A dependency treebank of Chinese Buddhist texts

John Lee and Yin Hei Kong
The Halliday Centre for Intelligent Applications of Language Studies, Department of Linguistics and Translation, City University of Hong Kong, Hong Kong

Abstract

We present a dependency treebank of Buddhist Chinese texts, containing more than 50K characters drawn from four sutras in the Chinese Buddhist Canon. With dates of composition that span almost five centuries, these sutras bear witness to the evolution of the Chinese language. The treebank has been annotated using the part-of-speech tagset of the Penn Chinese Treebank, and the Stanford Dependencies for Chinese with slight modifications. The article first discusses the texts and the annotation framework of this treebank, and reports on inter-annotator agreement. It then describes the search platform, to which the treebank has been imported, and applies the treebank to an open question in Chinese historical linguistics—the emergence of the Chinese copula.

Correspondence:
The Halliday Centre for Intelligent Applications of Language Studies, Department of Linguistics and Translation, City University of Hong Kong, Kowloon, Hong Kong.
E-mail:
jsylee@cityu.edu.hk

1 Introduction

The Chinese Buddhist Canon, a corpus of over 52 million characters, consists of translations of Buddhist texts from Indic languages into medieval Chinese from the 2nd to the 11th centuries CE. As scriptures for Chinese Buddhists, these texts have been studied, recited, and chanted by monks and laymen since their compilation.

While the cultural impact and the religious significance of these texts have always been evident, their linguistic significance has only been recognized more recently. With 1,514 individual texts composed over a millennium, the Canon is a linguistic data set that reflects the evolution of the Chinese language, including the development of vernacular words and usages into literary forms (Karashima 1996), the process of disyllabication and changes in lexical meanings, as well as syntactic structures (Zhu 2010; Jiang and Hu 2013). In fact, the composition of the Canon itself shaped many of these processes, comparable in some ways to the effect of Luther's translation of the Bible on the German language.

The sheer volume of this corpus means that no scholar can read and absorb all of its content, or analyze any linguistic phenomenon over the entire corpus with manual methods. Digitized versions of the Canon have now enabled computation of n-gram counts and distributions (Lancaster 2010), but similar analyses on part-of-speech (POS) and sentence structures are impossible without syntactic annotations.

As a first step toward this goal, we have manually created a treebank for four sutras from the Canon, totaling more than 50K characters. This is the first treebank for Medieval Chinese, containing word segmentation, POS tags, and syntactic analyses in the dependency framework.

Our treebank is designed to facilitate both fundamental and applied linguistics research, and provide intellectual access to the Canon. First, it reduces the time and effort needed to collect quantitative evidence for topics in Chinese historical

Digital Scholarship in the Humanities, Vol. 31, No. 1, 2016.
doi:10.1093/llc/fqu048 Advance Access published on 6 September 2014

linguistics, and allows researchers to replicate and verify results. The study in Section 6 on the emergence of the Chinese copula serves as an illustration. Second, the linguistic annotations can potentially improve the performance of downstream natural language processing applications, such as authorship attribution (Hung *et al.*, 2010) and detection of social networks (Bingenheimer *et al.*, 2011) in Buddhist literature. Third, this resource serves as reading aid for a major work in Chinese literature. To date, only punctuations have been added to the original medieval Chinese text to assist modern readers. To support reading of Christian and Islam scriptures in their original Greek and Arabic, dependency treebanks for the New Testament (Haug and Jøhndal 2008) and the Qur'an (Dukes and Buckwalter 2010) have been compiled; our treebank should likewise help both students and laymen understand the texts better and faster. To this end, we have also incorporated the English definitions from the Soothill-Hodous Dictionary of Chinese Buddhist Terms into the treebank.

The rest of the article is organized as follows. In the next section, we discuss existing corpora of historical texts, with a focus on Chinese, and the differences between Classical Chinese and the language in our corpus. In Section 3, we state the sources of our text. In Section 4, we outline the annotation framework of our treebank, then evaluate it in terms of inter-annotator agreement in the following section. Finally, in Section 6, we describe the search platform for accessing the treebank, and apply it to an open research question in Chinese historical linguistics—the emergence of the Chinese copula.

2 Relevant Research

We first provide an overview of digital corpora of historical texts (Section 2.1), then describe differences between Classical Chinese and the language found in Buddhist Chinese texts (Section 2.2).

2.1 Corpora of historical texts

The field of historical linguistics has increasingly drawn on digital corpora. These corpora boast a growing number of languages, from Old English (Taylor *et al.*, 2003; Kroch *et al.*, 2004) to Early New High German (Demske *et al.*, 2004) and Medieval Portuguese (Rocio *et al.*, 2000). As linguistic annotations on these texts become available, there is growing interest in the use of data-driven methods to analyze the evolution of languages. Currently, the largest resource is the Google Books Ngram Corpus, which contain syntactically analyzed texts in multiple languages, with composition dates that cover more than five centuries (Lin *et al.*, 2012). Its annotations consist of automatically produced POS tags and unlabeled dependencies. It can be used, for example, to detect changes in spelling (Lin *et al.*, 2012), in word usage, and word sense over time (Mihalcea and Nastase 2012). With labeled dependencies, the Perseus Latin Dependency treebank has been leveraged to study selectional preferences (Bamman and Crane 2008), with the ultimate goal of automatically generating dynamic lexica.

For Classical Chinese, the two major diachronic corpora, the Academia Sinica Ancient Chinese Corpus (Wei *et al.*, 1997) and the Sheffield Corpus of Chinese (Hu *et al.*, 2005), cover a wide range of time and genres. Both corpora have been annotated with word boundaries and POS tags, but not with syntactic structures. To the best of our knowledge, only two treebanks have been constructed to date for Classical Chinese. One is a constituent-based treebank, with 1000 sentences from the pre-Qin period, i.e. the third century BCE and before (Huang *et al.*, 2002). The other is a dependency treebank with 32K characters of Tang poetry, consisting of works of three prominent poets from the eighth century CE (Lee and Kong 2012). The present treebank is the first for Medieval Chinese. It is not only larger in size, but also more varied—the authors spanned almost five centuries, and the texts contain both poetry and prose, dialogues, narratives, and doctrines.

2.2 Comparison with Classical Chinese

With these treebanks of Classical Chinese as our starting point, we must take into account differences in medieval Chinese, and more specifically the language found in Buddhist Chinese texts.

According to Zhu (2010), it has two basic differences, reflecting its blending with two linguistic phenomena. The first difference is the result of blending with Indic languages, the source languages of the Chinese Buddhist Canon. This can be seen in the large number of transcribed or transliterated words from the Indian original; and in many cases, word-to-word translation from the source sentence yielded unusual word order for Chinese. The language is also influenced by the grammar in the Indian original, in terms of the more frequent use of passive sentences, and the vocative case.

The second difference is a consequence of blending with the oral and dialectal elements in contemporary, secular Chinese. This can be shown in various lexical and grammatical aspects, of which foremost is perhaps the greater number of polysyllabic words: while most words in Classical Chinese consist of single characters, medieval Chinese developed many two-character compound nouns, verbs, and adjectives. Other aspects, to name a few, include the use of novel interrogative articles, e.g. *wei* 爲; the use of demonstrative pronouns (Zhu 2008); the range of passive markers, e.g. diminished frequency of the pattern '*jian* 見 V', increased frequency of '*wei* 爲 R *suo* 所 V', and novel passive markers, e.g. *suojian* 所見; the emergence of the use of *zhuo* 著 as aspect marker; as well as the increased popularity of the copula (Section 6), which can sometimes be located after the predicate. More detail on these new phenomena can be found in Zhu (2009) and Jiang and Hu (2013).

3. Text Editions and Sources

We first describe the editions of the Chinese Buddhist Canon (Section 3.1), then motivate our selection of sutras (Section 3.2).

3.1 Editions

There are both block print and manuscript editions of the Chinese Buddhist Canon. The earliest of the block prints was the Northern Song Kaibao Edition. A set of rubbings from these blocks was given to Korea, where a new set was made. This set, unfortunately, was destroyed by the Mongol invaders, and a second set was made. This second set, which still exists today at Haein Monastery, is the most complete set of printing blocks available for the Chinese canon (Lancaster and Park 1979). In the 19th century, prints from this second set were used in Tokyo to make a metal type print. In turn, this Tokyo edition was taken by the Tendai School as the model for the modern printed version known as the Taishō Edition. A digital edition of the Taishō is part of the Chinese Electronic Tripitaka Collection produced by the Chinese Buddhist Electronic Text Association (CBETA).

The Taishō Edition, however, does not represent the whole of the text glyphs found in the Korean block prints. Only 10,000 characters were available to the publishers and thus many substitutions of similar characters had to be made. When the digital version of the Korean Edition was made in the 1990s, every glyph found in the blocks was reproduced. Thus, this electronic edition is more accurate for detailed linguistic analysis than the metal type version of the Taishō. The Korean Edition forms the basis of the search and visualization platform developed by Lancaster (2010), and also serves as the underlying text of our treebank.

3.2 Texts

Our treebank contains more than 50K characters drawn from the *Diamond Sutra* (K13; T235), the *Nagasena Bhiksu Sutra* (K1002; T1670), the *Surangama Sutra* (K426; T945), and the *Vimalakirti Sutra* (K120; T474). The first two have been annotated in full, while the latter two have been partially annotated. Some statistics are shown in Table 1. There are a total of 4991 sentences, with an average length of 10.1 words per sentence.

Table 1 The four sutras in our treebank

Text	Number of tokens	Date
Vimalakirti Sutra (selections)	10,583	223—228 CE
Nagasena Bhiksu Sutra	14,393	317—420 CE
Diamond Sutra	6,128	401 CE
Surangama Sutra (selections)	18,908	705 CE

Volumes 1 to 3 of the *Surangama Sutra*, and the first five chapters of the *Vimalakirti Sutra* have been annotated; the other two sutras have been annotated in full.

The choice of these sutras is intended to represent a variety of literary genres, content, authors, and period of composition, so as to be useful for historical linguistics research and for training automatic parsers. The four authors for these sutras spanned almost five centuries; one sutra is drawn from each of the four traditional parts of the Canon: (1) the Collected Sutras (經集部), mostly written in the form of 'question-and-answer', to which belongs the *Vimalakirti Sutra*, the earliest text in the Canon; (2) the Treatises (論集部), with many metaphors and conversations, to which belongs the *Nagasena Sutra*; (3) the Esoteric Teachings (密教部), to which belongs the *Surangama Sutra*; and finally, (4) the Perfection of Wisdom (般若部), to which belongs the *Diamond Sutra*, one of the most important texts in the whole Canon. The *Surangama Sutra* contains much verse, while the others are prose.

4 Treebank Annotation

In this section, we describe the various types of annotations in the treebank, including metadata, punctuation, word segmentation, POS tagging, and dependency labeling.

4.1 Metadata

We associated with each sutra its year(s) of translation and its author/translator, all obtained from the catalog of the *Tripitaka Koreana* (Lancaster and Park 1979).

4.2 Punctuation

The original text had no punctuation. We used as our basis the punctuations of the text in the *Chinese Electronic Tripitaka Collection* from CBETA. Since the text is from the *Taishō* Edition rather than the Korean, we first aligned the two versions, then automatically inserted their punctuation to our text; the annotators then edited the punctuations as required.

4.3 Word segmentation

Since there is not yet a scholarly consensus on a precise definition of 'wordhood' in Classical Chinese (Feng 1998), treatment of word segmentation varies widely from corpus to corpus. Some did not perform word segmentation (Huang *et al.*, 2002); the tagged corpus of the book *Huainanzi* (Lau *et al.*, 2013) followed the guidelines of the Penn Chinese Treebank (Xue *et al.*, 2005), originally developed for Modern Chinese; the Academia Sinica Ancient Chinese Corpus (Wei *et al.*, 1997) and the Sheffield Corpus of Chinese (Hu and McLaughlin 2007) adopted their own principles.

Since we are not qualified to lay down any definitive criterion for word segmentation, we follow Lee's (2012) strategy in striving for a theory-neutral annotation method. Specifically, we offer two levels of segmentation, with the lower level aiming at the smallest possible unit of analysis, and the higher level identifying a coarser unit, thus enabling the user to select the most suitable level of granularity of wordhood, according to his/her research objective at hand.

Hence, at the lower level, we analyzed characters individually whenever possible. Two or more characters were deemed to form one word only when they do not have any 'internal structure', for example with foreign loanwords (e.g. *bi qiu* 比丘 'monk'), numbers (*shi wu* 十五 'fifteen'), reduplications (e.g. *qin qin* 駸駸 'quickly'), and bound morphemes (e.g. *you ran* 油然 'spontaneously'). For transliterations, we segmented the string according to the boundaries of the underlying parts of the compounds in the Indian original, e.g. ((般若)(波羅蜜)) 'wisdom – perfection' for prajñāpāramitā. This stringent criterion at the lower level yielded fewer multi-character words than would be expected for most other guidelines.

On the other hand, at the higher level, we marked word boundaries for compound nouns and verbs, geographical terms, as well as Buddhist terms with religious meaning. To ensure consistency, we used the Soothill-Hodous *Dictionary of Chinese Buddhist Terms*,[1] which contains about 15,000 entries, as our reference: every word that is a dictionary entry is marked at this higher level (if not already at the lower). A total of 3,030 words in our corpus were marked as a result. Consider the string *da bi qiu* 大比丘 'Grand Bhiksu'. Two words are recognized at the lower level: *bi qiu* 比丘 'monk', as a foreign

loanword, is considered one word; *da* 大 'grand' is also analyzed as a separate word (i.e., an adjective that modifies *bi qiu*). However, at the higher level, since the term *da bi qiu* is included in the abovementioned dictionary, all three characters are segmented as one word, resulting in the analysis ((*da* 大) (*bi qiu* 比丘)). This reflects the fact that the expression refers not to any senior bhiksu or *bi qiu*, but specifically to the Arhats, a group with the highest status (Cheng Kuan 2005).

This annotation method allows users to query the treebank at different levels. For example, the lower-level annotation would enable the retrieval of all words that modify *bi qiu* 比丘 'monk'; meanwhile, the higher-level annotation can identify all occurrences of the term *da bi qiu* 大比丘 'Grand Bhiksu' in the corpus. Our method is similar in spirit to the treatment of the Chinese Penn Treebank for certain compound verbs such as *zou shang lai* 走上來 'walk up' (Xia 2000), which was segmented at two levels as ((*zou* 走 'walk') (*shang lai* 上來 'up')).

4.4 POS tags

Similar to word segmentation, POS tagsets also vary according to corpus. The best-known tagset for Modern Chinese, developed for the Penn Chinese Treebank (Xue *et al.*, 2005), is adopted by the corpus of *Huainanzi* (Lau *et al.*, 2013) and by a treebank of Classical Chinese poems (Lee and Kong 2012). As for the two major corpora of Classical Chinese, the Academia Sinica Ancient Chinese Corpus (Wei *et al.*, 1997) and the Sheffield Corpus of Chinese (Hu and McLaughlin 2007), each has its own POS tagset, likely reflecting the research questions of interest to the respective groups.

We chose to follow Lau *et al.* (2013) and Lee and Kong (2012) in adopting the guidelines of the Penn Chinese Treebank, for two reasons. First, by adopting this widely used tagset, we will be able to leverage the considerable amount of modern Chinese data for training automatic taggers. Second, it will also facilitate contrastive studies with Modern Chinese, for which much data are encoded with the Penn guidelines.

Similar to the principle espoused by the Penn Chinese Treebank, we assign POS tags not according to the meaning of the word, but to syntactic distribution (Xia 2000), i.e. the role the word plays in the sentence. A case in point, which will be revisited in Section 6, is the use of adverb as copula. In a sentence with a nominal predicate but without the copula *shi* 是 'to be', most scholars accept that adverbs such as *nai* 乃 'to be' and *jie* 皆 'to be all' can function as copular verbs (Pulleyblank 1995).[2] These words are labeled as 'copula' (VC), rather than 'adverb' (AD), in such context.

While a number of tags specific to Modern Chinese (e.g. 'Resultative de5 的' (DER) and 'Manner de5 的' (DEV)) are no longer applicable in our corpus, we also found no need to introduce any new POS tag. This is consistent with the discussion in Section 2.2 of differences between Classical Chinese and Buddhist Chinese, which include lexical changes and new grammatical structures, but no new POS. To continue with our running example, *da bi qiu* is tagged as ((大/JJ) (比丘/NN))/NR, i.e. a proper noun (NR), with the internal structure where an adjective, *da* 'big' (JJ), modifies a common noun *bi qiu* 'monk' (NN). Table 3 lists the ten most frequent POS tags in our treebank.

4.5 Syntactic structure

Of the two existing treebanks of Classical Chinese, one is encoded with constituent structures (Huang *et al.*, 2002), and the other with dependencies (Lee and Kong 2012). In this work, we chose to follow the dependency framework for two main reasons. First, due to influences from the Indian originals, many sentences have unusual word order (Section 2.2), and a small minority has non-projective structures; the dependency framework is more flexible in treating these phenomena. Second, since our treebank is also intended to serve as reading aid, explicit grammatical relations are expected to be helpful.

Following Lee and Kong, we adopted the Stanford Dependencies for Chinese (Chang *et al.*, 2009) as our basis. To account for grammatical structures in Classical Chinese, Lee and Kong introduced four new dependency relations. These four relations are also present in our text, and an example for each is listed in Table 2. Beyond these

Table 2 The first four dependency relations were added to the Stanford Dependencies for Chinese (Chang *et al.*, 2009) by Lee and Kong (2012) to account for grammatical structures in Classical Chinese; the fifth, ExD, was added to annotate vocatives

Relation	Example
Locative modifier (lmod)	天竺 名 象 爲 那[a] 'Sindhu' 'call' 'elephant' 'as' 'naga' '[Since] elephants were called naga in Sindhu' lmod(名 'called', 天竺 'Sindhu')
Indirect object (iobj)	我 今 實 言 告 汝[b] 'I' 'now' 'true' 'word' 'tell' 'you' 'Right now I would like to impart this truth to you' iobj(告 'tell', 汝 'you')
Oblique objects (obl)	寢 疾 于 床[c] 'sleep' 'sickness' 'at' 'bed' 'Sleep on [his] bed because of sickness' obl(寢 'sleep', 疾 'sickness')
Noun phrase as adverbial modifier (npadvmod)	人 眾 日 多[d] 'people' 'crowd' 'day' 'many' 'The crowd grew day by day' npadvmod(多 'many', 日 'day')
External dependency (ExD)	不 也 世尊[e] 'no' <particle> 'World-Veneratedship' 'No, your World-Veneratedship' ExD(不 'no', 世尊 'World-Veneratedship')

The notation <label>(<head>, <child>) is used to represent the dependency relation being illustrated.
[a]Taken from *Nagasena Sutra*.
[b]Taken from *Diamond Sutra*.
[c]Taken from *Vimalakirti Sutra*.
[d]Taken from *Nagasena Sutra*.
[e]Taken from *Diamond Sutra*.

Table 3 The ten most frequent POS tags and dependency relations in our treebank

POS tag	Count	Dependency relation	Count
NN common noun	11,633	dobj direct object	4,540
VV other verb	9,802	nsubj nominal subject	4,217
AD adverb	4,425	advmod adverbial modifier	3,915
NR proper noun	3,119	conj conjunct	3,341
PN pronoun	2,104	nn noun compound modifier	2,421
JJ other noun-modifier	1,946	dep	2,254
P preposition	1,387	amod adjectival modifier	1,933
VE 有 as the main verb	1,123	ccomp clausal complement	1,931
DT determiner	1,105	prep prepositional modifier	1,373
VA predicative adjective	831	det determiner	1,094

The POS tagset is based on that of the Penn Chinese Treebank (Xue *et al.*, 2005), and the dependency relations on those of the Stanford Dependencies for Chinese (Chang *et al.*, 2009).

four, we introduced a relation for annotating vocatives, which are much more frequent in Buddhist sutras than in contemporary secular texts (Zhu 2010). Following the Prague Dependency Treebank, we use the relation 'External Dependency' (ExD), with the vocatives depending on their verbal heads. For the existing relations, the most significant change is the use of 'topic'

(top) in annotating non-copular sentences; this will be described in detail in Section 6. Table 3 lists the ten most frequent dependency relations in our treebank.

5 Annotator Agreement

Three annotators compiled this treebank: two had an academic background in Classical Chinese, the other in Buddhist Studies. To evaluate inter-annotator agreement, two annotators independently performed word segmentation, POS tagging, and dependency labeling for the first six segments of the *Diamond Sutra*, which consists of just over 800 characters.

5.1 Word segmentation

Using one annotator as the gold, the precision and recall for word boundary identification are 0.97 and 0.99, respectively, with no crossing brackets. These figures compare well with the agreement for Modern Chinese; for example, in an experiment reported by Sproat *et al.* (1996), the average human agreement rate was 0.76. This may not be a fair comparison, however, since most words in ancient Chinese consist of single characters, making the segmentation task much easier. Reflecting the process of disyllabication in Medieval Chinese (Section 2.2), about 10.5% of the words in our treebank contain two or more characters, an increase from the 6.5% reported for Classical Chinese poems (Lee 2012); this frequency is still much lower than in Modern Chinese.

The annotators disagreed on the wordhood of 19 two-character strings, which included nouns, verbs, and adjectives. While one annotator believed that the two characters express one single meaning (e.g. *xu wang* 虛妄 'vain and delusive') and hence should be considered one compound word, the other thought they had distinct meanings (e.g. 'vain' and 'delusive', respectively, representing two different qualities) and hence constituted two separate words. Another example is *si liang* 思量 'think and measure'. Most of these are entries in the Soothill-Hodous *Dictionary of Chinese Buddhist Terms*, and so both interpretations are in practice accommodated.

5.2 POS tags

The inter-annotator agreement rate on POS tags[3] was 0.98. For nouns, the most frequent confusions are demonstrative pronouns (PN) versus determiners (DT), and common nouns (NN) versus temporal (NT) or proper nouns (NR). For verbs, the disagreement typically occurred in a sequence of two verbs (e.g. *hezhang gongjing* 合掌 恭敬 'join palms' 'pay respect'). One annotator would interpret the sequence as two separate actions ('join palms and paid respect'), thus assigning VV to both; the other would understand this as one action, with one of the verbs as supplying the manner, i.e. acting as an adverb (e.g. 'join palms reverently') for the other, thus labeling one as VV and the other as AD.

5.3 Dependency labels

The inter-annotator agreement rate of labeled head selection (i.e. agreement on the POS tag, the identity of the head, and the dependency label)[4] is 0.87. Leaving aside disagreements on head selection and dependency label that trickle down from disputed POS, there are three main problematic areas. First, since discourse relations tend to be unmarked in Classical Chinese, in a complex sentence with two clauses, it is often unclear which one is the main clause (and hence its verb should be the root) and which one is subordinate (and hence its verb depend on the other clause).

Second, for some compound words, especially those with religious meaning, the relation between the two characters can be difficult to pin down. For example, in the compound verb *zhuxiang* 住相 'dwell on appearance', one can plausibly defend the labeling of the noun *xiang* 相 'appearance' as the direct object (dobj), oblique object (obl), or even locative modifier (lmod) of the verb *zhu* 住 'dwell'.

Third, and the most frequent, is the confusion between 'noun compound modifier' (nn), 'associative modifier' (assmod), and 'conjunct' (conj), all common relations between two nouns. In Modern Chinese, the distinction is clear: 'associative

modifier' is used only when the two nouns are linked by *de* 的, the genitive marker, and 'conjunct' is used only when they are linked by a coordinating conjunction. In both Classical and Medieval Chinese, these function words can be omitted and therefore a subjective judgment is often required. For example, the noun compound *fu de* 福德 may have any one of the above relations depending on one's understanding it as 'blissful virtue', 'blessing and virtue', or 'virtue of bliss', etc.

6 Application

This section applies our treebank to the study of a major development in the history of the Chinese language—the emergence of the copula. We first describe the search platform (Section 6.1), which enables Humanities scholars without technical background to replicate our study; we then give some linguistic background on the copula (Section 6.2), explain our methodology (Section 6.3), and finally analyze our results (Section 6.4).

6.1 Corpus platform

Our treebank is accessible on the web via the ANNIS system, an open source, browser-based corpus search platform for richly annotated corpora (Zeldes *et al.*, 2009). Figure 1 shows our treebank in the ANNIS interface.

Conceptually, a corpus in ANNIS can be thought of as a graph with nodes and edges, where nodes represent tokens and annotations, and edges represent relations between them. As with most typical treebanks, the tokens are stored as nodes with multiple annotation layers including word boundaries and POS tags; and the dependency relations are stored as edges between nodes. This architecture is flexible for adding other layers; for example, we have added two additional layers, one for the two-level segmentation and POS tagging (Section 4), and another for storing definitions from the Soothill-Hodous *Dictionary of Chinese Buddhist Terms*.

The ANNIS Query Language facilitates search on patterns in dependency trees. Figure 1, for example, shows the search results for the query on all sentences with a word with the POS tag 'copula' (VC)

Fig. 1 Results of a query in ANNIS (Zeldes *et al.*, 2009) that searches for all sentences in the treebank with a copula and a nominal predicate

that serves as head of a noun in the relation 'attributive' (attr); i.e. a copular sentence (most frequently, *shi* 是 'to be') with a nominal predicate. A handful of similar queries suffice to compute the statistics in the study below.

6.2 Background

When the predicate of an English sentence consists of a noun phrase, the copula 'to be' links it to the subject, yielding 'N1 is N2'. Modern Chinese also has a similar copula, *shi* 是 'to be'. In Classical Chinese, however, there is normally no copula (Pulleyblank 1995). Over the centuries, the copula gradually gained in popularity, and in Modern Chinese almost all nominal predicates have a copula. How and when the copula emerged remains an open question in Chinese linguistics (Wang 1998; Feng 2003).

The Chinese Buddhist Canon is uniquely positioned to provide evidence on this question because

it contains a large quantity of text that is consecutive in time. Further, since it contains more vernacular elements than contemporary secular literature, it can be expected to provide more accurate chronology.

To observe the development of the copula, we need to compare the number of copular and non-copular sentences. Since current digital versions of the Canon do not yet have POS tags, it is difficult to locate the copula because the character *shi* 是 'to be' also frequently serves as a determiner or a pronoun—a naive string search would not do. But even POS tags alone cannot identify sentences with noun predicates that lack copulas; syntactic annotations are necessary. These obstacles explain why most previous studies on this topic can only provide a limited number of anecdotal examples.

In the most exhaustive study to date, Xie (2006) manually counted the number of copular and non-copular sentences in two Classical Chinese texts, *Zuozhuan* 左傳 and *Shiji* 史記. The ratio of non-copular sentences decreased from 99.3% in *Zuozhuan*, composed in late fifth century BCE, to 87.4% in *Shiji*, finished in 122 BCE.

6.3 Methodology

In the Stanford Dependencies for Chinese, the copular sentence is labeled with the dependency relations `top` ('topic') and `attr` ('attributive'). The relation `top` has N1 as child and the copula as head, while `attr` has N2 as child and the copula as head. When the copula is absent, there is no `attr` relation, and we assign N2 as the head of `top` instead. While it is possible to use an empty node to represent the 'missing' copula (from the modern perspective), we have opted for this simpler solution. See Figures 2 and 3 for examples of both copular and non-copular sentences.

In ANNIS, then, in order to retrieve copular sentences, it suffices to search for `attr` relations whose head words are tagged as `VC` ('copula'); as for non-copular sentences, to search for `top` relations whose heads are not tagged as `VC`. Portions of the search results are shown in Figure 1. In addition to *shi*, we also consider *wei* 爲 'to be' as copula, as well as a number of adverbs, such as *nai* 乃 'to be' and *jie* 皆 'to be all', that function as copula.

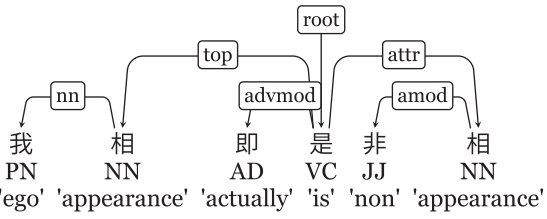

Fig. 2 A dependency tree for a copular sentence, 'Ego appearance is actually non-appearance', which is in the form 'N1 is N2'. Both N1 (the first *xiang* 相 'appearance') and N2 (the second *xiang* 相 'appearance') depend on the copula *shi* 是 'is', via the `top` and `attr` relations

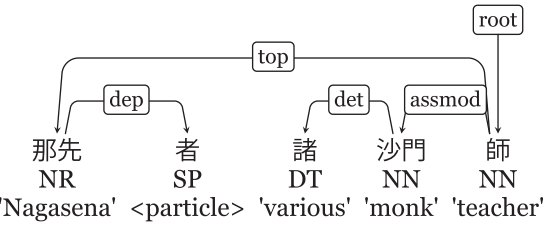

Fig. 3 A dependency tree for a non-copular sentence 'Nagasena is the teacher of various monks', which is in the form 'N1 N2'. In our annotation, N1 (*naxian* 那先 'Nagasena') depends directly on N2 (*shi* 師 'teacher') via the `top` relation

Finally, we compare our results with modern Chinese data, using 1.3 million words from a subset of the Penn Chinese Treebank (LDC2000T48 and LDC2007T36). We identify all constituents marked as 'nominal predicate' (`NP-PRD`), and examine whether any 'copula' (`VC`) precedes them.

In an ideal diachronic study, the date of composition should be the only variable, with all texts being similar in all other aspects. In the present study, there are two other aspects that can affect the number of copular sentences observed. First, the texts under consideration belong to different genres, where the copula may have been used to different degrees. Thus, the statistics from historical works such as *Zuozhuan* and *Shiji* may not be directly comparable to those from the sutras, or to the newswire text from the Penn Chinese Treebank. Second, even among the sutras, there are variations

Table 4 The proportion of non-copular sentences, out of all sentences with nominal predicates

Text	Non-copular sentences (%)
Zuozhuan	99.3
Shiji	87.4
Vimalakirti Sutra	22.6
Nagasena Bhiksu Sutra	13.2
Diamond Sutra	1.0
Surangama Sutra	4.8
Modern Chinese	4.9

The texts are displayed in chronological order, from earliest to the most recent. Results for *Zuozhuan* and *Shiji* are due to Xie (2006). Modern Chinese refers to LDC2000T48 and LDC2007T36.

in style and in degrees of influence from the Indian original. For example, the *Diamond Sutra* is particularly formulaic in its structure, and the origins of the *Surangama Sutra* are disputed. These limitations should be borne in mind when interpreting the results in the next section.

6.4 Analysis

Counting both copular and non-copular sentences, our treebank has a total of 797 sentences with nominal predicates. Table 4 shows the overall trend of the diminishing popularity of the non-copular sentence. Once the norm in Classical Chinese literature, such as *Zuozhuan* (99.3%), their usage appeared to have dropped sharply in three centuries—from 87.4% at about 100 BCE, when *Shiji* was composed, down to 22.6% at around 200 CE, the time of the *Vimalakirti Sutra*. A note of caution is that this drop may be amplified by the difference in literary genre, since Chinese Buddhist texts tend to contain more vernacular elements (Karashima 1996), including the copula, compared to secular literature in the same period.

The proportion of non-copular sentences continues to drop for the next two sutras in chronological order, the *Nagasena Sutra* and the *Diamond Sutra*. The most recent sutra in the treebank, the *Surangama Sutra*, has a similar proportion as Modern Chinese (both about 5%), as sampled in the Penn Chinese Treebank. The outlier in this analysis is the *Diamond Sutra*, where the use of the copula is relatively abundant. A possible reason is that this short and intense text contains more emphatic affirmations and negations, conveyed via the copula.

7 Conclusion

We have presented the first dependency treebank of Medieval Chinese, which contains four sutras drawn from the Chinese Buddhist Canon, the largest and one of the most significant bodies of text in ancient China. These sutras represent a variety of genres, authors, and periods of compositions.

We have described the construction process of the treebank by motivating the use of the dependency framework and a two-level method for segmentation and POS tagging. In an evaluation, we have shown a high degree of inter-annotator agreement and analyzed the areas of confusion.

Further, we have demonstrated the treebank's potential for diachronic linguistics research by applying it to a long-standing research question—the emergence of the Chinese copula. Our results, which can be easily replicated on a state-of-the-art search platform, illustrate the gain in frequency of the copula in Medieval Chinese, as one progresses from the earliest sutra to the most recent.

In future work, we plan to use the treebank to train a dependency parser, in order to automatically annotate the rest of the Chinese Buddhist Canon, as well as other genres of Buddhist literature. The resulting treebank would be able to address research questions in more breadth and provide quantitative analysis to many open questions in Chinese historical linguistics.

Funding

The work described in this article was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 155412).

References

Bamman, D. and Crane, G. (2008). *Building a Dynamic Lexicon from a Digital Library, Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*. Pittsburg, PA.

Bingenheimer, M., Hung, J.-J., and Wiles, S. (2011). Social network visualization from TEI data. *Literary and Linguistic Computing*, **26**(3): 271–8.

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

Cheng, K. (2005). *The Diamond Prajna-Paramita Sutra: An Annotated Edition with Chinese Text*. Taipei: Vairocana Publishing Co. Ltd.

Demske, U., Frank, N., Laufer, S., and Stiemer, H. (2004). *Syntactic Interpretation of an Early New High German Corpus, Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT)*. Tübingen, Germany.

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

Feng, S. (1998). Prosodic structure and compound words in classical Chinese. In Packard, J. (ed.), *New Approaches to Chinese Word Formation*. Berlin: Mouton de Gruyter.

Feng, S. (2003). Copular in classical Chinese assertive sentences [in Chinese]. *Research in Ancient Chinese Language*, **58**: 30–6.

Haug, D. and Jøhndal, M. (2008). *Creating a Parallel Treebank of the Old Indo-European Bible Translations, Proceedings of the LREC Workshop on Language Technology for Cultural Heritage Data (LaTeCH)*. Marrakech, Morocco.

Hu, X. and McLaughlin, J. (2007). *The Sheffield Corpus of Chinese*. Sheffield: Technical Report, University of Sheffield.

Hu, X., Williamson, N., and McLaughlin, J. (2005). Sheffield Corpus of Chinese for diachronic linguistic study. *Literary and Linguistic Computing*, **20**(3): 281–93.

Huang, L., Peng, Y., Wang, H., and Wu, Z. (2002). *PCFG Parsing for Restricted Classical Chinese Texts, Proceedings of the 1st SIGHAN Workshop on Chinese Language Processing*. Taipei, Taiwan.

Hung, J.-J., Bingenheimer, M., and Wiles, S. (2010). Quantitative evidence for a hypothesis regarding the attribution of early Buddhist translations. *Literary and Linguistic Computing*, **25**(1): 119–34.

Jiang, S. and Hu, C. 蔣紹愚 胡敕瑞 (2013). 漢譯佛典語法研究論集 [in Chinese]. Beijing: The Commercial Press.

Karashima, S. (1996). On vernacularisms and transcriptions in early Chinese Buddhist scriptures. *Vernacularisms in Medieval Chinese Texts: Sino-Platonic Papers*, **71**. Philadelphia: University of Pennsylvania.

Kroch, A., Santorini, B., and Diertani, A. (2004). *Penn-Helsinki Parsed Corpus of Early Modern English*. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html.

Lancaster, L. (2010). *From Text to Image to Analysis: Visualization of Chinese Buddhist Canon, Proceedings of Digital Humanities*. London, UK.

Lancaster, L. and Park, S. (1979). *The Korean Buddhist Canon: A Descriptive Catalogue*. Berkeley: Berkeley University Press.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**(1): 159–74.

Lau, K. T., Song, Y., and Xia, F. (2013). *The Construction of a Segmented and Part-of-speech Tagged Archaic Chinese Corpus: A Case Study on Huainanzi [in Chinese], Proceedings of the 12th China National Conference on Computational Linguistics (CNCCL)*. Suzhou, China.

Lee, J. (2012). *A Classical Chinese Corpus with Nested Part-of-Speech Tags. Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Avignon, France.

Lee, J. and Kong, Y. H. (2012). *A Dependency Treebank of Classical Chinese Poems, Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL)*. Montreal, Canada.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., and Petrov, S. (2012). *Syntactic annotations for the Google Books Ngram Corpus, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*. Jeju, Korea.

Mihalcea, R. and Nastase, V. (2012). *Word Epoch Disambiguation: Finding How Words Change Over Time, Proceedings of the 50th Annual Meeting of the*

Association for Computational Linguistics (ACL). Jeju, Korea.

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

Rocio, V., Alves, M. A., Lopes, J. G., Xavier, M. F., and Vicente, G. (2000). Automated creation of a medieval portuguese partial treebank. In Abeillé, A. (ed.), *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer Academic Publishers, pp. 211–27.

Sproat, R., Shih, C., Gale, W., and Chang, N. (1996). A stochastic finite-state word-segmentation – algorithm for Chinese. *Computational Linguistics*, **22**(3): 377–404.

Taylor, A., Warner, A., Pintzuk, S., and Beths, F. (2003). *York-Toronto-Helsinki Parsed Corpus of Old English Prose*. Heslington: University of York.

Wang, L. (1998). *Classical Chinese* 古代漢語 [in Chinese]. Beijing: Zhonghua Press.

Wei, P.-C., Thompson, P. M., Liu, C.-H., Huang, C.-R., and Sun, C. (1997). Historical corpora for synchronic and diachronic linguistics studies. *Computational Linguistics and Chinese Language Processing*, **2**(1): 131–45.

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

Xie, Z. 解植永 (2006).A comparative research on declarative sentences in Zuozhuan and Shiji上古判斷句比較研究 [in Chinese]. *Journal of Chongqing University of Arts and Sciences*, **5**(3): 45–9.

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.

Zeldes, A., Ritz, J., Luüdeling, A., and Chiarcos, C. (2009). *ANNIS: A Search Tool for Multi-Layer Annotated Corpora*. Liverpool: Proceedings of Corpus Linguistics.

Zhu, Q. Z. (2008). The inessive structure in archaic and medieval Chinese: An evolutionary study of inessive demonstrative uses from archaic to early modern Chinese. In Xu, D. (ed.), *Space in Language of China*. New York: Springer, pp. 249–66.

Zhu, Q. Z. (2009). 佛教漢語研究 *[in Chinese]*. Beijing: The Commercial Press.

Zhu, Q. Z. (2010). On some basic features of Buddhist Chinese. *Journal of International Association of Buddhist Studies*, **31**(1-2): 485–504.

Notes

1 Accessible at http://mahajana.net/texts/kopia_lokalna/soothill-hodous.html

2 We acknowledge that some linguists consider these adverbs as semi-copular verbs only (Wang 1998).

3 The kappa is 0.978, indicating a high degree of agreement (Landis and Koch 1977).

4 The kappa is 0.868, indicating a high degree of agreement (Landis and Koch 1977).