Syntactic patterns in classical Chinese poems: A quantitative study

John Lee and Yin Hei Kong
Department of Linguistics and Translation, City University of Hong Kong, Hong Kong SAR, China

Mengqi Luo
Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan, China

Abstract

It is widely believed that different parts of a classical Chinese poem vary in syntactic properties. The middle part is usually parallel, i.e. the two lines in a couplet have similar sentence structure and part of speech; in contrast, the beginning and final parts tend to be non-parallel. Imagistic language, dominated by noun phrases evoking images, is concentrated in the middle; propositional language, with more complex grammatical structures, is more often found at the end. We present the first quantitative analysis on these linguistic phenomena—syntactic parallelism, imagistic language, and propositional language—on a treebank of selected poems from the *Complete Tang Poems*. Written during the Tang Dynasty between the 7th and 9th centuries CE, these poems are often considered the pinnacle of classical Chinese poetry. Our analysis affirms the traditional observation that the final couplet is rarely parallel; the middle couplets are more frequently parallel, especially at the phrase rather than the word level. Further, the final couplet more often takes a non-declarative mood, uses function words, and adopts propositional language. In contrast, the beginning and middle couplets employ more content words and tend toward imagistic language.

Correspondence:
John Lee, Department of Linguistics and Translation, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR, China.
E-mail:
jsylee@cityu.edu.hk

1 Introduction

'Coupling' is a universal principle behind poetic structure (Levin, 1962), and has been recognized in diverse languages such as Hebrew (Kugel, 1981) and Russian (Jakobson, 1966). It is also a pervasive phenomenon in classical Chinese poems, which are composed of 'couplets'—each couplet is a pair of adjacent lines, with an identical number of characters. Most poems contain either two or four couplets. This study focuses on those with four couplets, of which Table 1 shows an example.

Traditionally, the four couplets are characterized as *qi* 'begin, arise', *cheng* 'continue, elaborate', *zhuan* 'turn', and *he* 'conclude, enclose'. The first couplet typically begins by setting the time and place. In Table 1, for example, it mentions two rivers as the backdrop. The second couplet elaborates with pairs of parallel images, for instance 'cloud'/'night' and 'heaven'/'moon'. The third couplet makes a turn,

Digital Scholarship in the Humanities, Vol. 33, No. 1, 2018.
doi:10.1093/llc/fqw059 Advance Access published on 7 January 2017

Table 1 A four-couplet poem entitled 'Yangtze and Han' (江漢) by Du Fu[1]

Couplet	1st char	2nd char	3rd char	4th char	5th char
First couplet	江	漢	思	歸	客
	jiang	*han*	*si*	*gui*	*Ke*
	'Yangtze'	'Han'	'think'	'return'	'stranger'
	By the Jiang and Han rivers broods a homeward traveler,				
	乾	坤	一	腐	儒
	qian	*kun*	*yi*	*fu*	*ru*
	'heaven'	'earth'	'one'	'withered'	'scholar'
	Between heaven and earth is one worthless scholar.				
Second couplet	片	雲	天	共	遠
	pian	*yun*	*tian*	*gong*	*yuan*
	'patch'	'cloud'	'heaven'	'with'	'far'
	A lone cloud, and the sky (and I) join in being faraway,				
	永	夜	月	同	孤
	yong	*ye*	*yue*	*tong*	*gu*
	'eternal'	'night'	'moon'	'with'	'alone'
	A long night, and the moon (and I) share the loneliness.				
Third couplet	落	日	心	猶	壯
	luo	*ri*	*xin*	*you*	*zhuang*
	'set'	'sun'	'heart'	'still'	'strong'
	The setting sun—yet I remain ambitious at heart,				
	秋	風	病	欲	蘇
	qiu	*feng*	*bing*	*yu*	*su*
	'autumn'	'wind'	'sickness'	'about to'	'revive'
	The autumn wind—from illness I will recover.				
Fourth couplet	古	來	存	老	馬
	gu	*lai*	*cun*	*lao*	*ma*
	'antiquity'	'come'	'keep'	'old'	'horse'
	From antiquity all the old horses that people kept,				
	不	必	取	長	途
	bu	*bi*	*qu*	*chang*	*tu*
	'no'	'need'	'take'	'long'	'road'
	Not always were chosen for long distances.				

usually with a different set of images, such as 'sun'/'wind' and 'heart'/'sickness' in our running example. The fourth couplet concludes, often with a subjective and personal statement, 'to move toward a closure and to make a well-rounded whole by joining beginning and end' (Cai, 2008, p. 168). The metaphorical 'old horse' echoes the 'stranger' and 'scholar' in the first couplet to refer to the poet himself.

Although not all poems prescribe exactly to this template, most are said to exhibit syntactic properties that reflect its outline. Specifically, the middle part (i.e. the second and third couplets) is expected to be parallel and to use imagistic language. In contrast, the final part (i.e. the fourth couplet) is 'by convention non-parallel and, as such, particularly conducive to a realistic portrayal of the poet's present condition' (Cai, 2008, p. 168). It is expected to use propositional language, which 'appeals primarily to the understanding' rather than the imagination, and demands an 'intellectual response' rather than a 'sensory' one (Kao and Mei, 1971, pp. 121–2).

This article presents a quantitative study on these linguistic phenomena, with two main contributions. First, we propose computable criteria for detecting syntactic parallelism and imagistic and propositional language. The criteria include not only lexical and part-of-speech (POS) information, which have already been considered in a number of poetry studies (He *et al.*, 2007; Kaplan and Blei, 2007; Kao & Jurafsky, 2012), but also syntactic structures.

Second, as a complement to numerous qualitative studies (e.g. Cai, 2008; Owen, 1985; Yu, 1987), we quantify the extent to which these phenomena are present in different parts of a poem, drawing statistics from a large treebank.

The rest of the article is organized as follows. We sketch the current understanding on parallelism, imagistic language and propositional language in the next section. After presenting in Section 3 the treebank that serves as our data set, we discuss in Sections 4 and 5 the POS and syntactic patterns that are indicative of these phenomena, and analyze their distributions in a poem. We then conclude in Section 6.

2 Background

2.1 Parallelism

A common literary device in classical Chinese literature, parallelism is generally defined as a pair of sentences that have the same number of characters and similar syntax (Chen, 1957). In poetry, this device manifests itself mostly in parallel couplets. In general, the two lines in the couplet must have the same sentence structure; characters occupying the same positions on the two lines must have matching POS, and have related meaning (Feng, 1990).

Our study focuses on the first two requirements.[2] Generally speaking, whether two words 'match' in POS depends on the POS taxonomy, for which there is no accepted standard.[3] If the tagset is overly coarse, it would overestimate the rate of parallelism. If the tagset is too fine-grained, it would fail to discern parallel couplets and underestimate the rate. At the least, content words should correspond to content words, and function words to function words (Tan, 2003). Broadly speaking, content words (*shizi*) are words, such as nouns, verbs, and adjectives, that carry semantic content and form an open class; function words (*xuzi*) are those that express grammatical relationships (Peyraube, 2016). Further, it is suggested that 'a noun corresponds to a noun, a verb corresponds to a verb, and an adjective corresponds to an adjective' (Wang, 1994, p. 142).

It is also difficult to pin down when two lines have the 'same' sentence structure. Feng (1990) expects them to belong to the same broad sentence type, e.g. subject-predicate, verb-object, or conditional. More explicitly, Wang (1994, pp. 182–51) enumerates more than 300 types of sentence structures. Each structure consists of a sequence of letters representing POS, with major phrases separated by hyphens. The type 'fN'-'V/F'-'fN', for example, describes the couplet in Table 2. In both lines, the first two characters form a noun phrase (fN), where an adjective (f) modifies the head noun (N). The third character can be a verb (V) or an adjective (F). The last two characters form a noun phrase (fN) with a similar structure as the first two.

Table 2 A parallel couplet annotated with POS, using the tagset proposed in Wang (1994)

泉	聲	咽	危	石
quan	*sheng*	*yan*	*wei*	*shi*
'fountain'	'sound'	'gurgle'	'precipitous'	'rock'
f	N	V	f	N
'The fountain sound gurgles above precipitous rocks'				
日	色	冷	青	松
ri	*se*	*leng*	*qing*	*song*
'sun'	'color'	'cold'	'blue'	'pine'
f	N	F	f	N
'Sun rays are chilly amidst the blue pines'[4]				

2.2 Imagistic language and propositional language

Classical Chinese poetry employs two conventions of syntax—the 'imagistic language' and the 'propositional language'. These two conventions correspond to two modes of expression, the lyric and the narrative: 'where imagistic language dominates, the lyric mode will emerge; where propositional language dominates, the narrative mode prevails' (Levy, 1988, p. 26).

The imagistic language, as its name suggests, uses images to 'evoke a mental picture or to recall a physical sensation' (Kao and Mei, 1971, p. 120). The closest examples from the Western tradition are perhaps the 'Imagist' poems, such as those by Ezra Pound; they tend to be dominated by noun phrases (NPs), and do not always contain complete sentences. Likewise, imagistic couplets in classical Chinese poems are rich in NPs, often juxtaposed with no linkage between them, resulting in

fragmentary, 'discontinuous' syntax. For example, the first couplet in Table 1 simply consists of two NP, headed by 'stranger' and 'scholar'.

Rather than projecting images, the propositional language makes assertions. Hence, it uses more function words to specify grammatical relations between content words, yielding whole sentences in 'continuous' syntax. For example, the fourth couplet in Table 1 consists of one complete sentence, with its subject (*ma* 'horse') in the first line and its main verb in the second (*qu* 'take'); the main verb is further qualified by two function words, a negation adverb (*bu*, 'not') and a modal verb (*bi*, 'need'). Kao and Mei (1971, p. 60) observed that a poem generally progresses from the imagistic language in the middle couplets, to the propositional language in the final couplet. Yu (1987, p. 194) claims, in contrast, that the traditional form of regulated verse is tripartite, where 'middle scenic couplets are enclosed with propositional statements'. In Section 5, we will develop a number of criteria, based on POS tags and syntactic structure, to detect where imagistic language and propositional language are present in a poem.

2.3 Quantitative analyses and NLP for classical Chinese poetry

There has been increasing interest in applying natural language processing (NLP) techniques to perform quantitative analysis on classical Chinese poetry. Past research has analyzed textual patterns in the *Complete Tang Poems* on various topics, including colors, sentiment, and personal names (Hou and Frank, 2015; Lee and Wong, 2012; Liu *et al.*, 2015). Natural language generation techniques have been applied to automatically compose couplets, using algorithms that take parallelism into account (Jiang and Zhou, 2008; Zhang and Lapata, 2014).

However, there is not yet any large-scale, quantitative study on the degree to which classical Chinese poems adhere to requirements on parallelism. Most previous studies have provided only qualitative discussions (e.g. Owen, 1985; Yu, 1987). Cao (1998) created a database with 1,000 couplets for a study on parallel couplets, but did not report any formal annotations on their syntactic structure. Other quantitative studies relied on relatively small numbers of samples. For example, Yuan (2005, pp. 236–9) analyzed thirty-nine poems by Du Fu, and Huang (2006, pp. 100–7) computed the percentage of parallel couplets in twenty-eight poems by Du Mu.

The use of imagistic language in classical Chinese poems is well known. Huang (2004) constructed and compared ontologies based on the *Three Hundred Tang Poems* and poems by Su Shi. Lo (2008) built an ontology of imagery for classical Chinese poems. Using this ontology, Fang *et al.* (2009) built a parser that identifies imagistic language. However, the distribution of imagistic language and propositional language—i.e. whether their presence is concentrated in certain parts of a poem—has been limited mostly to qualitative analysis, notably by Kao and Mei (1971) and Levy (1988). This article fills this gap with an empirical study based on a treebank of classical Chinese poems.

3 Data

Often considered the pinnacle of classical Chinese poetry, the *Complete Tang Poems* (Peng, 1960) consists of poems by over 2,000 poets from the Tang Dynasty (618–907 CE). From this anthology, 971 poems have been syntactically analyzed in a treebank (Lee and Kong, 2012). These poems include the complete works by Wang Wei and Meng Haoran, and selections from Du Fu.[5] Wang and Du are regarded by many to be among the greatest Chinese poets of all times; Meng is often associated with Wang, due to the similarity in the style and content of their poems. We now outline the three kinds of annotations provided in the treebank.[6]

3.1 Part of speech

A POS tag is given to each word. For example, in Fig. 1, the word *xin* 'heart' is assigned the tag NN (common noun), and the word *zhuang* 'strong' is assigned VA (adjectival verb). The Penn Chinese Treebank (Xue *et al.*, 2005) is a widely used digital resource for Modern Chinese. Our treebank has largely adopted its word segmentation and POS tagging

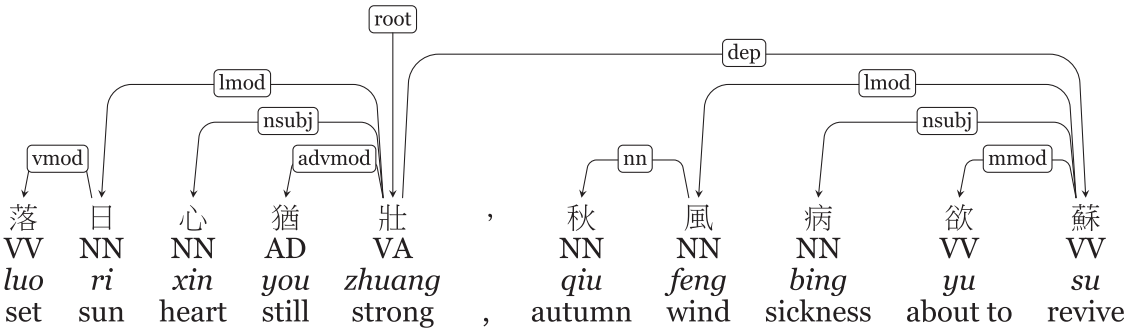

Fig. 1 Dependency parse tree for the third couplet in Table 1

guidelines. Tags specific to Modern Chinese, such as those for the usage of 得 *de* and 地 *de*, are discarded (Lee, 2012). Others are mapped to corresponding word classes in classical Chinese; for example, P ('preposition') is assigned on coverbs (Pulleyblank, 1995). To facilitate detection of parallelism, we also assign POS tags to individual characters within proper names. For example, rather than labeling the entire geographic entity 黃河 *huang he* 'Yellow River' as a proper noun, we annotate 黃 *huang* 'Yellow' as an adjective and 河 *he* 'river' as noun, so that its parallelism with common nouns such as 獨樹 *du shu* 'lone tree' would be apparent. Appendix Table A1 lists the POS tags referenced in this article.

3.2 Dependency relations

A parse tree is constructed for every couplet. Following the framework of dependency grammar (Gerdes *et al.*, 2014; Mel'cuk, 1998), the tree describes the syntactic structure of a sentence using a set of dependency relations. A dependency relation specifies how a word, called a 'child', modifies another word, its 'head'. For example, in Fig. 1, the word *xin* 'heart' is the child of the word *zhuang* 'strong' in the relation 'nsubj'; this means that *xin* modifies *zhuang* as its noun subject. The only word that has no head is the 'root' of the sentence, typically the main verb or adjectival verb, such as *zhuang* 'strong' (Fig. 1) or *qu* 'take' (Fig. 2). Using the forty-four relations in the Stanford dependencies for Modern Chinese (Chang *et al.*, 2009), the treebank added four more to account for phenomena unique to the classical language.[7] Appendix Table A2 lists the dependency relations referenced in this article.

The dependency framework overlaps in some aspects with the analysis by Wang (1994, pp. 183–4). The notation 'fN' in Table 2, for example, encodes the head–child relation between the noun and the adjective, though it does not specify the type of relation (i.e. adjectival modifier); it would have been explicitly labeled as 'amod' (adjectival modifier) in the treebank. Head–child relations are not encoded at the higher levels, for example, between the two 'fN' and 'V' in Table 2.

3.3 Poem types

The *Complete Tang Poems* contains two main types of poems. The 'recent-style' poems are regulated with syntactic, structural, and tonal rules, whereas the 'ancient-style' poems are unregulated. In the treebank, the type of each poem was annotated on the basis of its tonal patterns.

Recent-style poems are further subdivided into those with two couplets, called 'quatrains', and those with four couplets, called 'regulated verse'. Since quatrains and ancient-style poems do not exhibit regular patterns of parallelism, this study focuses on regulated verse. Of the 971 poems in the treebank, 617 are regulated verse; our data set thus includes 617 samples each for the first, second, third, and fourth couplets.

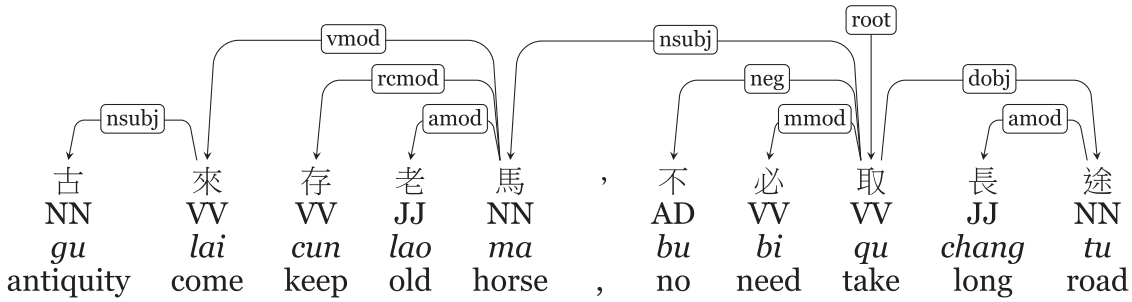

Fig. 2 Dependency parse tree for the fourth couplet in Table 1

4 Parallelism

According to Cai (2008, p. 165), a 'majority' of regulated verse 'begin with a non-parallel couplet, continue through two parallel couplets, and end with another non-parallel couplet'. In this section, we measure the extent to which the treebank reflects this pattern. To do so, we must first define a computable criterion for parallelism. We start with simple criteria for POS tag matching (Sections 4.1, 4.2), and then refine them with dependency relations (Section 4.3).

4.1 Exact POS match

The simplest criterion for parallelism is to require each character in the first line of a couplet to have the same POS tag as the character at the same position in the second line. The couplet below would satisfy this requirement:

明	月	松	間	照
ming	*yue*	*song*	*jian*	*zhao*
JJ	NN	NN	LC	VV
'bright'	'moon'	'pine'	'middle'	'shine'
'The bright moon shines amidst the pines'				
清	泉	石	上	流
qing	*quan*	*shi*	*shang*	*liu*
JJ	NN	NN	LC	VV
'clear'	'spring'	'rock'	'top'	'flow'
'The clear spring flows on the rocks'				

In both the first and second lines of this couplet, adjectives occupy the first position, with the words *ming* and *qing* both tagged as JJ; common nouns occupy the second position, with *yue* and *quan* both tagged as NN; and so on. Under this definition of parallelism, 20.4% of the second couplet and 20.8% of the third are parallel (Fig. 3), far higher than the rate of the first couplet (8.0%) and the final couplet (1.5%). This contrast is consistent with the common understanding that the middle part of a poem has more parallel couplets than the beginning or the end.

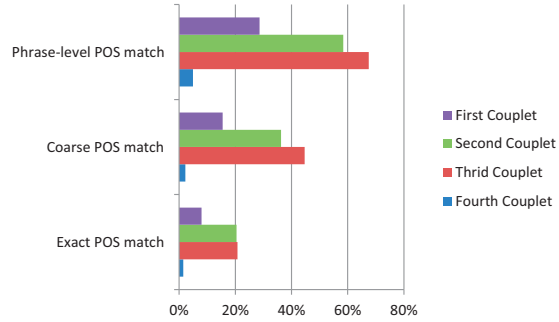

Fig. 3 Rate of parallelism under three different criteria: exact POS match (Section 4.1), coarse POS match (Section 4.2), and phrase-level POS match (Section 4.3)

4.2 Coarse POS match

Although the rate of exact POS match is highest among the middle couplets, it is barely over 20%, a far cry from a 'majority'. The fine-grained classification of nouns in the POS tagset is partly responsible. Different tags are assigned to common nouns (NN), proper nouns (NR), temporal nouns (NT), and pronouns (PN), even though all of these can be

Table 3 Mismatched POS tags in 'almost-parallel' couplets

Mismatched POS	Frequency (%)	Mismatched POS	Frequency (%)
Adjective (JJ)/Noun (N∗)	32.9	Cardinal Number (CD)/Adjective (JJ)	4.8
Adverb (AD)/Verb (V∗)	16.9	Localizer (LC)/Noun (N∗)	4.6
Noun (N∗)/Verb (V∗)	14.8	Preposition (P)/Verb (V∗)	3.4
Adjective (JJ)/Verb (V∗)	9.8	Localizer (LC)/Verb (V∗)	2.3
Adverb (AD)/Noun (N∗)	8.2	Determiner (DT)/Adjective (JJ); Cardinal Number (CD)/Ordinal Number (OD)	1.1

considered parallel.[8] We therefore collapsed all noun-related tags—NN, NR, NT, PN, as well as measure word (M)—into a general noun tag 'N∗'.

The POS tagset also distinguishes adjectival verbs (VA), copular verbs (VC), and existential verbs (VE) from other verbs (VV). We likewise consolidated all verb-related tags—VA, VC, VE, and VV—into a general verb tag 'V∗'.[9] Adjectival verbs are included in this general tag because, like other verbs, they also function as predicates (Pulleyblank, 1995). Thus, the word *zhuang* 'strong' in Fig. 1 is considered parallel to its counterpart, *su* 'revive', since both are V∗.

Under this 'coarse POS match', the rate of parallelism naturally increases for couplets in all positions. Among the first couplets, 15.5% are now parallel (Fig. 3). Parallel couplets remain very rare at the final position (2.2%). The rate of parallelism among the middle couplets (36.3% and 44.7%) continues to be significantly higher.

4.3 Phrase-level POS match

Coarse POS match still classifies a plurality of couplets as non-parallel. To see if the criterion should be further moderated, we examined the most frequent pairs of mismatched POS tags. We retrieved these pairs only from 'almost-parallel' couplets—i.e. those couplets with only one character pair that fails coarse POS match—since they were more likely meant to be parallel. Table 3 lists the most frequent pairs of mismatched POS tags among the 438 'almost-parallel' couplets.

In most pairs, although the two characters differ in POS, they both serve as modifiers in the same type of phrase. Consider the words *luo* 'set' and *qiu* 'autumn' in Fig. 1. Even though one is a verb and the other is a noun, they both modify the head noun of a NP (*ri* 'sun' and *feng* 'wind') and can be considered parallel. We thus propose a third criterion, 'Phrase-level POS match', to take this phenomenon into account.

NP-level POS match. The head noun of a NP may be modified by a number, a proper noun, an adjective, a color, a verb, or another noun.[10] Despite their different POS, these modifiers may be considered parallel in the context of an NP.[11]

This linguistic construction is responsible for the most frequent mismatch—adjective (JJ)/noun (N∗)—where both the adjective and the noun serve as modifiers in their respective NPs (Table 3). This construction also produces most of the mismatches for CD/JJ, DT/JJ, and CD/OD, and some for JJ/V∗. Hence, we propose 'NP-level POS match': two NPs are considered parallel when adjectives (JJ), determiners (DT), numbers (CD or OD), nouns (N∗), or verbs (V∗) modify their head nouns; or, more precisely, when they are the children of the head noun in the 'amod', 'det', 'nummod', 'ordmod', 'nn', or 'vmod' relation. Thus, *luo* 'set' and *qiu* 'autumn' in Fig. 1 are now parallel since, though one is a verb and the other a noun, they both modify the nouns following them.

An NP may also be headed by a localizer, which is typically modified by a noun child.[12] When this NP corresponds in a couplet to another NP with a head noun, the result is the mismatched pair LC/N∗ (Table 3). We also recognize a noun (N∗) to be parallel to the localizer (LC) in NP-level POS match.[13]

VP-level POS match. In a verb phrase, the head verb may be modified by a modal verb, an adverb, or a noun.[14] Again, despite their different POS, these modifiers may be considered parallel in this context.[15]

This linguistic construction produces the second most frequent mismatch, adverb (AD)/verb (V∗) (Table 3). It is also responsible for some of the

mismatches in AD/N∗ and N∗/V∗. Hence, we propose 'VP-level POS match': two VPs are parallel when adverbs (AD), nouns (NN), or verbs (V∗) modify their head verbs; or, more precisely, when they are children of verbs in a 'advmod', 'npadvmod', 'tmod', 'lmod', 'neg', 'dobj', 'mmod', 'rcomp', 'ccomp', 'conj', or 'dep' relation. For example, in Fig. 1, the two verb phrases 猶壯 *you zhuang* 'still strong' and 欲蘇 *yu su* 'about to revive' would be considered parallel, even though the former consists of an adverb-verb sequence, and the latter a verb-verb sequence. Likewise, the two verb phrases 抱琴 *bao qin* 'to hold a violin' and 垂釣 *sui diao* 'to fish with a drooping head' would also be considered parallel,[16] even though the former has a verb modified by a noun (*qin* 'violin'), and the latter has verb modified by an adverb (*sui* 'with a drooping head').

Another frequent mismatched pair is preposition (P)/verb (V∗), two closely related POS (Table 3). In classical Chinese, verbs are sometimes used as prepositions (Wang, 1994, p. 184), and in fact prepositions are considered coverbs (Pulleyblank, 1995). In a majority of the P/V∗ instances, the preposition and the verb are heads of their respective PP and VP, taking a noun as its object.[17] In VP-level POS match, we consider verbs and prepositions to be parallel.

4.4 Discussion

Under phrase-level POS match, 58.4% of the second couplets and 67.5% of the third couplets are considered parallel (Fig. 3). These rates, significantly higher than those of other couplets, affirm parallelism as a distinctive feature for the middle couplets. It is not, however, a rigid requirement; over one-third of them contain non-parallel elements. The most common mismatch, such as N∗/V∗, JJ/V∗, and AD/N∗, all involve a NP corresponding to a verb phrase, featured in couplets with partial parallelism (Owen, 1985, p. 90).

Among the final couplets, only 4.9% are parallel, confirming that a poet of regulated verse 'normally should not end a poem with a parallel couplet' (Cai, 2008, p. 165). However, it was the case that a poet 'could choose to begin with a parallel one' (Cai, 2008, p. 165), as was done in Table 1. In fact, 28.6% of the first couplets are parallel. Slightly more than half of these parallel first couplets belong to the so-called *touchun* (偷春) poem, where a parallel first couplet is followed by either a non-parallel second or third couplet (Wei, 2007). In the rest, the poet seemed to have preferred parallelism even though it was optional.

5 Imagistic Language and Propositional Language

Different parts of a poem can be distinguished not only by their degree of parallelism, as shown in the last section, but also by other linguistic characteristics. We now analyze their differences with respect to POS (Section 5.1), use of images (Section 5.2), syntactic continuity (Section 5.3), and mood (Section 5.4), to assess the extent to which 'the middle couplets . . . are imagistic in language and discontinuous in rhythm; the final couplet is propositional in language and continuous in rhythm' (Kao and Mei, 1971, p. 60).

5.1 Part of speech

The first three couplets all use more common nouns than the fourth (Fig. 4). Table 4 shows the ten most distinctive words for each couplet, as measured by log likelihood (Rayson, 2008). Common nouns, such as *shan* 'hill', *yu* 'rain', and *shui* 'water', dominate the lists for the first three couplets. In contrast, the list for the final couplet contains mostly function words, including a determiner (*he* 'what'), a

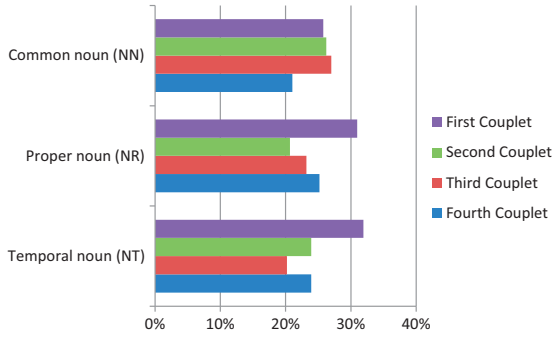

Fig. 4 Percentage of common nouns, proper nouns, and temporal nouns found in each of the four couplets

Table 4 Most distinctive words for the four couplets

First couplet		Second couplet		Third couplet		Fourth couplet	
Word	POS	Word	POS	Word	POS	Word	POS
山 *shan* 'hill'	NN	雨 *yu* 'rain'	NN	水 *shui* 'water'	NN	何 *he* 'what'	DT
南 *nan* 'south'	JJ	葉 *ye* 'leaf'	NN	酒 *jiu* 'wine'	NN	莫 *mo* 'not'	AD
秋 *qiu* 'autumn'	NN	雲 *yun* 'cloud'	NN	過 *guo* 'pass'	VV	是 *shi* 'be'	VC
野 *ye* 'wild'	JJ	心 *xin* 'heart'	NN	歌 *ge* 'song'	NN	應 *ying* 'should'	VV
州 *zhou* 'state'	NN	隨 *sui* 'follow'	VV	燕 *yan* 'swallow'	NN	看 *kan* 'look'	VV
道 *dao* 'road'	NN	開 *kai* 'open'	VV	石 *shi* 'rock'	NN	未 *wei* 'yet'	AD
公 *gong* 'sir'	NN	花 *hua* 'flower'	NN	香 *xiang* 'fragrant'	JJ	君 *jun* 'you'	PN
萬 *wan* 'ten thousand'	CD	松 *song* 'pine'	NN	藏 *cang* 'hide'	VV	誰 *shei* 'who'	PN
宮 *gong* 'palace'	NN	鳥 *niao* 'bird'	NN	舟 *zhou* 'boat'	NN	首 *shou* 'head'	NN
府 *fu* 'home'	NN	樹 *shu* 'tree'	NN	新 *xin* 'new'	JJ	意 *yi* 'sentiment'	NN

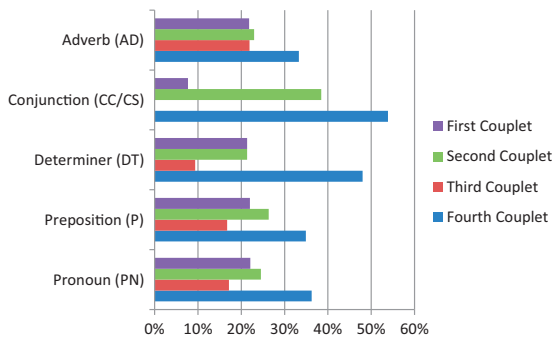

Fig. 5 Percentage of adverbs and function words found in each of the four couplets

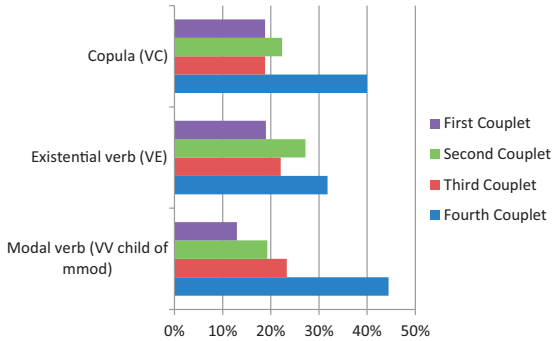

Fig. 6 Percentage of existential verbs, copular verbs, and modal verbs found in each of the four couplets

modal verb (*ying* 'should'), a copula (*shi* 'be'), two pronouns (*jun* 'you', *shei* 'who'), and two adverbs (*mo* 'not', *wei* 'yet'). Figures 5 and 6 confirm the concentration of function words in the final couplet. It has a higher share of conjunctions, determiners, prepositions, modal verbs, copulas, existential verbs, as well as adverbs, than any other couplet. These statistics corroborate the claim, on the one hand, that the middle couplets employ imagistic language, which prefers content words over function words (Cai, 2008, p. 163); and on the other hand, that the final couplet employs propositional language, which uses function words to specify relations between content words, and makes assertions with copulas and existential verbs.

So far, the first couplet largely resembles the middle couplets in their abundance of common nouns and paucity of function words. It outnumbers the middle couplets, however, in terms of proper nouns and temporal nouns (Fig. 4). Their

prevalence is consistent with the observation that 'time and place are usually indicated in the beginning of a poem' (Kao and Mei, 1971, p. 61). The first couplet in Table 1, for example, mentions Han and Jiang—both proper names referring to rivers—to establish the setting of the poem. Then, as the poem progresses to the middle couplet, 'references to a specific time and place seldom occur' (Cai, 2008, p. 165), giving way to timeless images. Although the most distinctive words for the first couplet are not proper or temporal nouns (Table 4), most of them are associated with locations (e.g. *shan* 'hill', *zhou* 'state', *dao* 'road', *gong* 'palace', *nan* 'south') and time (e.g. *qiu* 'autumn').

5.2 Images

The overuse of common nouns and underuse of function words tend to result in fragmentary, 'discontinuous' syntax. According to Kao and Mei (1971, p. 63), discontinuity 'isolates' NPs, and hence highlights them as images. In the extreme case, an entire couplet simply consists of NPs with no linkage between them. One such example is the first couplet in Table 1, which has but two juxtaposed NPs, headed by the nouns *ke* 'stranger' and *ru* 'scholar'. This structure can be identified by the root of its dependency tree. In most cases, the tree's root is the main verb, such as *qu* 'take' in Fig. 2. In the case where the sentence is simply an NP, the root is the head noun, as in *ke* 'stranger' in the first couplet in Table 1. As shown in Fig. 7, the first couplet has the most number of instances of common nouns serving as root; it is followed by the middle couplets, with the final couplet having the fewest instances. This indicates a greater presence of the imagistic language in the first three couplets.

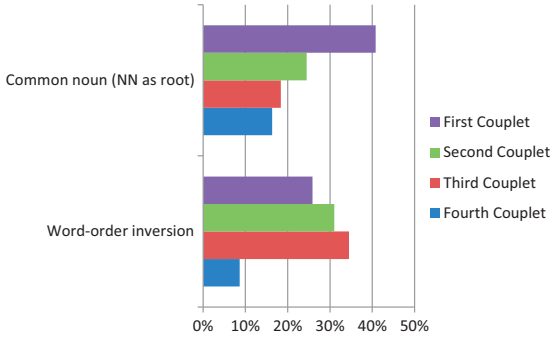

Fig. 7 Percentage of NP-only couplets and instances of word-order inversion (Section 5.2) found in each of the four couplets

Dislocation is another literary device for isolating an NP (Kao and Mei, 1971, p. 66). The NP is put at an unexpected location, usually moved forward, to seize the attention of the reader. For example, the second couplet in Table 1 highlights the noun *yue* 'moon' by placing it in front of *tong* 'with', even though a preposition normally precedes the prepositional object. Most instances of dislocation in the treebank break the normal word order of subject-verb-object in classical Chinese. In over 90% of the cases, the object is moved forward to yield either object-subject-verb or subject-object-verb; verb-object-subject accounts for most of the rest. Word-order inversion is considerably more common in the first three couplets (Fig. 7), again suggesting their greater emphasis on images compared to the final couplet.

5.3 Syntactic continuity

The poem in Table 1 progresses from syntactic discontinuity to continuity. The first couplet, an incomplete sentence with juxtaposed NPs, exemplifies syntactic discontinuity. In each of the middle couplets, its two lines feature complete sentences, but without any grammatical connection between them. The final couplet lies at the other end of the spectrum. Its two lines form one continuous sentence, since its main verb (*qu* 'take') is in the second line but its subject is (*ma* 'horse') in the first; therefore, the couplet can be properly understood only when its two lines are read together. This kind of couplet is an example of syntactic continuity, which may be defined as 'two lines in enjambment creating a continuous rhythm' (Kao and Mei, 1971, p. 120).

Dependency relations that link the two lines in a couplet—with the head in one line and the child in the other, forming a 'run-on line' (Kao and Mei, 1971, p. 57)—are indicative of syntactic continuity. In the case of the couplet in Fig. 2, the dependency relation 'noun subject' (nsubj) links the two lines. Another common structure places the main verb in

the first line and its clausal complement in the second, with the relation 'clausal complement' (ccomp) linking the two. Many sentences with this structure are 'pivotal sentences', where a noun serves simultaneously as a direct object of the main verb and as the noun subject in the complement.[18] Consider the couplet 請看石上藤蘿月, 已映洲前蘆荻花 'Please look, on the wall, the moon in the ivy/already, by the shores of the isle, lights the blossom on the reeds'.[19] The word *yue* 月 'moon' serves as the object of the verb *kan* 看 'look' in the first line, and also as the subject of the verb *yin* 映 'light' in the second. Figure 8 shows how often the 'nsubj' and 'ccomp' relations link the two lines of a couplet. The phenomenon of syntactic continuity is found considerably more often in the final couplet than elsewhere.

5.4 Mood

The final couplet contains 44% of the modal verbs, more than any other couplet (Fig. 6). The final couplet in Table 1, for example, contains a modal verb, *bi* 'need'. Another modal verb, *ying* 'should', is among the ten most distinctive words for the final couplet (Table 4). More generally speaking, there are more non-declarative sentences in the final couplet, which 'often departs from the simple declarative mood, in which case its mood may be interrogative, hypothetical, exclamatory, or imperative', to 'speak the voice of the poet' (Kao and Mei, 1971, p. 60). These moods are characteristic of the propositional language: it appeals to one's understanding, rather than one's imagination, as the imagistic language does.

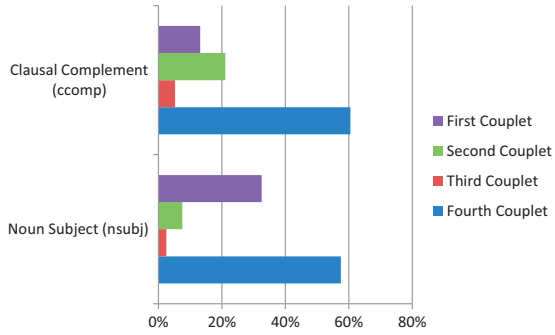

Fig. 8 Percentage of instances of syntactic continuity (Section 5.3) found in each of the four couplets

Besides modal verbs, questions and imperatives also tend to be packed into the final couplet. The word with the highest likelihood score for the final couplet is the interrogative pronoun *he* 'what, which' (Table 4). It is joined among the top ten by another interrogative pronoun, *shei* 'who'. The second most distinctive word for the final couplet is the adverb imperative *mo* 'not' (Table 4), which modifies a verb in the imperative mood. Also among the top ten is the second person pronoun *jun* 'you', often used in the imperative mood.

6 Conclusion

This article has presented the first quantitative study on syntactic properties in classical Chinese poems based on a dependency treebank. The study focused on the distribution of parallelism, imagistic language and propositional language in four-couplet poems in the 'regulated verse' genre. We proposed computable criteria to detect these linguistic phenomena. Unlike previous studies, these criteria leverage not only POS information but also syntactic structures.

For parallelism, we found that word-to-word, exact POS matching may not be the best criterion. Using a coarser tagset and phrase-level matching, our analyses showed that the middle part of a poem is often parallel, while the final part is rarely so. Parallelism seems to be optional for the first couplet, whose rate of parallelism lies in between.

For the use of imagistic language and propositional language, we explored a number of linguistic features—frequency of content words and function words, isolation of nouns, syntactic continuity, and non-declarative moods—that are indicative of each. Our analyses suggested that imagistic language is more present in the first three couplets, whose most distinctive words are common nouns; the first couplet resembles the middle ones, but uses more proper nouns and temporal nouns. In contrast, propositional language is prevalent in the final couplet, whose most distinctive words are function words.

From the vantage point of a much larger set of data, our results have substantiated previous qualitative studies on parallelism, imagistic language, and propositional language (Cai, 2008; Kao and Mei, 1971). In many other areas of research, such as semantic parallelism and imagery in poetic visions, existing qualitative analyses and conjectures may similarly be complemented by corpus-based methods. It is hoped that the quantitative approach in this article can be further developed to investigate other aspects of classical Chinese poetry and beyond, and contribute to the dialog between corpus linguistics and literary studies.

Funding

The work described in this article was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 11606515), and by a Strategic Research Grant (Project No. 7004338) from City University of Hong Kong. The third author completed this work while visiting City University of Hong Kong.

References

Cai, Z.-Q. (2008). *How to Read Chinese Poetry*. New York: Columbia University Press.

Cao, F. 曹逢甫 (1998). *A Linguistic Study of the Parallel Couplets in Tang Poetry*. Technical Report. Linguistics Graduate Institute, National Tsing Hua University, Taiwan.

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

Chen, W. 陳望道 (1957). 修辭學發凡 *[An investigation to Rhetoric]*. Taipei: Kaiming shuju.

Ding, B.-G., Huang, C.-N., and Huang, D.-G. (2005). Chinese main verb identification: from specification to realization. *Computational Linguistics and Chinese Language Processing*, **10**(1): 53–94.

Fang, A. C., Lo, F., and Chinn, C. K. (2009). *Adapting NLP and Corpus Analysis Techniques to Structured Imagery Analysis in Classical Chinese Poetry*. In *Proceedings of the Workshop on Adaptation of Language Resource and Technology to New Domains*, pp. 27–34.

Feng, X. 馮興煒 (1990). *Duiou zhishi* 對偶知識 *[Knowledge of Parallelism]*. Beijing: Luyou jiaoyu chubanshe.

Gerdes, K., Hajičová, E., and Wanner, L. (2014). *Dependency Linguistics: Recent Advances in Linguistic Theory Using Dependency Structures*. Amsterdam: John Benjamins Publishing Company.

He, Z., Liang, W., Li, L., and Tian, Y. (2007). SVM-based Classification Method for Poetry Style. In *Proceedings of IEEE International Conference on Machine Learning and Cybernetics*. Hong Kong, China: IEEE.

Hou, Y. and Frank, A. (2015). Analyzing Sentiment in Classical Chinese Poetry. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Beijing, China, Tokyo, Japan, pp. 15–24.

Hu, X., Williamson, N., and McLaughlin, J. (2005). Sheffield corpus of Chinese for diachronic linguistic study. *In Literary and Linguistic Computing*, 20(3): 281–293.

Huang, C.-R. (2004). Text-based Construction and Comparison of Domain Ontology: A Study Based on Classical Poetry. *In Proceedings of 18th Pacific Asia Conference on Language, Information and Computation*, Tokyo, Japan, pp. 17–20.

Huang, L. 黃麗敏 (2006). *The Study of Classical Poems of Tu-mu.* 杜牧古體詩研究 Master's Thesis, National Sun Yat-sen University, Taiwan.

Huang, L., Peng, Y., Wang, H., and Wu, Z. (2006). Statistical part-of-speech tagging for classical Chinese. *Lecture Notes in Computer Science*, 2448: 296–311.

Jakobson, R. (1966). Grammatical parallelism and its Russian facet. *Language*, **42**: 429.

Kao, J. and Jurafsky, D. (2012). A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. In *Proceedings of Workshop on Computational Linguistics for Literature*.

Kao, Y. and Mei, T. (1971). Syntax, diction, and imagery in T'ang Poetry. *Harvard Journal of Asiatic Studies*, Cambridge, MA: Harvard-Yenching Institute, **31**: 49–136.

Kaplan, D. and Blei, D. (2007). A Computational Approach to Style in American Poetry. In *Proceedings of IEEE Conference on Data Mining*. Omaha, Nebraska: IEEE.

Kugel, J. (1981). *The Idea of Biblical Poetry*. New Haven: Yale University.

Lee, J. (2012). A Classical Chinese Corpus with Nested Part-of-Speech Tags. In *Proceedings of Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*.

Lee, J. and Kong, Y. H. (2012). *A Dependency Treebank of Classical Chinese Poems*. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Lee, J. and Wong, T. S. (2012). Glimpses of Ancient China from Classical Chinese Poems. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*.

Levin, S. (1962). *Linguistic Structures in Poetry*. The Hague: Mouton.

Levy, D. (1988). *Chinese Narrative Poetry: The Late Han Through T'ang Dynasties*. Durham: Duke University Press.

Liu, C.-L., Wang, H., Cheng, W.-H., Hsu, C.-T., and Chiu, W.-Y. (2015). Color Aesthetics and Social Networks in Complete Tang Poems: Explorations and Discoveries. In *Proceedings of 29th Pacific Asia Conference on Language, Information and Computation*, pp. 132–41.

Lo, F. (2008). The Research of Building a Semantic Category System Based on the Language Characteristic of Chinese Poetry. In *Proceedings of the 9th Cross-Strait Symposium on Library Information Science* (in Chinese), Wuhan, China.

Jiang, L. and Zhou, M. (2008). Generating Chinese Couplets Using a Statistical MT Approach. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pp. 377–84.

Mel'cuk, I. (1998). *Dependency Syntax: Theory and Practice*. Albany, NY: State University of New York Press.

Owen, S. (1985). *Traditional Chinese Poetry and Poetics: Omen of the World*. Madison, WI: The University of Wisconsin Press.

Peng, D. 彭定求 (1960). *Quan Tang shi* 全唐詩 *[The Complete Shi Poetry of the Tang]*. Beijing: Zhonghua shuju.

Peyraube, A. (2016). *Ancient Chinese*. In Chan, S.-W. (ed.), *Routledge Encyclopedia of the Chinese Language*. London: Routledge.

Pu, Q. 浦起龍 (1961). *Du Du xin jie* 讀杜心解 *[My Insights of Reading Du Fu's Poems]*. Beijing: Zhonghua shuju.

Pulleyblank, E. (1995). *Outline of Classical Chinese Grammar*. Vaucouver: UBC Press.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4): 519–49.

Tan, R. 譚汝爲 (2003). *Shige xiuci jufa yu jianshang* 詩歌修辭句法與覽賞 *[The Rhetoric and Appreciation of Chinese Poems]*. Macau: Associação da Ciência Linguística de Macau.

Tong, P. 佟培基 (2000). *Meng Haoran shiji jian zhu* 孟浩然詩集箋注 *[The Commentaries on Meng Haoran's Poems]*. Shanghai: Shanghai Guji Chubanshe.

Wang, L. 王力 (1994). *Hanyu shiluxue* 漢語詩律學 *[The Metric of Chinese Poems]*. Hong Kong: Zhonghua shuju.

Wei, Q. 魏慶之 (2007). *Shiren yuxie* 詩人玉屑 *[Essence from Poets]*. Beijing. Zhonghua shuju.

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: 207–38.

Yu, P. (1987). *The Reading of Imagery in the Chinese Poetic Tradition*. Princeton, NJ: Princeton University Press.

Yuan, X. 袁行霈 (ed.) (2005). *The History of Chinese Literature*, vol.2《中国文学史(第二卷)》(in Chinese). Beijing: Higher Education Press.

Zhang, X and Lapata, M. (2014). Chinese Poetry Generation with Recurrent Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 670–80.

Notes

1 English translations are taken from Cai (2008, p. 174).
2 A quantitative study on the third requirement, related meaning, would require semantic annotation on the poems. Another important ingredient of parallelism, tonal correspondence, has been treated elsewhere (e.g. Wang, 1994).
3 Various POS tagsets have been proposed for ancient Chinese; see for example Hu *et al.* (2005), Huang *et al.* (2006), and Wang (1994).
4 Alternatively, 'Precipitous rocks cause fountain sound to gurgle, sun rays chill blue pines' (Kao and Mei, 1971, pp. 67–8).
5 There are 380 poems by Wang, taken from Vol. 125—6 from Peng (1960); 270 poems by Meng, taken from

Vol. 1—3 from Tong (2000); 321 poems by Du, all regulated verse from Vol. 3—5 from Pu (1961).

6 The treebank is available for research purpose upon request to the first author.

7 The four additional relations were locative modifier, oblique objects, NP as adverbial modifier, and indirect objects. The interested reader is referred to Lee and Kong (2012) for details.

8 For example, Wang (1994) considered pronouns and common and proper nouns to be parallel (p. 186, 227), and did not differentiate temporal nouns from other nouns (p. 183).

9 For example, Wang (1994) considered verbs to be parallel to adjectives used as predicates (Table 2), and did not differentiate existential and copular verbs from other verbs (p. 183).

10 Wang (1994) gives as example the number *yi* 'one' in *yi qiu* 'one autumn' — 秋; the proper noun *jing* 'Jing' in *jing men* 'Gate of Jing' 荆門; the adjective *ming* 'bright' in *ming xing* 'bright star' 明星; the color *bai* 'white' in *bai fa* 'white hair' 白髮; the verb *luo* 'setting' in *luo ri* 'setting sun' 落日; and the noun *qiu* 'autumn' in *qiu shui* 'autumn waters' 秋水.

11 For example, Wang (1994) considers adjective to be parallel to color (p. 238), number (p. 190), proper nouns (p. 226, 251), common noun (p. 246), verb (p. 190), and pronouns to be parallel to numbers (p. 237), when they serve as noun modifiers.

12 For example, the localizer *xia* 'under' is modified by *tian* 'heaven' to form *tian xia* 'all under heaven' 天下

13 Cf. Wang (1994, pp. 191, 239, 240).

14 Wang (1994) gives as examples the adverb *bu* 'not' in *bu jin* 'do not end' 不盡, and the noun *bian* 'boundary' in *wu bian* 'have no boundary' 無邊

15 For example, Wang (1994) considers verb phrases with either a 'verb-noun' sequence or an 'adverb-verb' to be parallel (p. 243).

16 Taken from the third couplet of a poem by Meng Haoran: 抱琴來取醉, 垂釣坐乘閒

17 Consider the couplet 鳥過煙樹宿,螢傍水軒飛 taken from《閑園懷蘇子》by Meng Haoran. The preposition *bang* 傍 'along' takes the noun *shui* 水 'water' as its prepositional object, and the verb *guo* 過 'pass' takes the noun *yan* 煙 'mist' as its direct object.

18 While the annotation in the treebank considers the second verb as a clausal complement, many alternative structures have also been proposed (Ding *et al.*, 2005).

19 English translation taken from Kao and Mei (1971, p. 116).

Appendix

Table A1 Part-of-speech tags

AD	Adverb
CC	Coordinating conjunction
CD	Cardinal number
CS	Subordinating conjunction
DT	Determiner
JJ	Adjective
LC	Localizer
M	Measure word
NN	Other noun
NT	Temporal noun
NR	Proper noun
OD	Ordinal number
P	Preposition
PN	Pronoun
PU	Punctuation
VA	Adjectival predicate
VC	Copula
VE	Existential verb
VV	Other verb

Table A2 Dependency relations

advmod	Adverbial modifier
amod	Adjective modifier
ccomp	Clausal complement
conj	Conjunct
dep	Default dependency
det	Determiner
dobj	Direct object
lmod	Locative modifier
mmod	Modal verb modifier
neg	Negation
nn	Noun compound modifier
npadvmod	Noun as adverbial modifier
nsubj	Nominal subject
nummod	Cardinal number modifier
ordmod	Ordinal number modifier
rcmod	Relative clause modifier
rcomp	Resultative complement
tmod	Temporal modifier
vmod	Verb modifier