Glimpses of Ancient China from Classical Chinese Poems

John LEE WONG Tak-sum
Halliday Centre for Intelligent Applications of Language Studies
Department of Chinese, Translation and Linguistics
City University of Hong Kong
{jsylee,tswong}@cityu.edu.hk

ABSTRACT

While our knowledge about ancient civilizations comes mostly from studies in archaeology and history books, much can also be learned or confirmed from literary texts. Using natural language processing techniques, we present aspects of ancient China as revealed by statistical textual analysis on the Complete Tang Poems, a 2.6-million-character corpus of all surviving poems from the Tang Dynasty (AD 618—907). Using an automatically created treebank of this corpus, we outline the semantic profiles of various poets, and discuss the role of seasons, geography, history, architecture, and colours, as observed through word selection and dependencies.

KEYWORDS : Classical Chinese, poetry, dependency parsing, word selection, semantics.

Proceedings of COLING 2012: Posters, pages 621–632,
COLING 2012, Mumbai, December 2012.

1 Introduction

In Classical Chinese literature, the prestige and popularity of poetry can hardly be overstated. Scholars aspired to master poem composition, not only for career advancement but also as the vehicle for personal expression and social commentary. Common people also liked to memorize, chanted, or even composed poems. The Tang Dynasty (AD 618—907) is widely viewed as the zenith of the art of poetry.

All surviving Tang poems have been preserved in an anthology called the Complete Tang Poems[1]. The whole corpus consists of around 2.6 million Chinese characters, drawn from more than 40,000 poems, composed by 2510 authors, as well as some anonymous ones. The ten most prolific poets, by number of characters, are shown in Table 1.

The sheer size of this corpus means that it would be difficult for any single scholar to analyse all poems by reading. Using a recently compiled digital treebank, we present the first study that exploits the entire corpus to answer questions about semantic content and word usage in the Complete Tang Poems. After outlining previous research (Section 2), we describe our data (Section 3), and then present our textual analysis (Section 4).

Poet	# characters	Poet	# characterss
Bái Jūyì 白居易	187964	Hán Yù 韓愈	41471
Dù Fǔ 杜甫	105930	Guàn Xiū 貫休	40306
Lǐ Bái 李白	84465	Qí Jī 齊己	38635
Yuán Zhěn 元稹	66426	Lù Guīméng 陸龜蒙	36590
Liú Yǔxí 劉禹錫	47880	Mèng Jiāo 孟郊	32446

TABLE 1 – The ten most prolific poets in the Complete Tang Poems.

2 Previous Research

2.1 Text Corpora of Classical Chinese

There has been increasing interest in corpus-based research on historical languages (Crane & Lüdeling, 2012). Large-scale corpora for Classical Chinese include the Academia Sinica Ancient Chinese Corpus (Wei et al., 1997), the corpus at the Centre for Chinese Linguistics Corpus at Peking University, the Chinese Ancient Text Database at the Chinese University of Hong Kong (Ho, 2002), and the Sheffield Corpus of Chinese (Hu et al., 2005). Linguistic annotations, if available in these corpora, are limited to part-of-speech (POS) tags. With this constraint, most previous corpus-based studies focused on character frequency distribution (Zhū, 2004; Zhāng, 2004; Qín, 2005), including a concordance for the Complete Tang Poems (Shǐ, 1990).

In terms of syntactic annotations, only two treebanks are currently available: a constituent treebank on 1000 sentences from the pre-Qin period (Huang et al., 2002), and a dependency treebank on a small subset of the Complete Tang Poems (Lee & Kong, 2012). This latter treebank will be used as training data to automatically produce dependency trees for the entire Complete Tang Poems, on which our word analysis will be based.

[1] In Chinese, 全唐詩 *Quántángshī*, (or *Ch'üan T'ang Shi*). The anthology was compiled by a team of scholars in 1705. Our digital version is downloaded from http://www.xysa.com/quantangshi/t-index.htm

2.2 Studies on the Complete Tang Poems

Research on syntactic and semantic issues in the Complete Tang Poems is a venerable subfield in Classical Chinese philology, with a vast literature. We seek to demonstrate a new route of investigation that can be complementary to traditional scholarship: by interrogating the treebank, one can quickly and objectively see broad trends on the entire corpus, which can help validate previous studies based on smaller sample, or point to interesting patterns for further in-depth analysis by hand.

A case in point is the semantic classification scheme of Wáng Lì, listed in Table 2. Wáng proposed 22 semantic categories (Wáng, 1989, p. 184−203), mostly for nouns but also some function words. As part of our analysis, we will apply these categories on the Complete Tang Poems to create semantic profiles of various poets (Section 4.1).

Category	**Representative words**	**Category**	**Representative words**
Celestial	天 sky 日 sun 風 wind	Body Parts	心 heart 目 eye 翼 wing
Seasonal	年 year 秋 fall 晝 day	Human emotions	談 talk 笑 smile 愛 love
Geographic	山 hill 池 pool 道 path	Human relationships	父 dad 王 king 僧 monk
Architectural	房 room 門 door 店 shop	Pronouns	吾 I 汝 you 誰 who
Products of civilization	車 car 弓 bow 杯 cup	Locations	東 east 後 back 上 up
Clothing	衣 cloth 帽 hat 甲 armour	Numbers	一 one 幾 some 半 half
Food	酒 wine 飯 rice 菜 veg	Colours	紅 red 金 gold 素 plain
Instruments	筆 pen 書 book 琴 piano	Calendar coordinates	甲 1st 乙 2nd 丙 3rd 丁 4th
Literary	詩 poem 歌 song	Adverbs	怎 how 不 not 只 only
Flora	木 tree 李 plum 根 root	Conj. & prep.	與 and 於 at 之 of
Fauna	馬 horse 鳥 bird 魚 fish	Particles	也 *yě* 乎 *hū* 然 rán

TABLE 2 – Semantic categories for nouns in the Complete Tang Poems (Wáng, 1989).

3 Data

A dependency treebank covering a subset of the Complete Tang Poems has been built (Lee & Kong, 2012). It consists of about 32,000 words, annotated with part-of-speech (POS) tags and dependency labels, derived from the Penn Chinese Treebank (Xue et al., 2005) and Stanford dependencies for Modern Chinese (Chang et al., 2009).

Using this treebank as training data, we performed POS tagging[2] on the whole Complete Tang Poems with TreeTagger (Schmid, 1995), and dependency parsing with the Minimum-Spanning Tree (MST) Parser (McDonald et al., 2006). On ten-fold cross-validation on the treebank itself, the average UAS and LAS of dependency parsing are 84.3% and 75.6% respectively[3].

[2] Although word segmentation is provided in the treebank, "in general the syllable, written with a single character, and the word correspond in Classical Chinese" (Pulleyblank, 1995, p. 8); most words to be analysed in this paper (Section 4) are indeed single characters.

[3] Similar experiments with MaltParser (Nivre et al., 2009) yielded similar accuracy rates.

4 Analysis

We first analyze the semantic content of the Complete Tang Poems both globally and by author (section 4.1), then use dependency information to glean aspects of the seasons, geography, architecture, history and use of colours in Ancient China (section 4.2).

4.1 Semantic Profile

To identify the main themes of the poems, we compute the distribution of the semantic categories listed in Table 2; the result is shown in Table 3. The five most frequent categories are 'Geographic', 'Adverbs', 'Celestial', 'Human emotions', and 'Seasonal'. For the most prolific poets, at least four of these five categories also rank among their individual top five, indicating that the topics of interest are rather uniform among Tang poets. Overall, aspects of nature ('Geographic', 'Celestial', 'Seasonal', etc.) dominate the attention of the poets, compared to aspects of humans ('Human emotions', 'Human relationships', 'Body parts', etc.).

Category	**Freq.**	**Category**	**Freq.**	**Category**	**Freq.**
Geographic	11.79%	Flora	4.94%	Conj. and prep.	2.54%
Adverbs	9.49%	Pronouns	4.87%	Clothing	1.19%
Celestial	9.48%	Body parts	4.84%	Instruments	1.08%
Human emotions	6.59%	Colours	4.46%	Food	0.78%
Seasonal	6.58%	Architectural	3.76%	Calendar	0.70%
Numbers	6.19%	Products	3.40%	Particles	0.59%
Locations	5.52%	Fauna	3.38%		
Human relationships	5.15%	Literary	2.66%		

TABLE 3 – Distribution of Wáng Lì's semantic categories in the Complete Tang Poems, based on the 864 example characters provided by Wáng. They cover 49% of the tokens in the corpus.

The absolute counts, however, mask some interesting underlying tendencies. To see the extent to which individuals deviate from the average distribution in Table 3, we calculate the z-score for each poet's own distribution. Further, we compute the TF-IDF of words, considering the complete works of each poet as a "document".

As shown in Table 4(a), Bái Jūyì wrote more than the average poet on human themes (e.g., 'Body parts', 'Food'), and less on 'Celestial' and 'Geographic', two of the most common categories related to nature (Table 3). This tendency is confirmed by his words with the highest TF-IDF, listed in Table 5(a), such as 病 bìng 'sick', 衰 shu*āi* 'decline' and 憂 *yōu* 'worry', describing the harshness of life. Another set of high-TF-IDF words involve drinking, such as 杯 *bēi* 'glass', 飲 *yǐn* 'drink', and 酒 ji*ǔ* 'wine'. These statistics concur with the general observation that Bái uses the theme of drinking to illustrate his loneliness and miserableness (Zuǒ, 2011).

Being disrupted by the Ān Lùshān Rebellion, Dù Fǔ was known for his anti-war stance, concern about his country's decline, and sympathy for the common people (Lú, 2009). These themes are confirmed by his set of words about warfare and turmoil, listed in Table 5(b), and his relative disinterest, like Bái Jūyì, in the common themes in nature – in his case, 'Seasonal' and 'Celestial'.

(a) Bái Jūyì 白居易		**(b) Dù Fǔ 杜甫**	
Body Parts	1.42	Human relationships	0.92
Food	1.32	Fauna	0.59
Conj. and prep.	1.23	Calendar	0.57
Pronouns	1.07	Food	0.54
Numbers	1.07	Literary	0.53
Adverbs	0.76	Pronouns	0.48
Architectural	-0.64	Seasonal	-0.52
Fauna	-1.17	Celestial	-0.57
Celestial	-1.32	Flora	-0.93
Geographic	-1.35	Human emotions	-1.08
(c) Lǐ Bái 李白		**(d) Wáng Wéi 王維**	
Colours	1.28	Particles	1.75
Human relationships	0.93	Clothing	1.30
Food	0.79	Human relationships	1.30
Conj. and prep.	0.68	Locations	1.19
Pronouns	0.65	Architectural	0.60
Celestial	0.57	Food	0.35
Flora	-0.65	Numbers	-0.65
Calendar	-0.76	Celestial	-0.83
Architectural	-0.85	Seasonal	-0.93
Seasonal	-1.78	Human emotions	-1.21

TABLE 4 – Semantic categories with the highest and lowest z-scores of four well-known poets. The higher the score, the more the poet exceeds the average in the use of the category.

Poet	Characters	Topic
(a) Bái Jūyì	病 ‘sick’ 衰 ‘decline’ 憂 ‘worry’ 苦 貧 臥	Harshness of life
	杯 ‘glass’ 飲 ‘drink’ 酒 ‘wine’ 歡 醉	Drinking
	弦	Warfare
(b) Dù Fǔ	衰 ‘decline’ 老 ‘old’ 病 ‘sick’	Harshness of life
	胡 兵 亂 泥 失 骨 軍 夫 戰	Warfare and turmoil
(c) Lǐ Bái	胡 劍 陵 嘆 悲 夫	Warfare
	笑 美 顏 女	Women
	杯 ‘glass’ 飲 ‘drink’	Drinking
(d) Wáng Wéi	戶 隱 雞 鳴 田 井 村 川 悠 門	Isolation
	戰	Warfare

TABLE 5 – Characters with the highest TF-IDF in the works of four poets, grouped into main topics.

In contrast, under the pen of the poet Lǐ Bái, 'Celestial', already a popular category (Table 3), is employed even more frequently. This likely reflects his extensive use of the moon as imagery. His poems are also well recognized for vivid colours and the drinking theme ('Colours' and 'Food' in Table 4(c)), with the characters 杯 *bēi* 'glass' and 飲 *yǐn* 'drink' achieving some of the highest TF-IDF scores.

Lastly, as shown in Table 4(d), the top category for Wáng Wéi is 'Particles', no doubt a result of his frequent use of 兮 *xī* , a particle mainly used in archaic poems. This is a style of which Wáng is known to be fond.

4.2 Word selection

We now exploit dependency information to investigate word selections, centering on three common areas: the seasons, the cardinal directions, and the colours.

4.2.1 Seasons

Among the four seasons, mentions of 春 chūn 'spring' and 秋 *qiū* 'autumn' overwhelmingly outnumber those of 夏 xià 'summer' and 冬 *dōng* 'winter', by a factor of more than ten to one. As seen from the written record in Shang dynasty (circa BC 17c. − 1046), only "spring" and "autumn" were attested in oracle bone inscriptions but "summer" and "winter" were not (Chén, 1988, p. 226 − 227). Thus, the discrepancy may be explained by the fact that the concepts of 'spring' and 'autumn' are much older, and also that these two seasons were bound up with many activities in ancient China. Given this discrepancy, it is more appropriate to use mutual information (MI) than absolute counts to detect significant word selections.

Notable word co-occurrences with the highest MI are shown in Table 6. Reflecting the natural order, both 'summer' and 'autumn' are predominately associated with plant words; 'spring' is associated with significantly fewer ones, and 'winter', hardly any. By the same reasoning, one might expect the word 暉 huī 'sunshine' to relate most strongly with 'summer'. Its relation with 'spring' is in fact stronger since, when poets pay tribute to spring as mother nature, as it were, they often depict the spring sun which is gentle, comforting, and caring for the sprouting of the plants after a severe winter. This tribute also explains the high MI of the direction 'east' for the word 風 fēng 'wind' (section 4.2.2), as wind usually blows from the east during spring. In contrast, summer is more frequently described with words such as 酷 kù 'extreme' and 暑 *shǔ* 'heat', rather than 'sunshine'.

Since peasants formed the majority of the population (Murphey, 1996, p. 5), agriculture was a common way of life. Agricultural activities were highly regulated by the seasons, and naturally the word 耨 nòu 'raking' is significantly related with 'spring', and 稼 jià 'harvest' with 'autumn'. Another major means of subsistence was hunting, especially in the winter, when cooked meat was especially coveted. It is no coincidence that 狩 shòu 'hunting' has the highest MI with 'winter'.

There are two words that both mean 'sleep', namely 睡 shuì and 蟄 zhé. A glance at Table 6, however, shows that the former is highly correlated with 'spring', whereas the latter with 'winter'. The reason is that shuì in general refers to humans, while zhé refers to animals, which tend to go into hibernation during winter.

Scholars were not immune from seasonal cycles, either. National examinations were held annually at the capital city, and passing these exams was critical in climbing the career ladder. Since the examinations were held in spring, the words 闈 wéi 'examination' and 榜 bǎng 'result' only collocate with that season. Candidates who failed the examination sometimes stayed in the capital to take remedial lessons, therefore 課 kè 'lesson' is often modified by 'summer'.

Spring			Summer			Autumn			Winter		
Ch	MI	Meaning	Ch	MI	Meaning	Ch	MI	Meaning	Ch	MI	Meaning
晼	3.41	beauty	汭	4.51	bend of river	旻	3.81	autumn sky	狩	5.56	hunting
闈	2.67	examination	蘖	4.21	sprout	韭	3.56	chives	菁	4.70	flower
韭	2.55	chives	課	4.07	lesson	穟	2.77	ripe grain	蟄	3.19	sleep
耨	2.41	raking	酷	3.39	extreme (hot)	稼	2.70	harvest	筍	3.03	bamboo shoots
暉	2.39	sunshine	筱	3.12	bamboo	荼	2.40	vegetable	霰	2.83	ice
酎	2.33	vintage wine	卉	3.04	grass	蔬	2.34	vegetable	蕊	1.98	bud
醪	2.29	mellow wine	筎	2.95	bamboo	草	1.91	grass			
風	2.22	wind	苗	2.87	hunting; seed	芋	1.74	taro			
釀	2.10	brew	菜	2.77	vegetable	菰	1.70	taro			
草	2.06	grass	葛	2.68	arrowroot						
醁	1.97	wine	木	2.64	tree						
苜	1.91	clover	暑	2.63	heat						
霖	1.87	heavy rain	麥	2.56	wheat						
畦	1.86	field	果	2.50	fruit						
睡	1.85	sleep	萼	2.29	calyx						
榜	1.72	exam result	蕊	2.18	bud						
蔬	1.67	vegetable	筍	1.85	bamboo shoots						
筍	1.44	bamboo shoots	蘚	1.60	moss						

TABLE 6 – Characters with the highest mutual information (MI) with each of the four seasons. Two characters are considered to co-occur when they are connected by a dependency relation. Characters occurring less than 10 times are excluded.

4.2.2 Cardinal directions

Like the seasons, the four cardinal directions – 東 *dōng* 'east', 南 nán 'south', 西 *xī* 'west', and 北 *běi* 'north' – appear frequently in poems, contributing the bulk of the counts towards the category 'Locations'. Table 7 lists several sets of words with similar meaning but drastically different co-occurrences with the directions. They reveal facets of culture, history and geography of Ancient China.

Geography. The verbs 流 liú 'flow' and 逝 shì 'pass' both like to head eastward. In China, most rivers flow from mountains in the west towards the Pacific Ocean in the east. Since liú and shì tend to be associated with rivers, 'east' is the natural direction for them. Now, given that the ocean is located in the east, one might wonder why 海 hǎi 'sea' has such high MI with 'north'. In fact, in most contexts, the term refers to the remote area in the north far away from the central

plain. Likewise, 南國 nánguó ‘south country’ refers to the remote area in the south, and so 南省 nánshěng ‘south province’.

History. The words 都 *dū* and 京 *jīng* both mean ‘capital’, yet they have diametrically opposing directions, namely ‘east’ and ‘west’. In many dynasties, China had a main capital in the west and also a secondary capital in the east; for example, in the Tang dynasty, they were Cháng’ān and Luòyáng, respectively. The word *jīng* usually refers to the main capital, while *dū* refers to the secondary. Since the Tang capital was located in the west, when an emperor went out on a 巡 xún ‘patrol’ to tour his domain, he was likely to go ‘east’ or ‘south’. Also, seen from the capital, barbarians on the fringes of the empire were labelled with the name of the tribe that dwelled in that direction during the archaic period. These were 狄 dí in the north, 蠻 mán in the south, 戎 rúng in the west, and 夷 yí in the east or south.

Architecture. The distributions of the cardinal directions also tell us about architectural design. While ‘east’ and ‘west’ are the dominant directions of 廂 *xiāng* ‘side-room’, ‘north’ has the highest MI with 堂 táng ‘hall’. The reason lies with the design of quadrangle courtyards, a common type of residence in ancient China. In a typical courtyard, the main house, or hall, faced the north, while the side-rooms were located along the east-west axes. Furthermore, a small building is often built in the west for moon-viewing. Hence, the word 樓 lóu ‘building’ is most likely to be modified by ‘west’.

Topic	Co-occurring word	East	South	West	North
Geography	流 liú ‘flow’	**2.46**	-0.43	0.55	0.89
	逝 shì ‘pass’	**2.41**	0.28	/	/
	海 hǎi ‘sea’	**1.95**	0.98	0.85	**1.55**
	國 guó ‘nation’	-0.83	**2.76**	0.06	-1.93
	省 shěng ‘province’	0.85	**2.02**	1.42	1.35
	風 fēng ‘wind’	**2.01**	0.75	1.53	1.72
History	都 *dū* ‘capital’	**2.37**	0.51	0.78	0.09
	京 *jīng* ‘capital’	1.84	0.52	**2.78**	1.33
	巡 xún ‘patrol’	**2.24**	**2.41**	1.86	0.19
	夷 yí ‘tribe’	**0.82**	**0.95**	0.70	-0.28
	蠻 mán ‘tribe’	0.37	**1.24**	/	0.18
	戎 rúng ‘tribe’	/	-1.06	**2.22**	-0.50
	狄 dí ‘tribe’	/	/	/	**3.72**
Architecture	廂 *xiāng* ‘side-room’	**4.17**	3.21	**4.45**	/
	堂 táng ‘hall’	1.93	-1.33	-0.49	**2.52**
	樓 lóu ‘building’	0.72	1.37	**2.00**	1.09

TABLE 7 – Word co-occurrences with the four cardinal directions that have high mutual information. Two characters are considered to co-occur when they are connected by a dependency relation. Characters occurring less than 10 times are excluded.

4.2.3 Colours

Two common words in Classical Chinese both refer to the black colour, namely, 黑 h*ēi* and 玄 xuán. The former tends to be used in negative contexts, and the latter one in positive ones, sometimes indicating an auspicious sign (Yìng, 2004, p.13).

To verify this hypothesis, we compute the mutual information (MI) of characters co-occurring with h*ēi* or xuán. Table 8 lists those characters with the highest MI. Most co-occurrences with xuán involve an exalted or noble entity, such as 玄圃 xuán*pǔ* 'palace of the gods', 玄貺 xuánkuàng 'present from emperor', 玄豹 xuánbào 'leopard' (a rare and thus valuable animal), 玄宗 xuánzōng 'idea on Buddhism', and 玄晏 xuányàn 'ritual'. In contrast, those involving h*ēi* are mostly everyday objects (e.g., 'rice') including some with negative sentiment such as 黑紗 h*ēi*sh*ā* 'funeral cloth' and 黑蛺 h*ēi*jiá 'bug'. These observations lend evidence to the usage of these two characters described in (Yìng, 2004).

玄 xuán 'black'				**黑 h*ēi* 'blank'**			
freq.	Ch	MI	Meaning	freq.	Ch	MI	Meaning
13	牝	5.92	root of everything	9	煤	5.74	ash
207	圃	4.86	gods' palace	170	貂	5.28	sable
48	貺	4.33	present from emperor	42	蛺	4.20	bug
149	豹	4.30	leopard (valuable)	149	米	3.62	rice
434	暉	4.29	sun/moon	129	壤	3.08	fertile earth
541	宗	4.24	idea on Buddhism	277	蛟	3.00	dragon
293	晏	3.91	ritual	337	紗	2.81	cloth for funeral
49	奧	3.90	difficult	176	蟻	2.76	ant
363	兔	3.51	moon	176	鉛	2.76	graphite
234	覽	3.44	foresight	356	裘	2.75	fur coat
39	祉	3.44	kindness from ruler	4090	頭	2.71	young-age
179	冕	3.30	clothes of ruler	3845	龍	2.32	dragon

TABLE 8 – Word co-occurrences with the two words for 'black', h*ēi* and xuán.

Conclusion and Perspectives

This paper presents textual analysis on the entire Complete Tang Poems. We described the overall semantic range of the corpus, as well as the semantic profiles of various poets, via a semantic classification scheme and TF-IDF scores. We then used dependency relations and mutual information to investigate word selections involving the four seasons, the four cardinal directions and the black colour. Our observations lend statistical evidence to previous scholarly assertions, but also reveal aspects of Chinese geography, history, and architecture.

Our analyses represent a new avenue of scholarly enquiry over this treasure trove of Classical Chinese, but they have touched only the tip of an iceberg. It is hoped that the automatically produced treebank will provide useful syntactic features for other research topics, such as the readability of poems (Zhāng et al., 2009) and authorship questions (Matsuoka, 2003).

Acknowledgments

This project was supported in part by a Strategic Research Grant (#7002549) from City University of Hong Kong.

References

Chang, P.-C., Tseng, H., Jurafsky, D. and Manning C. D. (2009). Discriminative Reordering with Chinese Grammatical Relations Features. In Proc. 3rd Workshop on Syntax and Structure in Statistical Translation.

Chén M. (1988). *Yīnxū Běncí Zòngshù*. Zhonghua Book Company, Peking.

Crane, G., and Lüdeling, A. (2012). Introduction to the Special Issue on Corpus and Computational Linguistics, Philology, and the Linguistic Heritage of Humanity. Journal on Computing and Cultural Heritage, 5(1).

Ho, C. W. (2002). CHANT (CHinese ANcient Texts): A Comprehensive Database of All Ancient Chinese Texts up to 600 AD. Journal of Digital Information, 3(2).

Hu, X., Williamson, N., and McLaughlin, J. (2005). Sheffield Corpus of Chinese for Diachronic Linguistic Study. Literary and Linguistic Computing, 20(3):281−293.

Huang, L., Peng, Y., Wang, H., and Wu. Z. (2002). PCFG Parsing for Restricted Classical Chinese Texts. In Proc. 1st SIGHAN Workshop on Chinese Language Processing.

Lee, J. and Kong, Y. H. (2012). A dependency treebank of Classical Chinese poems. In Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).

Lú, Y. (2009). A Brief Study on Dù Fǔ's Anti-War Poems. Journal of Yunmeng, 4:109−114.

Matsuoka, E. (2003). Are Táo Yuānmíng and Xiè língyùn from the Same Era?: Study in social context after the electronic publication of the Two-Fours. In 1st Literature and Information Technology International Conference, National Tsing Hua University of Taiwan and Yuan Ze University of Taipei, Republic of China.

McDonald, R., Lerman, K. and Pereira F. (2006). Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. In 10th Conference on Computational Natural Language Learning (CoNLL-X).

Murphey, R. (1996). East Asia: A New History. Addison-Wesley Educational Publishers Inc., New York.

Nivre, J., Kuhlmann, M. and Hall, J. (2009). An Improved Oracle for Dependency Parsing with Online Reordering. In Proc. 11th International Conference on Parsing Technologies (IWPT), pages 73−76.

Pulleyblank, E. (1995). Outline of Classical Chinese Grammar. UBC Press, Vancouver.

Qín, Q. (2005). A Statistical Study on Character Frequency of Pre-Qin Books. Studies in Language and Linguistics, 25(4):112−116.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In Proc. ACL SIGDAT-Workshop, Dublin, Ireland.

Shǐ, C. (1990). Indexes of Complete Tang Poems. Shanghai Guji, Shanghai.

Wáng, L. (1989). Versification in Chinese. Wáng Lì Collection, vol. 14, 15. Shandong Jiaoyu Chubanshe, Ji'nan.

Wei, P., Thompson, P. M., Liu, C., Huang, C., and Sun, C. (1997). Historical Corpora for Synchronic and Diachronic Linguistics Studies. Computational Linguistics and Chinese Language Processing, 2(1):131−145.

Xue, N., Xia, F., Chiou, F.-D., and Palmer, M. (2005). The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering, 11:207−238.

Yìng, L. (2004). "The whole poems in Tang dynasty" colour words and expressions research. M.A. thesis. Chongqing: Southwest University.

Zhū, Y. (2004). A statistical account of the frequency and distribution of the character usage in the ancient Chinese classics. Journal of the National Library of China, 49:91−93.

Zhāng, J., Ān P. and Sū N. (2009). An analysis of the entropy of the character frequency and the grading of readability by the general public of the Tang poems. Science and Technology Information, 2009(6):241−243.

Zhāng, Z. (2004). A large scale statistical report on the usage of characters in ancient Chinese classics. In 3rd Conference of Database of Chinese Literature and History.

Zuǒ, C. (2011). A Look at the experience and cognition of Bái Jūyì towards wine from Xiào Táo Qián Tí Shī Shíliù Shǒu. Young Litterateur, 24.