A Classical Chinese Corpus with Nested Part-of-Speech Tags

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

Abstract

We introduce a corpus of classical Chinese poems that has been word segmented and tagged with parts-of-speech (POS). Due to the ill-defined concept of a 'word' in Chinese, previous Chinese corpora suffer from a lack of standardization in word segmentation, resulting in inconsistencies in POS tags, therefore hindering interoperability among corpora. We address this problem with nested POS tags, which accommodates different theories of wordhood and facilitates research objectives requiring annotations of the 'word' at different levels of granularity.

1 Introduction

There has been much effort in enriching text corpora with linguistic information, such as parts-of-speech (Francis and Kučera, 1982) and syntactic structures (Marcus et al., 1993). The past decade has seen the development of Chinese corpora, mostly for Modern Chinese (McEnery & Xiao, 2004; Xue et al., 2005), but also a few for pre-modern, or "classical", Chinese (Wei et al. 97; Huang et al. 2006; Hu & McLaughlin 2007).

One common design issue for any corpus of Chinese, whether modern or classical, is word segmentation. Yet, no segmentation standard has emerged in the computational linguistics research community. Hence, two adjacent characters X1X2 may be considered a single word in one corpus, but treated as two distinct words X_1 and X_2 in another[1]; furthermore, the part-of-speech (POS) tag assigned to X_1X_2 in the first corpus may differ from the tag for X_1 and the tag for X_2 in the second. These inconsistencies have made it difficult to compare, combine or exploit Chinese corpora. This paper addresses this problem by proposing a new method for word segmentation and POS tagging for Chinese and applying it on a corpus of classical Chinese poems.

2 Research Objective

A Chinese character may either function as a word by itself, or combine with its neighbor(s) to form a multi-character word. Since the goal of part-of-speech (POS) tagging is to assign one tag to each word, a prerequisite step is word segmentation, i.e., drawing word boundaries within a string of Chinese characters. The general test for 'wordhood' is whether "the meaning of the whole is compositional of its parts"; in other words, X_1X_2 forms one word when the meaning of the characters X_1X_2 does not equal to the meaning of X_1 plus the meaning of X_2 (Feng, 1998). Consider the string 沙門 *sha men* 'Buddhist monk'. As a transliteration from Sanskrit, it bears no semantic relation with its constituent characters 沙 *sha* 'sand' and 門 *men* 'door'. The two characters therefore form one word.

From the point of view of corpus development, word segmentation has two consequences. First, it defines the smallest unit for POS analysis. It would be meaningless to analyze the POS of the individual characters as,

[1] This phenomenon can be compared with what is often known as multiword expressions (Sag et al., 2002) in other languages.

Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 75–84,
Avignon, France, 24 April 2012. ©2012 Association for Computational Linguistics

say, 沙/NN and 門/NN (see Table 1 for the list of POS tags used in this paper). Instead, the two characters *sha* and *men* together should be assigned one POS tag, 沙門/NN.

Second, word segmentation sets boundaries for automatic word retrieval. A simple string search for "*sha men*" on a non-segmented corpus might yield spurious matches, where *sha* is the last character of the preceding word, and *men* is the first character of the following one. In a word study on 本覺 *ben jue* 'original enlightenment' (Lancaster, 2010), based on a non-segmented corpus of the Chinese Buddhist Canon, the author needed to manually examine each of the 763 occurrences of the string *ben jue* in order to determine which of them are in fact the word in question, rather than accidental collocations of the two characters. Word boundaries resulting from word segmentation would have removed these ambiguities, expedited the search and enabled this kind of word studies to be performed on much larger scales.

There is not yet a scholarly consensus on a precise definition of 'wordhood' in Classical Chinese (Feng, 1998). Inevitably, then, treatment of word segmentation varies widely from corpus to corpus. Some did not perform word segmentation (Huang et al. 2006); others adopted their own principles (Wei et al. 1997; Hu & McLaughlin 2007). The lack of standardization not only hinders corpus interoperability, but also makes it difficult for any single corpus to cater to users with different assumptions about wordhood or different research objectives. What is regarded as one word for a user may be two words in the eyes of another. Consider two alternative analyses of the string 黃河 *huang he* 'Yellow River' in two research tasks. For retrieval of geographical references in a text, it should ideally be tagged as one single proper noun, 黃河/NR; to study parallelisms in poetry, however, it is better to be tagged as two separate words, 黃/JJ *huang* 'yellow' followed by 河/NN *he* 'river', in order not to obscure the crucial POS sequence 'adjective-noun' that signals parallelism in a couplet. To settle on any particular word segmentation criterion, then, is to risk omitting useful information.

We are not qualified to lay down any definitive criterion for word segmentation; rather, we advocate a theory-neutral approach through nested POS tags: characters are analyzed individually whenever possible, but annotated with hierarchical tags to recognize possible word boundaries.

3 Previous Work

In this section, we summarize previous practices in Chinese word segmentation (section 3.1) and part-of-speech tagging (section 3.2), then describe existing frameworks of multi-level tagging (section 3.3).

3.1 Word segmentation

As mentioned in Section 2, a common test for word segmentation is "compositionality of meaning". While there are clear-cut cases like *sha men*, many cases fall in the grey area. Indeed, even native speakers can agree on word boundaries in modern Chinese only about 76% of the time (Sproat et al., 1996). It is not surprising, then, that a myriad of guidelines for word segmentation have been proposed for various corpora of Modern Chinese (Liu et al., 1994; Chinese Knowledge Information Processing Group, 1996; Yu et al., 1998; Xia 2000; Sproat and Emerson, 2003). In the rest of this section, we first review the approaches taken in three classical Chinese corpora, developed respectively at Jiaotong University (Huang et al., 2006), University of Sheffield (Hu et al., 2005) and the Academia Sinica (Wei et al., 1997). We then describe in more detail a modern Chinese corpus, the Penn Chinese Treebank (Xue et al., 2005).

Corpus at Jiaotong University. This treebank consists of 1000 sentences of pre-Tsin classical Chinese. No word segmentation was performed. On the one hand, this decision may be supported by the fact that "in general the syllable, written with a single character, and the word correspond in Classical Chinese" (Pulleyblank, 1995). On the other hand, there are nonetheless a non-negligible number of strings for which it makes little sense to analyze their constituent characters. These include not only transliterations of foreign loanwords such as *sha men*, but also bound morphemes[2] and reduplications[3] (Pulleyblank, 1995). The lack of segmentation in this corpus also leads to the lack of word boundaries to support word retrieval.

[2] E.g., 然 ran, a suffix forming expressive adverbs such as 卒然 *cu ran* 'abruptly'

[3] E.g., 須 *xu* 'wait', which, via partial reduplication, derives 須臾 *xu yu* 'a moment'

Academia Sinica Ancient Chinese Corpus. With more than 500K characters, this is the largest word-segmented and POS-tagged corpus of classical Chinese. In the annotation process, a character is presumed to be a word in its own right; it is combined with other characters to form a word if they fall into one of the following categories: parallel and subordinating compounds; bisyllabic words; reduplications; and proper nouns. Two of these categories, namely, bisyllabic words and reduplications, are retained in our word segmentation criteria (see section 4.1). Proper nouns, as well as parallel and subordinating compounds, however, are treated differently (see section 4.2).

Sheffield Corpus of Chinese. This corpus has more than 109K characters of archaic Chinese and 147K characters of medieval Chinese. Word segmentation was performed by hand. Their criteria for word segmentation, unfortunately, do not seem to be publicly available.

The Penn Chinese Treebank. This widely used treebank of modern Chinese boasts an extensively documented word segmentation procedure (Xia, 2000), which rests on six principles. We follow their principle that complex internal structures should be segmented when possible (see section 4.2). We also retain a second principle that a bound morpheme forms a word with its neighbor[4], although morphemes in Classical Chinese are nearly always free forms (Feng, 1998).

A third criterion is the number of syllables. Consider a noun phrase N_1N_2 where the first noun (N_1) modifies the second (N_2). This noun phrase is considered one word if N_2 consists of one character, but two words if N_2 has two or more characters. For example, the string 北京大學 *bei jing da xue* ‘Peking University’ is segmented as two words *bei jing* ‘Peking’ and *da xue* ‘university’, since ‘university’ is made up of two characters; however, a similar string 北京市 *bei jing shi* ‘Beijing City’ is one word, since ‘city’ consists of just one character *shi*. Given the dominance of monosyllabic words in classical Chinese, a direct application of this principle would have resulted in a large number of multi-character words in our corpus.

Further, there are three linguistic tests. The “semantic compositionality” test has already been outlined in section 2 and is not repeated here. The “insertion test” asks whether another morpheme can be inserted between two characters X_1 and X_2; if so, then X_1X_2 is unlikely to be a word. The “XP-substitution test” asks if a morpheme can be replaced by a phrase of the same type; if not, then it is likely to be part of a word. Performing these tests requires intuition and familiarity with the language. Since no human is a native speaker of classical Chinese, we found it difficult to objectively and reliably apply these tests. Instead, we strive to accommodate different views of wordhood in our corpus.

3.2 Part-of-Speech Tagging

Following word segmentation, each word is assigned a part-of-speech (POS) tag. Most POS tagsets cover the major word categories, such as nouns, verbs, and adjectives; they differ in the more fine-grained distinctions within these categories. For examples, verbs may be further subdivided into transitive and intransitive; nouns may be further distinguished as common, proper or temporal; and so on. In general, a larger tagset provides more precise information, but may result in lower inter-annotator agreement, and hence reduced reliability.

Classical Chinese does not have inflectional morphology; this makes POS tags even more informative, but also makes inter-annotator agreement more challenging. As with other languages, the POS tagset is tailored to fit one’s research objective, as reflected in the wide-ranging levels of granularity in different corpora, from 21 tags in (Huang et al., 2006), 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). Our tagset is based on that of the Penn Chinese Treebank, which lies towards the lower end of this spectrum, with 33 tags.

3.3 Multi-level Tagging

In principle, any text span may be annotated at an arbitrary number of levels using, for example, stand-off annotation. In practice, most effort has concentrated on identifying named entities, such as (Doddington et al., 2004). While our corpus does specify word boundaries of multi-character proper nouns, it tackles all other forms of compounds in general (section 4.2).

Turning to the Chinese language in particular, we are by no means the first to point out inconsistencies in word segmentation and POS

[4] E.g., the morpheme 本 *ben* is bound to the character 人 *ren* ‘person’ in the word 本人 *ben ren* ‘oneself’

tags among different corpora. Annotators of the Penn Chinese Treebank, among others, also recognized this issue (Xia, 2000). As a remedy, a two-level annotation method is used on a number of grammatical constructions. Suppose it is uncertain whether X_1 and X_2 should be considered two separate words or one word. Under this method, X_1 and X_2 are first tagged individually (say, as pos_1 and pos_2), then tagged as a whole (say, as *pos*), and finally grouped together with a pair of brackets, resulting in the final form $(X_1/pos_1\ X_2/pos_2)/pos$. For instance, rather than simply tagging the string 走上來 *zou shang lai* ‘walk up’ as one verb 走上來/VV, the three-character word is further segmented internally as 走 *zou* ‘walk’ and 上來 *shang lai* ‘come up’, hence (走/VV 上來/VV)/VV. This method makes the interpretation more flexible: those who consider *zou shang lai* to be one word can simply ignore the details inside the brackets; others who view *zou* and *shang lai* as stand-alones can discard the brackets and retain their individual analyses.

This device is used in the Penn Chinese Treebank on only a narrow range of constructions to ensure compatibility with the Chinese Knowledge Information Processing Group (1996) and with (Liu et al., 1994). In contrast, it is generalized in our corpus as nested tags of arbitrary depth, and used systematically and extensively to mark alternate word boundaries.

Tag	Part-of-Speech
AD	Adverb
CD	Cardinal number
DER	Resultative *de5*
DEV	Manner *de5*
FW	Foreign word
IJ	Interjection
JJ	Other noun modifier
LC	Localizer
NN	Other noun
NR	Proper noun
NT	Temporal noun
P	Preposition
PN	Pronoun
SP	Sentence-final particle
VV	Other verb

Table 1: Part-of-speech tags of the Penn Chinese Treebank that are referenced in this paper. Please see (Xia, 2000) for the full list.

4 Corpus Design

This section describes our corpus at two levels, first the ‘strings without internal structures’ (section 4.1), which may be combined to form ‘strings with internal structures’ (section 4.2) and marked with nested brackets and tags.

4.1 Strings without internal structures

The lowest annotation layer marks the boundaries of what will be referred to as ‘strings without internal structures’. These are roughly equivalent to ‘words’ in existing Chinese corpora.

Segmentation criteria. Following the practice of the Academia Sinica Ancient Chinese Corpus, each character is initially presumed to be a monosyllabic word. The annotator may then decide that it forms a multi-character word with its neighbor(s) under one of the categories listed in Table 2. This set of categories represents a more stringent segmentation criterion than those in most existing corpora, such that the number of multi-character words is relatively small in our target text (see section 6).

Category	**Example**
Foreign loanwords	匈奴 *xiong nu* ‘the Xiongnu people’ e.g., 匈奴/NR 圍酒泉 ‘The Xiongnus surrounded the city of Jiuquan’
Numbers	十五 *shi wu* ‘fifteen’, 十六 *shi liu* ‘sixteen’ e.g., 少年十五/CD 十六/CD 時 ‘as a youth of 15 or 16 years of age’
Reduplications	駸駸 *qin qin* ‘quickly’ e.g., 車馬去駸駸/AD ‘the chariots went quickly’
Bound morphemes	油然 *you ran* ‘spontaneously’ e.g., 天油然/AD 作雲 ‘the sky spontaneously makes clouds’

Table 2: Categories of multi-character words that are considered ‘strings without internal structures’ (see Section 4.1). Each category is illustrated with one example from our corpus.

Part-of-speech tagging. Similar to the principle adopted by the Penn Chinese Treebank, POS tags are assigned not according to the meaning of the word, but to syntactic distribution (Xia, 2000), i.e. the role the word plays in the sentence. Compared to modern Chinese, it is a much more frequent phenomenon in the classical language for a word to function as different parts-of-speech in different contexts. For example, it is not uncommon for nouns to be used as verbs or adverbs, and verbs as adverbs (Pulleyblank, 1995). Consider two nouns 鐘 *zhong* 'bell' and 雲 *yun* 'cloud'. The former is used as a verb 'to ring' in the verse 深山何處鐘/VV 'where in the deep mountain [is it] ringing'; the latter serves as an adverb 'in the manner of clouds' in the verse 倏忽雲/AD 散 'quickly disperses like clouds'. They are therefore tagged as a verb (VV) and an adverb (AD). Likewise, when the verb 盡 *jin* 'exhaust' has an adverbial sense 'completely', such as in 送君盡/AD 惆悵 'saying farewell to you, I am utterly sad', it is tagged as such.

We largely adopted the tagset of the Penn Chinese Treebank. As the standard most familiar to the computational linguistics community, their tagset has been used in annotating a large volume of modern Chinese texts, offering us the possibility of leveraging existing modern Chinese annotations as training data as we seek automatic methods to expand our corpus. For the most part, the Penn tagset can be adopted for classical Chinese in a straightforward manner. For example, the tag PN (pronoun) is used, instead of the modern Chinese pronouns 我 *wo* 'I' and 你 *ni* 'you', for the classical equivalents 吾 *wu* 'I' and 爾 *er* 'you'. Similarly, the tag SP (sentence-final particles) is applied, rather than to the modern Chinese particles 吧 *ba* or 呀 *a*, to their classical counterparts 耳 *er* and 也 *ye*. In other cases, we have identified roughly equivalent word classes in classical Chinese. To illustrate, although the classical language has no prepositions in the modern sense, the P (preposition) tag is retained for words known as coverbs (Pulleyblank, 1995). A few tags specific to modern Chinese are discarded; these include DER, DEV, and FW (see Table 1).

4.2 Strings with internal structures

Since our criteria for 'strings without internal structures' are intentionally strict, they disqualify many multi-character strings that may fail the "semantic compositionality" test and are therefore commonly deemed words. These include proper names with analyzable structures, as well as parallel or subordinating compounds, which are considered 'strings with internal structures' in our corpus, and are annotated with nested tags.

Category	**Example**
Parallel compounds	
Similar meaning	君王 *jun wang* 'king' = 君 *jun* 'ruler' + 王 *wang* 'king' (君/NN 王/NN)/NN
Related meaning	骨肉 *gu rou* 'kin' = 骨 *gu* 'bone' + 肉 *rou* 'flesh' (骨/NN 肉/NN)/NN
Opposite meaning	是非 *shi fei* 'rumors' = 是 *shi* 'right' + 非 *fei* 'wrong' (是/JJ 非/JJ)/NN
Subordinating compounds	
Verb-object	識事 *shi shi* 'experience' = 識 *shi* 'understand' + 事 *shi* 'affairs' (識/VV 事/NN)/NN
Subject-verb	日落 ri luo 'sunset' = 日 *ri* 'sun' + 落 *luo* 'descend' (日/NN 落/VV)/NN
Adjectival modifier	少年 *shao nian* 'youth' = 少 *shao* 'few' + 年 *nian* 'year (少/JJ 年/NN)/NN
Noun modifier	家食 *jia shi* 'household food' = 家 *jia* 'house' + 食 *shi* 'food' (家/NN 食/NN)/NN

Table 3: Categories of multi-character words that are considered 'strings with internal structures' (see Section 4.2). Each category is illustrated with an example from our corpus. Both the individual characters and the compound they form receive a POS tag.

Segmentation criteria. All parallel and subordinating compounds are considered to be 'strings with internal structures'. A parallel compound is a two-character noun, verb and adjective "in which neither member dominates the other" (Packard, 1998) and it refers to one meaning despite having two characters. For example, the noun compound 骨肉 *gu rou*, formed from from 骨 *gu* 'bone' and 肉 *rou* 'flesh', means simply 'kin' rather than 'bone and flesh'. In practice, in our corpus, two characters are considered to be a parallel compound when they are of the same POS, and have similar, related, or opposite meaning, as shown in Table 3. The individual characters are 'strings without internal structure' and receive their own POS tags, while the compound also receives its own tag.

Subordinating compounds refer to those where "one member (the modifier) is subordinate to and modifies the other (the head)" (Packard, 1998). For example, the compound 少年 *shao nian* is made up of an adjective 少 *shao* 'few' modifying a noun 年 *nian* 'year', but together has the specialized meaning 'youth'. In our corpus, two characters are considered to form a subordinating compound when they have the verb-object or subject-verb relationship, or a modifier-head relationship, including adjectival modifiers and noun modifiers.

Proper names can also have internal structures, whenever the grammatical structure of their constituent characters may be discerned. The most common such proper names in our corpus are geographical names, such as 黄河 *huang he* 'Yellow River', where the adjective *huang* 'yellow' modifies the noun *he* 'river'. Another frequent type is personal names with titles, such as 始興公 *shi xing gong* 'Duke Shixing', where one noun modifies another.

Our definition of 'strings with internal structures' is deliberately broad. As a result, some of these strings would not be considered to be a word or compound by all or even most linguists. Many verb-object combinations, for example, may well fail the 'semantic compositionality' test. This is intentional: rather than searching for the perfect segmentation policy that suits everyone [5], the nested annotations allow the user to decide which level of tags is suitable for the research objective at hand.

Part-of-speech tagging. The nested annotations of 'strings with internal structures' not only mark the possible word boundaries, but also assign a POS tag at every level, since that tag is not always predictable from the tags of the constituent characters. Consider the verse in Table 4. There are two possible segmentations for the string 晚來 *wan lai*. As two separate words, *wan* 'evening' and *lai* 'come' form a clause meaning 'as the evening comes'; the whole verse may be translated 'the weather turns chilly as the evening comes'. Alternatively, they can be taken as a two-character word, i.e., simply a temporal noun 晚來/NT *wan lai* 'evening'. In this case, the proper translation would be 'the weather turns chilly at evening'. Notice that the tag NT (temporal noun) cannot be predicted from the tags at the lower level, NN (noun) and VV (verb).

Further, these nested tags indicate alternatives for future syntactic analysis. In dependency grammar, for instance, the adjectival verb *qiu* 'chilly' would be the head of the verb *lai*, which is the verb in the subordinate clause; in the second interpretation, however, it would be the head of a temporal modifier, *wan lai* 'evening'.

天	氣	晚	來	秋
tian	*qi*	*wan*	*lai*	*qiu*
'weather'		'night'	'come'	'chilly'
NN		NN	VV	JJ
		NT		

Table 4: POS annotations of an example sentence with a string, *wan lai* 'evening', that has internal structure. See Section 4.2 for two possible translations, and Table 1 for the meaning of the POS tags.

Verse 1				
獨	樹	臨	關	門
du	*shu*	*lin*	*guan*	*men*
'only'	'tree'	'upon'	'pass'	'entrance'
JJ	NN	VV	NN	NN
'a lone tree watches the entrance of the pass'				
Verse 2				
黃	河	向	天	外
huang	*he*	*Xiang*	*tian*	*wai*
'yellow'	'river'	'face'	'sky'	'outside'
JJ	NN	VV	NN	LC
NR				
'The Yellow River faces the outer sky'				

Table 5: POS annotations of a couplet, i.e., a pair of two verses, in a classical Chinese poem. See Table 1 for the meaning of the POS tags.

One significant benefit of nested annotation, especially in classical Chinese poetry, is the preservation of the underlying parallelism. Two consecutive verses, called a *couplet*, always have the same number of characters. Moreover, two characters at the same position in the two verses

[5] The verb-object combination, for example, is "among the hardest cases for the word definition" (Xia, 2000).

often have the same or related POS. Consider the couplet in Table 5. The first two characters of each verse, 獨樹 *du shu* 'lone tree' and 黃河 *huang he* 'Yellow River', respectively, are parallel; both are noun phrases formed by a noun modified by the preceding adjective.

In most existing corpora, *huang he* would be simply considered one word and assigned one tag, namely, a proper noun 黃河/NR. This treatment would, first of all, result in one verse having four words and the other five, making it difficult to analyze character correspondences. It also obscures the parallelism between the noun phrases *du shu* and *huang he*: both are JJ-NN, i.e. 'adjective-noun'. In contrast, our corpus annotates *huang he* as a string with internal structures (黃/JJ 河/NN)/NR, as shown in Table 5. Its outer tag (NR) preserves the meaning and boundary of the whole proper noun *huang he*, facilitating word searches; the inner tags support automatic identification of parallel structures.

In all examples above of 'strings with internal structures', the nested annotations have only a depth of one. In theory, the depth can be arbitrary, although in practice, it rarely exceeds two. An example is the string 細柳營 *xi liu ying* 'Little Willow military camp'. At the coarsest level, the three characters may be considered to form one proper noun, referring to a camp at the ancient Chinese capital. The string obviously has 'internal structures', composed of 營 *ying* 'military camp' and its location, the place name 細柳 *xi liu* 'Xiliu'. Furthermore, this place name has an evocative meaning, 'little willow', made up of the adjective *xi* 'little' and the noun *liu* 'willow'. As shown in Table 6, this analysis results in a three-level, nested annotation ((細/JJ 柳/NN)/NR 營/NN)/NR.

Furthermore, these three characters are the last characters in the second verse of a couplet. Table 6 also shows the annotations for the corresponding characters in the first verse, 新豐市 *xin feng shi* 'Xinfeng city'. Taken together, the annotations reveal the perfect symmetry of both noun phrases at every level of analysis.

5 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. 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.

Our method of word segmentation and POS tagging has been applied to the complete works by two Chinese poets in the 8th century CE, Wang Wei and Meng Haoran. Wang is considered one of the three most prominent Tang poets; Meng is often associated with Wang due to the similarity of his poems in style and content. Altogether, our corpus consists of about 32,000 characters in 521 poems.

Noun Phrase in Verse 2		
細	柳	營
xi	*liu*	*ying*
'little'	'willow'	'camp'
'Little Willow camp'		
JJ	NN	NN
NR		
NR		
Noun Phrase in Verse 1		
新	豐	市
xin	*feng*	*shi*
'new'	'abundance'	'city'
'City of New Abundance'		
JJ	NN	NN
NR		
NR		

Table 6: Part-of-speech annotations of the three-character strings 細柳營 *xi liu ying* 'Little Willow military camp' and 新豐市 *xin feng shi* 'Xinfeng city'. Both are 'strings with internal structures', with nested structures that perfectly match at all three levels. They are the noun phrases that end both verses in the couplet 忽過新豐市, 還歸細柳營.

6 Evaluation

Two research assistants, both of whom hold a Bachelor's degree in Chinese, have completed the annotations. To estimate inter-annotator agreement, the two annotators independently performed word segmentation and POS tagging on a 1,057-character portion of the poems of Wang. We measured their agreement on word segmentation, POS tags for 'strings without internal structures', and those for 'strings with internal structures'.

Word segmentation. This task refers to decisions on boundaries between 'strings without internal structure' (section 4.1). Given the rather stringent criteria, it is not surprising that only about 6.5% of the words in our texts contain more than one character. Among these, 75% consists of two characters.

Disagreement rate on the presence of word boundary between characters was only 1.7%. No comparable figure has been reported for classical Chinese word segmentation, but this rate compares favorably with past attempts for modern Chinese, e.g., an average of 76% inter-human agreement rate in (Sproat et al., 1996). This may be explained by the relatively small number of types of strings (see Table 2) that are considered to be multi-character words in our corpus.

POS tagging on strings without internal structures. We now consider the POS tags assigned at the lowest level, i.e. those assigned to strings without internal structures. After discarding characters with disputed word segmentation boundaries, the disagreement rate on POS tags was 4.9%. Three main areas of disagreement emerged.

One category is the confusion between verbs and adverbs, when the annotators do not agree on whether a verb has an adverbial force and should therefore be tagged as AD rather than VV. For example, the word 紆 *yu* 'bow' normally functions as a verb, but can also be used adverbially when referring to an attitude, 'respectfully', which is implied by bowing. When used in collocation with the word 顧 *gu* 'visit' in the verse 伏檻紆三顧 *fu jian yu san gu*, it can therefore mean 'prostrated on the threshold and respectfully (AD) paid visits three times' or 'prostrated on the threshold and bowed (VV) and paid visits three time'.

A second category is the confusion between measure word and a noun. The noun 簞 *dan* 'bowl' can collocate with the noun 食 *shi* 'food'. Taken together, *dan shi* can either mean 'a bowl of food' where *dan* is a measure word (M), or it can simply mean a specific kind of meal, in which case *dan* is a noun modifier (NN). Both interpretations have been supported by commentators.

The third is the confusion between adjective (JJ) and noun (NN), when the word in question modifies a noun that immediately follows. For example, for the noun phrase 命服 *ming fu* 'uniform with rank devices', it is clear that the first character 命 *ming* 'profession' modifies the second character 服 *fu* 'clothes'. The annotators did not agree, however, on whether *ming* is a noun modifier or an adjectival modifier. In the Penn Chinese Treebank POS guidelines (Xia, 2000), this question is resolved with the linguistic test: if the word is JJ, then it cannot be the head of a noun phrase. In practice, this test is difficult to apply for non-native speakers of a language. The annotator would have to decide whether he can compose a "good" classical Chinese that uses the word has an NP head.

POS tagging on strings with internal structures. Thirdly, we turn our attention to POS tags assigned at the higher levels of the nested structure. Of the 'strings with internal structures', about 73% consist of two characters; those longer than two characters are mostly proper names.

We measured inter-human agreement for the nested bracketing by taking each annotator in turn as 'gold', and calculated the precision and recall of the other. The average precision was 83.5%; the average recall also worked out to 83.5%. A significant source of error was disagreement over whether several characters form a proper name, and should therefore be bracketed and assigned the tag NR; these often involve knowledge of Chinese history and geography. In the remaining cases of discrepancies, the vast majority are direct consequences of differences in POS tagging. Lastly, among the strings with internal structures that have received identical bracketing, there was almost complete agreement between the annotators regarding their POS tags, except in a few isolated cases.

7 Conclusion

We have a described a novel method of word segmentation and POS tagging, tailored for the classical Chinese language, and designed to support interoperability between corpora. This method has been applied on about 32,000 characters, drawn from two well-known poets from the 8th century CE.

The corpus aspires to contribute to two areas of scholarly enquiry. First, it is expected to facilitate classical Chinese word studies by automating word retrieval (e.g., (Lancaster, 2010)), and will support investigations in other areas of classical Chinese philology, such as semantic and metaphorical coherence (Zhu & Cui, 2010), by supplying syntactic evidence. Second, it is intended to serve as training data for automatic POS taggers, to automate the analysis of the vast and growing digital collections of classical Chinese texts.

Acknowledgments

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

References

Pi-Chuan Chang, Michel Galley, and Chris Manning, 2008. Optimizing Chinese Word Segmentation for Machine Translation Performance. In *Proc. ACL 3rd Workshop on Statistical Machine Translation*.

Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu, 1996. SINICA CORPUS: Design Methodology for Balanced Corpora. In *Proc. Language, Information and Computation (PACLIC).*

Chinese Knowledge Information Processing Group, 1996. Shouwen Jiezi --- A study of Chinese Word Boundaries and Segmentation Standard for Information Processing (in Chinese). Technical Report, Academia Sinica, Taipei.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel, 2004. The Automatic Content Extraction (ACE) Program: Tasks, Data, and Evaluation. In *Proc. LREC.*

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.

Xiaolong Hu, N. Williamson and J. McLaughlin, 2005. Sheffield Corpus of Chinese for Diachronic Linguistic Study. In *Literary and Linguistic Computing* **20**(3):281---93.

Xiaoling Hu and Jamie McLaughlin, 2007. The Sheffield Corpus of Chinese. Technical Report, University of Sheffield, UK.

Liang Huang, Yinan Peng, Huan Wang and Zhengyu Wu, 2006. Statistical Part-of-Speech Tagging for Classical Chinese. In *Lecture Notes in Computer Science* **2448**:296-311.

Lewis Lancaster, 2010. Pattern Recognition and Analysis in the Chinese Buddhist Canon: A Study of “Original Enlightenment”. In *Asia Pacific World* 3rd series **60**.

Yuan Liu, Qiang Tan, and Kun Xu Shen, 1994. *Segmentation Standard for Modern Chinese Information Processing and Automatic Segmentation Methodology*. Qinghua University Press, Beijing, China.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini, 1993. Building a Large Annotated Corpus of English: the Penn Treebank. In *Computational Linguistics* **19**(2).

Anthony McEnery and Zhonghua Xiao, 2004. The Lancaster Corpus of Mandarin Chinese: a corpus for monolingual and contrastive language study. In *Proc. LREC.*

Jerome Lee Packard, 1998. New Approaches to Chinese Word Formation: Morphology, Phonology and the Lexicon in Modern and Ancient Chinese. In *Trends in Linguistics Studies and and Monographs*, Mouton de Gruyter.

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

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

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger, 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Lecture Notes in Computer Science* 2276/2002:189—206.

Richard Sproat, Chilin Shih, William Gale and Nancy Chang, 1996. A Stochastic Finite-state Word-Segmentation Algorithm for Chinese. In *Computational Linguistics* **22**(3).

Richard Sproat and Thomas Emerson, 2003. The First International Chinese Word Segmentation Bakeoff. In *Proc. 2nd SIGHAN Workshop on Chinese Language Processing*.

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

Fei Xia, 2000. *The Segmentation Guidelines for the Penn Chinese Treebank (3.0)*. University of Pennsylvania, PA.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer, 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. In *Natural Language Engineering* **11**:207-238.

Shiwen Yu, Xuefeng Zhu, Hui Wang, and Yunyun Zhang, 1998. *The Grammatical Knowledgebase of Contemporary Chinese: A Complete Specification* (in Chinese). Tsinghua University Press, Beijing, China.

Shiwen Yu, Huiming Duan, Xuefeng Zhu, and Bin Sun, 2002. 北京大學現代漢語語料庫基本加工规範 *Beijing daxue xiandai hanyu yuliaoku jiben* *jiagong guifan*. 中文信息學報 *Zhongwen Xinxi Xuebao* **5**:49--64.

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