, Handbook of Oriental Studies, China*. Brill, Leiden ; New York.

McCarthy, John J (1986). OCP effects: Gemination and antigemination. *Linguistic Inquiry* 17:2, 207–263.

Myers, Scott (1997). OCP Effects in Optimality Theory. *Natural Language & Linguistic Theory* 15:4, 847–892.

Phu, Van Han, Jerold Edmondson & Kenneth Gregerson (1997). Eastern Cham as a tone language. *Mon-Khmer Studies* 20, 31–43.

Prince, Alan S. & Paul Smolensky (1993). Optimality Theory: Constraint Interaction in Generative Grammar. *Optimality Theory in phonology* p. 3.

Ross, Malcolm D. (1993). Tonogenesis in the North Huon Gulf Chain. *Oceanic Linguistics Special Publications* 24, 133–153.

Shen, Zhongwei (2020). *A Phonological History of Chinese*. Cambridge University Press, Cambridge ; New York, NY.

Shih, Stephanie S. (2017). Phonological Influences in Syntactic Alternations. Gribanova, Vera & Stephanie S. Shih (eds.), *The Morphosyntax-Phonology Connection*, Oxford University Press, 223–252.

Shih, Stephanie S. & Kie Zuraw (2017). Phonological conditions on variable adjective and noun word order in Tagalog. *Language* 93:4, e317–e352.

Shih, Stephanie, Jason Grafmiller, Richard Futrell & Joan Bresnan (2015). Rhythm's role in genitive construction choice in spoken English. Vogel, Ralf & Ruben van de Vijver (eds.), *Rhythm in Cognition and Grammar*, De Gruyter Mouton, 207–234.

Sturgeon, Donald (2019). Chinese Text Project: A dynamic digital library of premodern Chinese. *Digital Scholarship in the Humanities* .

Tz'ù, Ch'u & David Hawkes (1959). *The Songs of the South: An Ancient Chinese Anthology*. Columbia University Press.

Zhèngzhāng, Shàngfāng (2003). *Old Chinese Phonology*. Chinese Modern Linguistics Series, Shanghai Jiaoyu Chubanshe, Shanghai.