A Dependency Treebank of Classical Chinese Poems

John Lee and Yin Hei Kong
The Halliday Centre for Intelligent Applications of Language Studies
Department of Chinese, Translation and Linguistics
City University of Hong Kong
{jsylee,yhkong}@cityu.edu.hk

Abstract

As interest grows in the use of linguistically annotated corpora in research and teaching of foreign languages and literature, treebanks of various historical texts have been developed. We introduce the first large-scale dependency treebank for Classical Chinese literature. Derived from the Stanford dependency types, it consists of over 32K characters drawn from a collection of poems written in the 8th century CE. We report on the design of new dependency relations, discuss aspects of the annotation process and evaluation, and illustrate its use in a study of parallelism in Classical Chinese poetry.

1 Introduction

Recent efforts in creating linguistically annotated text corpora have overwhelmingly focused on modern languages. Among the earliest and most well-known are the part-of-speech (POS) tagged Brown Corpus (Francis & Kučera, 1982), and the syntactically analyzed Penn Treebank (Marcus et al., 1993). However, the first digital corpus, which emerged soon after the invention of computers, had as its subject matter a collection of 13th-century texts --- in 1949, Roberto Busa initiated the POS tagging of the complete works of Thomas Aquinas, written in Latin.

In the past decade, Humanities scholars have begun to use digital corpora for the study of ancient languages and historical texts. They come in a variety of languages and genres, from Old English (Taylor et al., 2003) to Early New High German (Demske et al., 2004) and Medieval Portuguese (Rocio et al. 2000); and from poetry (Pintzuk & Leendert, 2001) to religious texts such as the New Testament (Haug & Jøhndal, 2008) and the Quran (Dukes & Buckwalter, 2010). They are increasingly being leveraged in teaching (Crane et al., 2009) and in research (Lancaster, 2010).

This paper describes the first large-scale dependency treebank for Classical Chinese. The treebank consists of poems from the Tang Dynasty (618 – 907 CE), considered one of the crowning achievements in traditional Chinese literature. The first half of the paper reviews related work (section 2), then describes the design of the treebank (section 3), its text and evaluation (section 4). The second half shows the research potentials of this treebank with a study on parallelism in (section 5).

2 Previous Work

Existing linguistic resources for Chinese is predominantly for the modern language. This section first describes the major Modern Chinese treebanks on which we based our work (section 2.1), then summarizes previous research in word segmentation and POS tagging, two pre-requisites for building a Classical Chinese treebank (section 2.2).

2.1 Modern Chinese

Most treebanks have been annotated under one of two grammatical theories, the phrase structure grammar, which is adopted by the Penn Treebank (Marcus et al., 1993), or dependency grammar, adopted by the Prague Dependency Treebank

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 191–199,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

(Hajic, 1998). The most widely used treebank for Modern Chinese, the Penn Chinese Treebank (Xue et al., 2005), belongs to the former kind.

Rather than encoding constituency information, dependency grammars give information about grammatical relations between words. Modern Chinese has been analyzed in this framework, for example at Stanford University (Chang et al., 2009). The dependency relations follow the design principles of those initially applied to English (de Marneffe and Manning, 2008), with a few added relations to accommodate Chinese-specific features, such as the "ba"-construction. Their POS tagset is borrowed from that of the Penn Chinese Treebank.

2.2 Classical Chinese

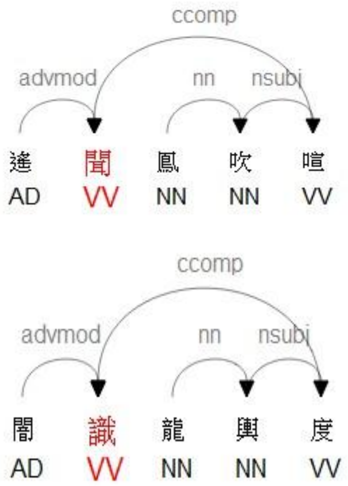
Like its modern counterpart, two pre-requisites for constructing a Classical Chinese treebank are word segmentation and part-of-speech tagging. In this section, we first summarize existing POS tagging frameworks, then describe the only current treebank of Classical Chinese.

Word boundaries and parts-of-speech tags have been added to the Academia Sinica Ancient Chinese Corpus (Wei et al., 1997) and the Sheffield Corpus of Chinese (Hu et al., 2005). Since there is not yet a scholarly consensus on word segmentation in Chinese (Feng 1998), it is not surprising that there are wide-ranging levels of granularity of the POS tagsets. They range from 21 tags in (Huang et al., 2002), 26 in the Peking University corpus (Yu et al., 2002), 46 in the Academia Sinica Balanced Corpus (Chen et al., 1996), to 111 in the Sheffield Corpus of Chinese (Hu et al., 2005). This treebank uses a system of nested POS tags (Lee, 2012), which accommodates different policies for word segmentation and maximize interoperability between corpora.

The only previous syntactic treebank for Classical Chinese is a constituent-based one (Huang et al., 2002), composed of 1000 sentences from pre-Tsin Classical Chinese. No word segmentation was performed for this treebank.

3 Treebank design

Although Classical Chinese is not mutually intelligible with Modern Chinese, the two share considerable similarities in vocabulary and grammar. Given the seminal work already achieved for Modern Chinese, our principle is to borrow from existing annotation framework as much as possible. For example, our POS tagset is based on that of the Penn Chinese Treebank, after a slight revision of its 33 tags (Lee, 2012). This approach not only gives users a familiar point of reference, and also makes the treebank interoperable with existing Modern Chinese resources. Interoperability allows the potential of bootstrapping with Modern Chinese data, as well as contrastive studies for the two languages.



遙	聞	鳳	吹	喧
'far'	'hear'	'phoenix'	'call'	'make noise'
[I] hear from afar the call of the phoenix making noise.				
闇	識	龍	輿	度
'faint'	'sense'	'dragon'	'carriage'	'come'
[I] faintly sense the dragon-decorated carriage coming.				

Figure 1. Dependency trees of two adjacent 5-character lines (forming a parallel *couplet*)[1]. The POS tags are based on (Xue et al., 2005); the dependency relations on (Chang et al., 2009). The two lines are perfectly parallel both in terms of POS and dependencies.

A dependency framework is chosen for two reasons. First, words in Classical Chinese poems, our target text (section 4), tend to have relatively free word order. Dependency grammars can handle this phenomenon well. Second, our treebank is expected to be used pedagogically, and we expect explicit grammatical relations between words to be helpful to students. These relations also encode

[1] From Wang Wei 《奉和聖制御春明樓臨右相園亭賦樂賢詩應制》

semantic information, which lend themselves to meaning extraction applications.

Our set of dependency relations is based on those developed at Stanford University for Modern Chinese (see section 2.2). Our approach is to map their 44 dependency relations, as much as possible, to Classical Chinese. Modern Chinese, a non-inflectional language, does not mark many linguistic features, including person, gender, and number, etc. It uses a small number of function words to encode other features, such as tense, voice, and case. Many of these function words do not exist in Classical Chinese. In particular, prepositions are rare[2]; instead, nouns expressing time, locations, instruments, indirect recipients, etc., modify the verb directly. This phenomenon prompted the introduction of two new relations "locative modifiers" (section 3.1) and "oblique objects" (section 3.2); and the re-instatement of two relations, "noun phrases as adverbial modifiers" (section 3.3) and "indirect objects", from the Stanford dependencies (de Marneffe and Manning, 2008) that are excluded from the Modern Chinese variant (Chang et al., 2009) . An overview is provided in Table 1.

Dependency	Stanford English	Stanford Modern Chinese	This paper
Direct object (dobj)	√	√	√
Indirect object (iobj)	√		√
Locative modifier (lmod)			√
Noun phrase as adverbial modifier (npadvmod)	√		√
Oblique objects (obl)			√
Concessive, temporal, conditional, and causal modifier (conc, temp, cond, caus)			√

Table 1. Comparison of our set of dependency relations with the Stanford dependencies for English (de Marneffe and Manning, 2008) and for Modern Chinese (Chang et al., 2009). All other relations from Stanford Modern Chinese are retained and are not listed here.

3.1 Locative modifiers

To indicate time, English usually requires a preposition (e.g., '*on* Monday'), but sometimes does not (e.g., 'today'). For the latter case, the bare noun phrase is considered a "temporal modifier" in a tmod relation with the verb in (de Marneffe and Manning, 2008).

Similarly, to indicate locations, a preposition is normally required in English (e.g., '*on* the hill'). However, in Classical Chinese, the preposition is frequently omitted, with the bare locative noun phrase modifying the verb directly. To mark these nouns, we created the "locative modifier" relation (lmod). Consider sentence (1) in Table 2. Although the word "hill" occupies the position normally reserved for the subject, it actually indicates a location, and is therefore assigned the lmod relation. In sentence (2), the locative noun 'alley' is placed after the verb.

3.2 Oblique objects

Oblique objects are a well-known category in the analysis of ancient Indo-European languages, for example Latin and ancient Greek. In the PROIEL treebank (Haug and Jøhndal, 2008), for example, the "oblique" (obl) relation marks arguments of the verb which are not subjects or non-accusative 'objects'. These are most commonly nouns in the dative or ablative case, as well as prepositional phrases. It is believed that oblique objects exist in Classical Chinese, but have been replaced by prepositional phrases in Modern Chinese (Li and Li, 1986).

The obl relation is imported to our treebank to mark nouns that directly modify a verb to express means, instrument, and respect, similar to the functions of datives and ablatives. They typically come after the verb. In sentence (6) in Table 2, the noun 'cup' is used in an instrumental sense to modify 'drunk' in an obl relation.

3.3 Noun phrase as adverbial modifier

A temporal modifier such as "today" is an example where a noun phrase serves as an adverbial modifier in English. This usage is more general and extends to other categories such as floating reflexives (e.g., it is *itself* adequate), and other PP-like NPs (e.g., two *times* a day). These noun phrases are marked with the relation npadvmod in (de Marneffe and Manning, 2008).

[2] Classical Chinese has a category of verbs called "coverbs" which function like prepositions, but are less frequently used. (Pulleyblank, 1995).

Locative modifier				
千	山	響	杜鵑	
‘thousand’	‘hill’	‘make sound’	‘bird’	
(1) Birds are singing on a thousand hills. lmod(‘make sound’, ‘hill’)				
五	馬	驚	窮	巷
‘five’	‘horse’	‘scare’	‘end’	‘alley’
(2) Five horses are scared at the end of the alley.				
lmod(‘scare’, ‘alley’)				
Indirect Objects				
寄	語	邊	塞	人
‘send’	‘word’	‘edge’	‘region’	‘person’
(3) [I] send a word to the person at the frontier.				
iobj(‘send’, ‘person’)				
Noun phrase as adverbial modifier				
風	物	自	瀟灑	
‘scene’	‘thing’	‘self’	‘natural, unrestrained’	
(4) The scenes are being natural and unrestrained in themselves.				
npadvmod(‘natural’, ‘self’)				
年	年	梁	甫	吟
‘year’	‘year’	‘Liang’	‘Fu’	‘song’
(5) [He sings] the Liangfu Song year after year.				
npadvmod(‘song’, ‘year’)				
Oblique objects				
同	醉	菊	花	杯
‘together’	‘drunk’	‘chrysan-themus’	‘flower’	‘cup’
(6) [We] get drunk together with the chrysanthemus cup.				
obl(‘drunk’, ‘cup’)				

Table 2. Example sentences illustrating the use of the dependency relations lmod (locative modifier), iobj (indirect object), npadvmod (noun phrase as adverbial modifier), and obl (oblique object)[3].

In Modern Chinese, this usage is less frequent[4], perhaps leading to its exclusion in (Chang et al., 2009). In contrast, in Classical Chinese, nouns function much more frequently in this capacity, expressing metaphoric meaning, reasons, moods, repetitions, etc., and typically preceding the verb (Li and Li, 1986). Sentences (4) and (5) in Table 2 provide examples of this kind, with the noun ‘self’ as a reflexive, and the noun ‘year’ indicating repetition.

3.4 Indirect objects

The double object construction contains two objects in a verb phrase. The direct object is the thing or person that is being transferred or moved (e.g., “he gave me a *book”*); the indirect object is the recipient (“he gave *me* a book”). In inflected languages, the noun representing the indirect object may be marked by case. Since Classical Chinese does not have this linguistic device, the indirect object is unmarked; we distinguish it with the “indirect object” label (iobj).

The iobj label exists in Stanford English dependencies (de Marneffe and Manning, 2008), but was not included in the Modern Chinese version (Chang et al., 2009), likely due to its infrequent appearance in Modern Chinese. It is re-instated in our Classical Chinese treebank. Sentence (3) in Table 2 provides an example, with ‘word’ as the direct object and ‘person’ as the indirect.

3.5 Absence of copular verbs

In a copular construction such as “A is B”, A is considered the “topic” (top) of the copular verb “is” (Chang et al., 2009). The copular, however, is rarely used in Classical Chinese (Pulleyblank, 1995). In some cases, it is replaced by an adverb that functions as a copular verb. If so, that adverb is POS-tagged as such (VC) in our treebank, and the dependency tree structure is otherwise normal. In other cases, the copular is absent altogether. Rather than inserting implicit nodes as in (Haug and Jøhndal, 2008), we expand the usage of the top relation. It usually connects the subject (“A”) to the copular, but would in this case connect it with the noun predicate (“B”) instead. In the example sentence below, the relation top(‘capable’, ‘general’) would be assigned.

將軍	武	庫	才
‘general’	‘weapon’	‘warehouse’	‘capable’

The general [is] a capable manager of the arsenal[5].

[3] The verses are from Wang Wei 《送梓州李使君》,《鄭果州相過》; Meng Haoran 《同張明府清鏡歌》,《宴包二融宅》,《與白明府游江》,《和賈主簿弁九日登峴山》.

[4] Mostly restricted to temporal and location modifiers.

3.6 Discourse relations

Two clauses may be connected by a discourse relation, such as causal or temporal. In English, these relations may be explicitly realized, most commonly by discourse connectives, such as ‘because’ or ‘when’. Even in the absence of these connectives, however, two adjacent clauses can still hold an implicit discourse relation. A detailed study, which resulted in the Penn Discourse Treebank (Prasad et al., 2008), found that explicit relations outnumber implicit ones in English, but the latter is nonetheless quite common and can be annotated with high inter-annotator agreement.

Temporal relation				
為	童	憶	聚	沙
‘be’	‘child’	‘remember’	‘gather’	‘sand’
(1) [When I] was a child, [I] remember [playing] a game with sand.				
`dep-temp`(‘remember’, ‘be’)				
Causal relation				
不	才	明	主	棄
‘not’	‘capable’	‘good’	‘ruler’	‘forsake’
(2) The good ruler does not appoint me [as an official], [because] I am not capable.				
`dep-caus`(‘forsake’, ‘capable’)				
Concessive relation				
國	破	山	河	在
‘country’	‘broken’	‘mountain’	‘river’	‘exist’
(3) [Although] the country is broken, the mountains and the rivers still stay.				
`dep-conc`(‘exist’, ‘broken’)				

Table 3. Example sentences illustrating the use of discourse labels for discourse relations[6].

In many ancient languages, explicit realization of discourse relations is less frequent. In Latin and Ancient Greek, for instance, these connectives are often replaced by a participial clause. The participle is marked only by the genitive or ablative case, leaving the reader to decide from context how it relates to the main clause. As a non-inflectional language, Classical Chinese cannot use this device, and instead typically constructs a complex sentence with a series of verbs without any marking (Pulleyblank, 1995). For example, sentence (2) in Table 3 literally says ‘not capable, good ruler forsake’; the onus is put on the reader to interpret the first two characters to form a clause that provides the reason for the rest of the line.

This condensed style of expression often erects a barrier for understanding. Although the focus of the treebank is on syntax rather than discourse, we decided to annotate these relations. Implicit connectives are more difficult to achieve inter-annotator agreement (Prasad et al., 2008); since they are mostly implicit in Classical Chinese, we adopted a coarse-grained classification system, rather than the hierarchical system of sense tags in the Penn Discourse Treebank. More precisely, it contains only the four categories posited by (Wang, 2003) --- causal, concessive, temporal, and conditional. Table 3 gives some examples.

When it is impossible to determine the discourse relation between two lines, the default “dependent” (`dep`) label is assigned. This label is originally used when “the system is unable to determine a more precise dependency relation between two words” (de Marneffe and Manning, 2008).

4 Data

Among the various literary genres, poetry enjoys perhaps the most elevated status in the Classical Chinese tradition. The Tang Dynasty is considered the golden age of *shi*, one of the five subgenres of Chinese poetry[7]. The *Complete Shi Poetry of the Tang* (Peng, 1960), originally compiled in 1705, consists of nearly 50,000 poems by more than two thousand poets. This book is treasured by scholars and the public alike. Even today, Chinese people informally compose couplets (see section 5), in the style of *shi* poetry, to celebrate special occasions such as birthdays. Indeed, NLP techniques have been applied to generate them automatically (Jiang and Zhou, 2008).

4.1 Material

This treebank contains the complete works, a total of over 32,000 characters in 521 poems, by two Chinese poets in the 8th century CE, Wang Wei and Meng Haoran. Wang, also known as the Poet-Buddha (*shifo* 詩佛), is considered one of the three most prominent Tang poets. Meng is often asso-

[5] From Meng Haoran 《與張折衝遊耆闍寺》

[6] From top to bottom, Meng Haoran 《登龍興寺閣》,《歲暮歸南山》, and Du Fu 杜甫《八陣圖》

[7] The other four genres are *ci*, *fu*, *qu*, and *sao*.

ciated with Wang due to the similarity of his poems in style and content.

Aside from the dependency relations, word boundaries and POS tags, the treebank contains a number of metadata. For each character, the tone is noted as either level (*ping* 平) or oblique (*ze* 仄). Each poem is also recorded for its title, author, and genre, which may be 'recent-style' (*jintishi* 近體詩) or 'ancient-style' (*gutishi* 古體詩).

This choice of our text stems from three motivations. Classical Chinese is typically written in a compressed style, especially so with poetry, where the word order is relatively flexible, and grammatical exceptions are frequent. These characteristics pose a formidable challenge for students of Classical Chinese, for whom Tang poetry often forms part of the introductory material. It is hoped that this treebank will serve a pedagogical purpose. Second, this challenging text makes it more likely that the resulting dependency framework can successfully handle other Classical Chinese texts. Third, Tang poetry is an active area of research in Chinese philology, and we aspire to contribute to their endeavor.

4.2 Inter-annotator agreement

Two annotators, both university graduates with a degree in Chinese, created this treebank. To measure inter-annotator agreement, we set apart a subset of about 1050 characters, on which both of them independently perform three tasks: POS tagging, head selection, and dependency labeling.

Their agreement rate is 95.1%, 92.3%, and 91.2% for the three respective tasks. For POS tagging, the three main error categories are the confusion between adverbs (AD) and verbs with an adverbial force, between measure words (M) and nouns (NN), and between adjectives (JJ) and nouns. The interested reader is referred to (Lee, 2012) for a detailed analysis.

These differences in POS tags trickle down to head selection and dependency labeling. In fact, all words which received different POS tags also received different dependency relations. To illustrate with a disagreement between adverb and verb, consider the following sentence. The word 恐 *kong* 'afraid' may be considered as an adverb, expressing the psychological state for the verb 'attract'; or, alternatively, as a verb in its own right. Depending on the decision, it bears either the relation advmod or root.

恐	招	負	時	累
'afraid'	'attract'	'burden'	'fame'	'affect'

[I am] afraid [I] will attract and be burdened by fame[8].

Some differences are genuine alternative annotations, resulting from a mixture of polysemy and flexible word order. Consider the sentence 簞食伊何 *dan shi yi he*, consisting of four characters meaning, in order, 'bowl / blanket', 'food', a copular or a particle, and 'what'. If the meaning 'bowl' and copular is taken, it means 'What food is contained in that bowl?' In this case, the relation clf is required for 簞 *dan*, and 伊 *yi* is the root word. Alternatively, if the meaning 'blanket' and particle is taken, it is interpreted as 'What food is placed on the blanket?' Here, *dan* takes on the relation nn, and the root word would be 何 *he* instead.

5 Application: Parallel Couplets

We now demonstrate one use of this treebank by analyzing a well-known but understudied feature of Classical Chinese poetry: the parallel couplets.

5.1 Introduction

Parallelism in poetry refers to the correspondence of one line with another; the two lines may bear similar or opposite meaning, and have comparable grammatical constructions. This phenomenon is perhaps most well known in classical Hebrew poetry, but it is also one of the defining features of Chinese poetry; "it pervades their poetry universally, forms its chief characteristic feature, and is the source of a great deal of its artificial beauty", observed Sir John Francis Davis, author of one of the earliest commentaries on Chinese poetry published in the West (Davis 1969).

The lines in a Chinese poem almost always contain the same number of characters, most commonly either five or seven characters. This exact equality of the number of characters makes it especially suited for expressing parallelism, which became a common feature ever since 'recent-style' poetry (section 4.1) was developed during the Tang

[8] From Wang Wei 《贈從弟司庫員外絿弟》

Dynasty. Unlike those in ‘ancient-style’, poems of this style are tonally regulated and assume a high degree of parallelism within a couplet, i.e., two adjacent lines. See Figure 1 for an example.

5.2 Methodology

The couplet in Figure 1 is undisputedly symmetric, both in terms of POS tags and dependency labels. The definition for parallelism is, however, quite loose; in general, the corresponding characters must ‘agree’ in part-of-speech and have related meaning. These are unavoidably subjective notions.

While a vast amount of Tang poems have been digitized, they have not been POS-tagged or syntactically analyzed in any significant amount. It is not surprising, therefore, that no large-scale, empirical study on how, and how often, the characters ‘agree’. There have been a study on 1,000 couplets (Cao, 1998), and another on a small subset of the poems of Du Mu (Huang, 2006), but neither clarify the criteria for parallelism. We undertake a descriptive, rather than prescriptive, approach, using only the treebank data as the basis.

Character-level parallelism. Even given the POS tags, this study is not straightforward. The naive metric of requiring exactly matched POS tags yields a parallel rate of only 74% in the corpus as a whole. This figure can be misleading, because it would vary according to the granularity of the POS tagset: the more fine-grained it is, the less agreement there would be. As a metric for parallelism, it has high precision but lower recall, and would only be appropriate for certain applications such as couplet generation (Jiang and Zhou, 2008).

Equivalence	**POS tags and dependency links**
Noun modifier	`CD`, `OD`, `JJ`, `DT`
Verbs	`BA`, `<verb>`, and `P` (head of `pobj` or `plmod`)
Adverbials	`AD`, `CS`, `<verb>` (head of `mmod`), `<noun>` (head of `npadvmod`)
Adjectival	`<noun>` (head of `nn` or `assmod`), `<verb>` (head of `vmod`), `JJ` (head of `amod`)
Nouns	`<noun>`, `<verb>` (head of `csubj` or `csubjpass`), `M` (except `clf`)

Table 4. Equivalence sets of POS tags for the purpose of parallelism detection. `<noun>` includes `NN`, `NT`, `NR`, `PN`; `<verb>` includes `VA`, `VC`, `VE`, `VV`.

By examining portions of the regulated verse where parallelism is expected, we derived five ‘equivalence sets’ of POS tags, shown in Table 4. Two tags in the same set are considered parallel, even though they do not match. In many sets, a tag needs to be qualified with its dependency relations, since it is considered parallel to other members in the set only when it plays certain syntactic roles. When applying these equivalence sets as well as exact matching, the parallel rate increases to 87%.

The algorithm is of course not perfect[9]. It cannot detect, for example, parallelism involving the use of a polysemous character with a ‘out-of-context’ meaning (*jieyi* 借義). For instance, the character 者 *zhe*, the fourth character in the second line in the couplet[10] “欲就終焉志，恭聞智者名,” means ‘person’. On the surface, it does not match its counterpart, 焉 *yan*, the fourth character in the first line, since *yan* is a sentence particle and *zhe* is a noun. However, the poet apparently viewed them as parallel, because *zhe* can also function as a sentence particle in other contexts.

Phrase-level parallelism. The character-level metric, however, still rejects some couplets that would be deemed parallel by scholars. Most of these couplets are parallel not at the character level, but at the phrase level.

A line in a ‘recent-style’ poem is almost always segmented into two syntactic units (Cai, 1998). A pentasyllabic (5-character) line is composed of a disyllabic unit (the first two characters) followed by a trisyllabic unit (the last three characters)[11]. Consider two corresponding disyllabic units, 抱琴 *bao qin* ‘hold’ ‘violin’, and 垂釣 *sui diao* ‘look down’ ‘fish’. They are tagged as *bao*/VV *qin*/NN and *sui*/AD *diao*/VV, respectively. There is a complete mismatch at the character level: *bao* is a verb but *sui* is an adverb; *qin* is a noun but *diao* is a verb. Taken as a whole, however, both units are verb phrases describing an activity (‘to hold a violin’ and ‘to fish while looking down’), and so the poet likely considered them to be parallel at the unit, or phrase, level.

[9] The quality of these equivalence sets were evaluated on 548 characters. The human expert agrees with the decision of the algorithm 96.4% of the time at the character level, and 94% of the time at the phrase level.

[10] From Meng Haoran 《陪張丞相祠紫蓋山，途經玉泉寺》

[11] Equivalently, the seven characters in the heptasyllabic regulated verse are segmented in a 4+3 fashion.

The dependencies provide a convenient way to gauge the level of parallelism at the phrase level. One can extract the head word in the corresponding units in the couplet (*bao*/VV and *diao*/VV in the example above), then compare their POS tags, using the algorithm for character-level parallelism describe above.

5.3 Results

The results are shown in Table 5. All couplets from an 'ancient-style' poem are considered "parallelism optional". A couplet from a 'recent-style' poem with eight or more lines[12] is either "parallelism not expected", if it is the first or last couplet in the poem; or "parallelism expected", if it is in the middle of the poem. We first determine whether a character is parallel to its counterpart in the couplet at the character level; if not, then we back off to the phrase level.

In the "parallelism expected" category, the couplets of Wang are highly parallel, at both the character (91%) and phrase levels (95%). This is hardly surprising, given that his poems are highly regarded. It is notable, however, that the proportion is still relatively high (57% at the character level) even among those couplets for which parallelism is not expected, suggesting that the poet placed a high view on parallelism. He also employed much parallelism (64% at the character level) in 'ancient-style' poems, perhaps to aim at a higher artistic effect.

Among the couplets of Wang which are not parallel at the phrase level, the most frequent combination is a verb phrase matching a noun phrase. The verb, as the second character, is modified by an adverb; the noun, also as the second character, is modified by an adjective. This implies that the "AD VV" vs. "JJ NN" combination may be considered to be parallel by poets at the time.

The trends for Meng are similar, with a significantly higher score for couplets expected to be parallel than those that are not (82% vs. 53% at the character level). Compared to Wang, however, both percentages are lower. One wonders if this has any correlation with Meng being commonly considered a less accomplished poet. Since the 'rules' for parallelism have never been codified, Meng may also have simply espoused a more coarse-grained view of parts-of-speech. This hypothesis would be consistent with the fact that, at the phrase level, the proportion of parallelism for Meng is much closer to that for Wang. This suggests that Meng was content with parallelism at the phrase level and emphasized less on matching character to character.

Couplet type	Metric	Wang	Meng
Parallelism expected	Char-level only	91%	82%
	+ Phrase-level	95%	91%
Parallelism not expected	Char-level only	57%	53%
	+ Phrase-level	71%	71%
Parallelism optional	Char-level only	64%	65%
	+ Phrase-level	78%	81%

Table 5. The proportion of characters that are parallel to their counterparts in the couplet (see section 5.2). The couplets are classified into three types, depending on the genre of poetry and their position in the poem (see section 5.3).

6 Conclusion

We have presented the first large-scale dependency treebank of Classical Chinese literature, which encodes works by two poets in the Tang Dynasty. We have described how the dependency grammar framework has been derived from existing treebanks for Modern Chinese, and shown a high level of inter-annotator agreement. Finally, we have illustrated the utility of the treebank with a study on parallelism in Classical Chinese poetry.

Future work will focus on parsing Classical Chinese poems of other poets, and on enriching the corpus with semantic information, which would facilitate not only deeper study of parallelism but also other topics such as imagery and metaphorical coherence (Zhu and Cui, 2010).

Acknowledgments

We thank Tin Ho Chan and Cheuk Ho Wan for their assistance in annotating this treebank. This work was supported by a Strategic Research Grant (#7002549) from City University of Hong Kong.

[12] These are known as the 'regulated verse' (*lushi* 律詩) and are subject to definite patterns of parallelism. Those with fewer lines are left out, since their patterns are less regular.

References

Zong-Qi Cai. 2008. *How to Read Chinese Poetry*. Columbia University Press, New York.

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

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher Manning. 2009. Discriminative Reordering with Chinese Grammatical Relations Features. *Proc. 3rd Workshop on Syntax and Structure in Statistical Translation.Psychological Association*.

Gregory Crane, Brent Seales, and Melissa Terras. 2009. Cyberinfrastructure for Classical Philology. *Digital Humanities Quarterly* 3(1).

John F. Davis. 1969. *The Poetry of the Chinese*. Paragon Book, New York.

Ulrike Demske, Nicola Frank, Stefanie Laufer and Hendrik Stiemer, 2004. Syntactic Interpretation of an Early New High German Corpus. *Proc. Workshop on Treebanks and Linguistic Theories (TLT).*

Kais Dukes and Tim Buckwalter, 2010. A Dependency Treebank of the Quran using Traditional Arabic Grammar. *Proc. 7th International Conference on Informatics and Systems (INFOS)*, Cairo, Egypt.

Shengli Feng. 1998. Prosodic Structure and Compound Words in Classical Chinese. In *New Approaches to Chinese Word Formation*, Jerome Packard (ed.), Mouton de Gruyter.

W. Nelson Francis and Henry Kučera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin.

J. Hajic. 1998. Building a syntactically annotated corpus: The Prague Dependency Treebank. Issues of Valency and Meaning, Charles University Press.

Dag Haug and Marius Jøhndal. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. *Proc. Language Resources and Evaluation Conference (LREC).*

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

Li-min Huang. 2006. *The Study of Classical Poems of Tu-mu*. Master's Thesis, National Sun Yat-sen University, Taiwan.

Liang Huang, Yinan Peng, Huan Wang, and Zhengyu Wu. 2002. PCFG Parsing for Restricted Classical Chinese Texts. *Proc. 1st SIGHAN Workshop on Chinese Language Processing*.

Long Jiang and Ming Zhou. 2008. Generating Chinese Couplets using a Statistical MT Approach. *Proc. COLING*.

Lewis Lancaster. 2010. Pattern Recognition and Analysis in the Chinese Buddhist Canon: A Study of "Original Enlightenment". *Pacific World* 3(60).

Zuonan Li 李作南 and Renhou Li 李仁厚. 1986. A comparison of Classical Chinese and Modern Chinese 古今漢語語法比較 (in Chinese). Nei Menggu Renmin Chubanshe, China.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. *Stanford typed dependencies manual*. California: Stanford University.

Dingqiu Peng. 1960. Quan Tang Shi 全唐詩. Zhonghua Shuju, Beijing.

Susan Pintzuk and Plug Leendert. 2001. York-Helsinki Parsed Corpus of Old English Poetry. http://www-users.york.ac.uk/~lang18/pcorpus.html

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. *Proc. LREC*.

Edwin Pulleyblank. 1995. *Outline of Classical Chinese Grammar*. UBC Press, Vancouver, Canada.

Vitor Rocio, Mário Amado Alves, J. Gabriel Lopes, Maria Francisca Xavier, and Graça Vicente. 2000. Automated Creation of a Medieval Portuguese Partial Treebank. In Anne Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora* (Dordrecht: Kluwer Academic Publishers), pp. 211-227.

Ann Taylor, Anthony Warner, Susan Pintzuk and Frank Beths. 2003. *York-Toronto-Helsinki Parsed Corpus of Old English Prose*. University of York.

Pei-chuan Wei, P. M. Thompson, Cheng-hui Liu, Chu-Ren Huang, and Chaofen Sun. 1997. Historical Corpora for Synchronic and Diachronic Linguistics Studies. *Computational Linguistics and Chinese Language Processing* 2(1):131—145.

Li Wang 王力. 2004. A sketch of the history of Chinese language 漢語史稿 (in Chinese). Zhonghua Shuju, Beijing.

Jiaolu Xu. 許嘉璐. 1992. *Classical Chinese* 古代漢語 (in Chinese). Higher Education Press, Beijing.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer, 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering* 11:pp.207—238.

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

Chunshen Zhu and Ying Cui, 2010. Imagery Focalization and the Evocation of a Poetic World. *Chinese Translators Journal*.