Natural Language Engineering **11** (2): 207–238. © 2005 Cambridge University Press
doi:10.1017/S135132490400364X Printed in the United Kingdom

The Penn Chinese TreeBank: Phrase structure annotation of a large corpus

NAIWEN XUE, FEI XIA,* FU-DONG CHIOU and MARTA PALMER
University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: {xueniwen,fxia,chioufd,mpalmer}@linc.cis.upenn.edu

(*Received 3 October 2002; revised 4 November 2003*)

Abstract

With growing interest in Chinese Language Processing, numerous NLP tools (e.g., word segmenters, part-of-speech taggers, and parsers) for Chinese have been developed all over the world. However, since no large-scale bracketed corpora are available to the public, these tools are trained on corpora with different segmentation criteria, part-of-speech tagsets and bracketing guidelines, and therefore, comparisons are difficult. As a first step towards addressing this issue, we have been preparing a large bracketed corpus since late 1998. The first two installments of the corpus, 250 thousand words of data, fully segmented, POS-tagged and syntactically bracketed, have been released to the public via LDC (www.ldc.upenn.edu). In this paper, we discuss several Chinese linguistic issues and their implications for our treebanking efforts and how we address these issues when developing our annotation guidelines. We also describe our engineering strategies to improve speed while ensuring annotation quality.

1 Introduction

The creation of annotated corpora has led to major advances in corpus-based natural language processing technologies. Most notably, the Penn English Treebank (Marcus, Santorini and Marcinkiewicz 1993) has proven to be a crucial resource in the recent success of English Part-Of-Speech (POS) taggers and parsers (Collins 1997, 2000; Charniak 2000), as it provides common training and testing material so that different algorithms can be compared and progress be gauged. Its success triggered the development of treebanks in a variety of languages. As displayed in a recent book on treebanks (Abeillé 2003), there are efforts in progress for Czech, German, French, Japanese, Polish, Spanish and Turkish, to name just a few. Specific to Chinese, however, most of the annotation effort has been devoted to word segmentation (tokenization) and POS tagging. Several segmented and POS tagged corpora have been developed, based on standards published in different Chinese-speaking regions,

* This work was done while the author was a graduate student at the University of Pennsylvania. The author currently is a research staff member at the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA.

most notably the Beijing University Institute of Computational Linguistics Corpus (PKU) (Yu, Zhu, Wang and Zhang 1998) and the Academia Sinica (Taiwan) Balanced Corpus (ABSC) (CKIP 1995). More recently, the LIVAC synchronous corpus[1] has been developed at City University of Hong Kong. However, there has been a general lack of syntactically annotated Chinese corpora which hinders the development of Chinese NLP tools and makes it difficult to compare results and measure progress in Chinese language processing. In fact, there was no publicly available syntactically bracketed Chinese treebank when the Penn Chinese Treebank was started in late 1998 to address this need. The first installment of the Penn Chinese Treebank (CTB-I hereafter), a 100 thousand words of annotated Xinhua[2] newswire articles, along with its segmentation (Xia 2000b), POS-tagging (Xia 2000a) and syntactic bracketing guidelines (Xue and Xia 2000), was released in the fall of 2000 (see the Appendix for the timeline). The second installment of the Penn Chinese Treebank (CTB-II hereafter)[3], containing an additional 150,000 words, and beginning to include Hong Kong News and *Sinorama*[4] articles in an attempt to diversify its data source, was released in the spring of 2003. The eventual goal of this on-going project is to build a large-scale Chinese corpus as a sharable resource that addresses the need for training and testing material in the Chinese NLP community.

Building a treebank requires tremendous human effort. To ensure high quality while maintaining reasonable annotation speed is a major challenge. In order to speed up the annotation, we use a series of NLP tools to preprocess the data at different stages of annotation. We also adopt several strategies to control the quality of the annotation: (i) a significant effort is devoted to the creation of clear, consistent, and complete annotation guidelines; (ii) all the annotation in the treebank is double checked by a second annotator; (iii) a gold standard is created and annotation accuracy and inter-annotator agreement are monitored; and (iv) the treebank goes through a final cleanup with semi-automatic tools before the release.

While the engineering strategies may be language-independent, creating a treebank for a particular language also requires a thorough study of the language itself, especially its morphology and syntax. The properties of the language should be taken into consideration when designing the overall annotation paradigm and writing annotation guidelines. For instance, Chinese written texts do not contain word delimiters. To build a treebank for Chinese, we need to break a sentence into a word sequence before adding POS tags and phrase structures. Chinese also lacks inflectional morphology, a property that complicates all aspects of Chinese text annotation: word segmentation, POS tagging, and syntactic bracketing. As a result, many diagnostic tests that work well for English do not work for Chinese, and new diagnostic tests have to be found when developing annotation guidelines. The multitude of differences between Chinese and Indo-European languages have led

[1] More information can be found at `www.rcl.cityu.edu.hk/english/livac`.
[2] Xinhua is the official news agency of the People's Republic of China.
[3] CTB-I is released by LDC as Chinese Treebank Versions 1.0 and 2.0. CTB-II is included in Chinese Treebank Version 3.0
[4] *Sinorama* is a Taiwan news magazine.

many Chinese linguists to doubt the feasibility of applying standard Western-style phrase structure analysis to Chinese. As a result, other recent efforts to build Chinese treebanks have adopted a different approach, putting more emphasis on providing semantic analysis. For example, Li, Li, Dong, Wang and Lu (2003) have elected to annotate dependency structures, along the lines of the Prague Dependency Treebank (Böhmová, Hajič, Hajicová and Hladká 2003). The Sinica Treebank (Chen, Huang, Chen, Luo, Chang and Chen 2003) also has a more semantic orientation, although it does provide simple syntactic analysis. The Penn Chinese Treebank represents the only attempt to provide full phrase structure for complete sentences in Chinese as the Penn English Treebank did for English. However, CTB goes further than the English Treebank in marking dropped arguments, providing argument/adjunct distinctions, and some NP-internal structure. Its efficacy for training statistical parsers has been validated by the development of several different systems (Bikel and Chiang 2000; Levy and Manning 2003; Luo 2003).

This paper is organized as follows. In section 2, we discuss several Chinese linguistic issues and our basic strategies for creating a high-quality treebank. In section 3, we address major problems that we encountered when creating three sets of annotation guidelines (for word segmentation, POS tagging and syntactic bracketing, respectively). In section 4, we briefly compare the design of our treebank with that of the Penn English Treebank (Marcus, Kim and Marcinkiewicz 1994) and the Sinica Treebank (Chen *et al.* 2003). In section 5, we discuss our approach to speed up the annotation and to control quality. Specifically, we describe how we use a word segmenter, a POS tagger, and a parser to speed up annotation, and use LexTract (Xia 2001) and CorpusSearch to find annotation errors. Section 6 concludes this paper, and describes future directions.

2 Linguistic issues and engineering strategies

In this section we first give an overview of the Penn Chinese Treebank as a treebanking task. Then we outline several Chinese linguistic issues that have to be addressed when preparing the guidelines. Next we discuss the engineering issues of this project and our basic strategies for addressing them as well.

2.1 An overview of the Penn Chinese Treebank

The data in the Penn Chinese Treebank are mostly newswire and magazine articles from Xinhua newswire, Hong Kong news and the *Sinorama* magazine. The structure of the original articles is maintained as much as possible without modification or editing. CTB-I, the first installment of the Penn Chinese Treebank, includes 325 articles of Xinhua newswire. Most of the articles focus on economic development from 1994 to 1998, while the remaining documents describe general political and cultural topics at the same period of time. The average sentence length is 28.7 words.[5]

[5] In our treebank, we use periods, exclamation marks, and questions marks to break a document into a sequence of sentences. We do not use commas, in contrast with the Academia Sinica Treebank. See section 4.2 for details.

Starting with CTB-II, we began to include data sources other than Xinhua newswire. CTB-II, the second installment of the treebank, contains an additional 150,000 words and includes 373 articles of Xinhua newswire (130,000 words), 55 articles of Hong Kong News (15,000 words), and two articles from *Sinorama* (6000 words). The average sentence length is 28.9 words. We are currently working on the third installment of the treebank which will continue to diversify our data sources.

The task of annotating sentences in this treebank can be broken into three subtasks: word segmentation, part-of-speech tagging and syntactic bracketing. This process is illustrated in (1): (1a) is an example Chinese sentence before annotation, (1b), (1c) and (1d) illustrate the same sentence after segmentation, POS-tagging and syntactic bracketing, respectively.[6] The actual annotation is carried out in two-stages: word segmentation and POS tagging are performed first, and phrase structures are added later.

2.2 Addressing Chinese linguistic issues

The development of a large-scale annotated Chinese corpus pushes to the forefront some fundamental issues in Chinese linguistics. In this section we outline a few of them, and discuss their implications for our annotation efforts. We focus on three issues: (i) the feasibility of the word segmentation task; (ii) the impoverished inflectional morphology of Chinese; and (iii) difficult constructions in Chinese syntax.

2.2.1 The feasibility of the word segmentation task

As demonstrated in Example 1, a Chinese sentence is a sequence of Chinese characters without natural delimitors between words. As a result, it has to be segmented into words before POS tags and phrase structures can be added. The feasibility of word segmentation as an annotation task for Chinese[7] has been a subject of considerable research interest. Sproat, Gale, Shih and Chang (1996), for example, reported experimental results that show native speakers of Chinese have a very low degree of agreement among them as to what a word is. In their experiments, six native speakers were asked to mark all the places where they might pause if they were reading the text aloud. The inter-judge agreement reported is only 76%. However, the experiments were set up in the context of a text-to-speech synthesis system and thus the results may not speak directly to the feasibility of a more general word segmentation task.

To test how well native speakers agree on word segmentation of written texts, we randomly chose 100 sentences (5060 *hanzi*) from the Xinhua newswire and

[6] The English gloss of the Chinese examples throughout this paper is not part of the annotation. It is included for the convenience of non-Chinese speakers.

[7] Even for languages which use delimiters between words, such as English, the distinction between a *word* and a non-word is not always clear-cut. For example, *pro-* normally cannot stand alone, therefore, it is like a prefix. However, it can appear in a coordinated structure, such as *pro-* and *anti-abortion*, and under the assumption that only words and phrases can be coordinated, it is a word. As a reviewer pointed out, deciding word boundaries is also a difficult task for other languages, such as Portuguese (Santos, Costa and Rocha 2003).

(a) Raw data:

他还提出一系列具体措施和政策要点。

(b) Segmented:

他	还	提出	一	系列	具体	措施	和	政策	要点	。
He	also	propose	one	series	concrete	measure	and	policy	essential	.

(He also proposed a series of concrete measures and essentials on policy.)

(c) POS-tagged:

他/PN 还/AD 提出/VV 一/CD 系列/M 具体/JJ 措施/NN 和/CC 政策/NN 要点/NN 。/PU

(d) Bracketed:

```
(IP (NP-SBJ (PN 他/he))
    (VP (ADVP (AD 还/also))
        (VP (VV 提出/propose)
            (NP-OBJ (QP (CD 一/one)
                          (CLP (M 系列/series)))
                     (NP (NP (ADJP (JJ 具体/concrete))
                              (NP (NN 措施/measure)))
                          (CC 和/and)
                          (NP (NN 政策/policy)
                              (NN 要点/essential))))))
    (PU 。))
```

Example 1. A sample Chinese sentence.

asked the participants of the First Chinese Language Processing Workshop, which was held at the University of Pennsylvania in 1998, to segment them according to their personal preferences.[8] We got replies from eight groups, and all but one of them hand corrected their output before sending it. To measure the agreement between each pair of the groups that did hand correction, we use three measures that are widely used to measure parsing accuracy: precision, recall, and the number of crossing brackets (Black, Abney, Flickinger, Gdoniec *et al.* 1991).[9] Following

[8] We did not give them any segmentation guidelines. Some participants applied their own guideline standards for which they had automatic segmenters while others simply used their intuitions.

[9] Given a candidate file and a Gold Standard file, the three metrics are defined as: precision is the number of correct constituents in the candidate file divided by the number of constituents in the candidate file; recall is the number of correct constituents in the candidate file divided by the number of constituents in the Gold Standard file; and the number of crossing brackets is the number of constituents in the candidate file that cross a constituent in a Gold Standard file.
If we treat each word as a constituent, a segmented sentence is similar to a bracketed sentence and its depth is one. To compare two outputs, we chose one as the Gold

Table 1. *Comparison of hand-corrected word segmentation results from seven groups*

	1	2	3	4	5	6	7	Average
1	–	90/88/6	90/90/4	83/88/3	92/91/3	91/91/3	92/84/9	90/89/5
2	88/90/3	–	87/90/3	80/88/14	89/90/4	86/89/3	89/83/7	87/88/6
3	90/90/3	90/87/5	–	82/88/2	89/88/5	89/89/4	89/82/10	88/87/5
4	88/83/9	88/80/10	88/82/7	–	92/86/7	86/81/9	87/74/16	88/81/10
5	91/92/3	90/89/4	88/89/4	86/92/9	–	90/90/4	92/85/8	90/90/5
6	91/91/3	89/86/6	89/89/4	81/86/3	90/90/4	–	91/83/10	89/88/5
7	84/92/1	83/89/2	82/89/2	74/87/4	85/92/1	83/91/1	–	82/90/2

Sproat *et al.* (1996), we calculate the arithmetic mean of the precision and the recall as one measure of agreement between each output pair, which produces an average agreement of 87.6%, much higher than the 76% reported in Sproat *et al.* (1996). Table 1 shows the results of comparing the output between each group pair. For each $x/y/z$ in the table, x and y are precision and recall rates, respectively, and z is the *total* number of crossing brackets in the 100 sentences.

The fact that the average agreement in our experiment is 87.6% and the highest agreement among all the pairs is 91.5% confirms the belief that native speakers do have significant disagreement on where word boundaries should be. On the other hand, on average there are only 5.4 crossing brackets in the 100 sentences, and most of these crossing brackets turned out to be human errors. This suggests that much of the disagreement is not critical and if native speakers are given good segmentation guidelines, consistent word segmentation can be achieved. There are several possible explanations for the discrepancy between our results and those reported in Sproat *et al.* (1996). One is that the instructions given to the judges are different. In our experiment, the judges were asked to segment the sentences into *words* based on their own definitions, while in their experiment, the judges were asked to mark all places where they might possibly pause if they were reading the text aloud. There are places in Chinese, such as the place between a verb and an aspect marker that follows the verb, where native speakers normally do not pause but would add word boundaries if asked to segment the sentence. Pragmatic factors can also influence a decision to pause which would be independent of word segmentation. Another reason why the degree of agreement in our experiment was much higher is that in our experiment all the judges were well-trained computational linguists who are familiar with both the linguistic and computational issues of the word segmentation task. Some judges had their own segmentation guidelines and/or segmenters. They either followed their guidelines or used their segmenters to automatically segment the

Standard, and evaluated the other output against it. As noted in Sproat *et al.* (1996), for two outputs J_1 and J_2, taking J_1 as the Gold Standard and computing the precision and recall for J_2 yields the same results as taking J_2 as the Gold Standard and computing the recall and the precision respectively for J_1. However, the number of crossing brackets when J_1 is the standard is not the same as when J_2 is the standard. For example, if the string is *ABCD* and J_1 segments it into *AB CD* and J_2 marks it as *A BC D*, then the number of crossing brackets is 1 if J_1 is the standard and the number is 2 if J_2 is the standard.

data and then hand corrected the output. As a result, their resulting segmentation is more consistent. Taken as a whole, the results show that word segmentation is feasible as an annotation task. Therefore, it is reasonable to assume that, given the same set of guidelines, the human agreement on segmentation would be well over 90%.

2.2.2 The impoverished morphological system

The second characteristic of Chinese that has far-reaching consequences on the annotation of Chinese text is the fact that Chinese has very little, if any, inflectional morphology. This general lack of morphological clues affects every aspect of Chinese text annotation: word segmentation, POS-tagging and syntactic bracketing. For instance, if there were abundant prefixes or suffixes in Chinese, they could be used to signal the beginning or the end of a word even in the absence of natural delimiters. Without these convenient means for word boundary detection, in the word segmentation guidelines we have to resort to phonological, syntactic, and semantic tests to decide on proper word boundaries.

The lack of inflectional morphology simplifies some aspects of the POS tagging task. For instance, lemmatization is usually not necessary in Chinese POS tagging. In general, however, this characteristic of the language makes the POS-tagging task harder. Determining the POS tag of a word becomes less straightforward because of the lack of morphological clues. For a word that is ambiguous between a noun and a verb, to determine its part-of-speech requires a careful analysis of its syntactic environment. (See section 3.2 for a detailed discussion of our methodologies for POS-tagging nouns and verbs.)

The lack of morphological clues also has implications for determining the subcategorization frames of verbs, which is crucial in deciding the syntactic structure of a clause. For a language like English, morphological clues can be used to signal the subcategorization frame of a verb, thus the syntactic structure of a clause. For instance, based on the morphological clues, it is easy to distinguish a verb with a sentential complement such as *say* from an object control verb such as *force*: the structural distinction between "*John said that he would come*" and "*John forced him to come*" can easily be made with the help of the complementizer *that*, the infinitive marker *to*, and the case marked pronouns *he* and *him*. In Chinese, similar distinctions can also be made between "说/say" and "逼/force". Although no morphological difference can be observed between (2a) and (2b), the structural differences become clear when we consider the contrast between (2c, 2e) and (2d, 2f). For example, "说/say" can take an existential construction as its complement but "逼/force" cannot, as illustrated in (2c) and (2d). Also, "说/say", but not "逼/force", can take a *bei*-construction as an object, as shown in (2e) and (2f).[10] To annotate the syntactic structures of these verbs accurately and consistently, more elaborate diagnostic tests such as the ones in (2) have to be provided in the guidelines.

[10] The *bei*-construction indicates passivization, but its structure is different from the English passive construction.

(∗ indicates ill-formedness)
(a) 张三/Zhangsan 说/say 他/he|him 来/come
"Zhangsan said he would come."
(b) 张三/zhangsan 逼/force 他/he|him 来/come
"Zhangsan forced him to come."
(c) 张三/Zhangsan 说/say 房间/room 里/inside 有/have 人/person
"Zhangsan said there was someone in the room."
(d) ∗ 张三/zhangsan 逼/force 房间/room 里/inside 有/have 人/person
(e) 张三/Zhangsan 说/say 李四/Lisi 被/Bei 打/hit 了/ASP
"Zhangsan said Lisi was hit."
(f) ∗ 张三/zhangsan 逼/force 李四/Lisi 被/Bei 打/hit 了/ASP

Example 2. Diagnostic tests for the object control verbs and sentential complement verbs (Xue and Xia 2000).

2.2.3 Difficult constructions in Chinese syntax

Our data are from a variety of different sources and it is not uncommon to find structures that have no existing analysis in the Chinese linguistic literature. This makes it impossible to simply follow some grammar book or books, as seems to be the case in the development of the Prague Dependency Treebank (Hajič 1998). There are also some notoriously difficult constructions where linguists have yet to agree on an analysis. Many of them, such as the *ba*-construction and the *bei*-construction, have been investigated for decades, but there is still no consensus on how they should be analyzed. For example, the word "把/BA" in the *ba*-construction has been argued to be a case marker, a secondary topic marker, a preposition, a verb, and so on in the literature (Bender 2000). The analysis of the construction will affect both POS tagging and syntactic bracketing. Clearly, the word is unique in a lot of respects and it has a different distribution from canonical verbs and prepositions. There is no strong evidence to support the position that Chinese has overt case markers, making it difficult to adopt the case marker analysis. After much discussion with linguistic experts, we decided to treat *ba* as a verb. To accommodate alternative analyses adopted by others (e.g., the Sinica Treebank), we set aside a unique POS tag *BA* for this word. In the bracketing guidelines we give detailed instructions as to how the construction should be annotated. As long as the construction is consistently annotated, the users of the treebank could easily map our analysis to the one that they prefer.

2.3 Addressing engineering issues

Building a treebank involves more than just addressing the linguistic issues. Since our goal is to create a sharable resource not just for our own consumption, but for the benefit of the entire Chinese NLP community, we also need to address three key engineering issues. These can be summarized as *quality*, *speed*, and *usefulness*. By *quality*, we simply mean *annotation accuracy* and *inter-annotator agreement*. To calculate annotation accuracy, we create a gold standard for a portion of the

Table 2. *The important factors in achieving our goal*

	Quality	Speed	Usefulness
Guidelines	√	√	√
NLP tools	√	√	
The team	√	√	√
Community involvement			√
Proper procedure	√		

treebank using double-blind annotation and adjudication, and evaluate an individual annotator's annotation against it. *Speed* is measured by the number of words (or characters) annotated per hour. The meaning of *usefulness* is three-fold: first, the potential users of the treebank may have different preferences for what the treebank should look like, depending on their applications and the particular algorithms used for the applications, among other things. When we design the treebank, we consider these preferences and try to accommodate them when possible. For instance, people who work on dependency parsers would prefer a treebank that contains dependency structures (like the Prague Dependency Treebank (Hajič 1998; Böhmová *et al.* 2003)), while others might prefer a phrase structure treebank (like the Penn English Treebank). Converting from phrase structure to dependency structure is more straightforward than going in the opposite direction (Xia and Palmer 2001), which was one of our motivations for choosing phrase structure. Our representation scheme also distinguishes arguments from adjuncts, which makes it easier to convert from phrase structure to LTAG (Xia, Han, Palmer and Joshi 2001), CCG (Hockenmaier 2003), or LFG (Cahill, McCarthy, Genabith and Way 2002). The second aspect of *usefulness* is related to linguistic analysis. It is common for people to disagree on the underlying linguistic theories and the particular analyses of certain linguistic phenomena in a treebank. A treebank should have rich annotation so that it is possible to convert the treebank into other annotation schemes. For instance, in our treebank we treat *ba* in the *ba-construction* as a verb, but we give it a special POS tag *BA* so that it is easy to convert our analysis into one where the word *ba* is treated as a preposition. The third aspect of *usefulness* is about the extendability of the treebank. In the future we will add not only more data from different genres, but also additional layers of annotations (such as predicate-argument structure and co-reference). The design of the treebank should facilitate such expansion.

It is obvious that the three objectives, *speed*, *quality* and *usefulness*, are sometimes in conflict and tradeoffs are needed. For instance, rich annotation will make a treebank more useful, but it will slow down the annotation process. In practice, we find the five factors in Table 2 to be crucial in achieving our goal. The importance of the first two factors (i.e., the guideline design and NLP tools) is well-known and we will discuss them in detail in sections 3 and 5. For the third factor, we find the background and the proper training of the annotators to be crucial in ensuring the high quality and good annotation speed. Besides the annotators, our team also includes several linguists and computational linguists: the linguists help us find plausible analyses for linguistic phenomena, whereas computational linguists are potential users of the treebank so they can provide feedback on using the treebank for training and testing various NLP tools. For the fourth factor, *community involvement*, in the first two years of the project alone, we held three meetings and two workshops, as shown in Figures 1 and 2 in the Appendix. The feedback from these meetings helped us better understand the needs and requirements from the Chinese NLP community. The last factor is proper *annotation procedure*. For the creation of CTB-I look words, we took the following steps:

1. Feasibility study: in this step, we identify major controversial topics and test whether consistent annotation is possible.
2. Creation of the first draft of the annotation guidelines based on comprehensive linguistic study: this step requires an extensive study of the literature and input from linguistics experts.
3. The first pass of the annotation: the cases that annotators find difficult to annotate often reveal problems in the guidelines. We revised the guidelines during the annotation process to fix these problems. Once the first pass was completed, the guidelines were effectively finalized.
4. The second pass of the annotation: in this step, one annotator corrects the output of the other annotator. Quality control is performed to monitor annotation accuracy and inter-annotator agreement.
5. Final cleanup of the data: we run tools to semi-automatically find annotation errors. After the cleanup, the guidelines are finalized and the data is ready to be released.

The two-pass process is necessary as the first draft of the guidelines is bound to miss certain phenomena. Annotators also need practice to become familiar with annotation tools, guidelines, and so on. We also found that spending sufficient time on the first draft of the guidelines greatly reduces the work of future revisions of both the guidelines and the treebank. Our experience is that very little revision of the guidelines was necessary in the development of CTB-II and beyond.

3 Designing guidelines

To build the Chinese treebank we need to create three sets of guidelines – segmentation, part-speech tagging and bracketing guidelines. Designing guidelines has been a crucial component of treebank development and is also the most difficult task in the treebank creation process. It is important because the guidelines specify the kind of information to be encoded in the annotated corpus and thus to a large extent determine the kind of knowledge that can be acquired by automatic systems that use the corpus. The guidelines are also an important tool for ensuring consistent annotation of the data. Real data are far more diverse and complicated than what is generally seen in the linguistics literature and devising an annotation scheme that provides linguistically plausible analyses for broad-coverage naturally occurring data is a huge challenge. We believe that guidelines should have the following properties:

- *Thoroughness*. As a key tool to ensure consistency of the annotation, the guidelines have to be specific and complete. This also means that, when we define the POS tagset and different linguistic structures, we should provide clean, solid diagnostic tests that are easy to follow. We also try to ensure that the guidelines cover all the possible structures that are likely to occur in the treebank so that the annotators do not need to come up with their own analyses. To achieve a level of generality that will easily extend to additional data, we do not limit our purview to just the data at hand. Instead, when we determine an analysis for a sentence in the corpus, we try to examine as many relevant examples as possible.

- *Theoretical neutrality*. Another desired goal is theoretical neutrality. Clearly we prefer that this corpus survives ever changing linguistic theories. While absolute theoretical neutrality is an unattainable goal, we approach this by building the corpus on the "safe" assumptions of theoretical frameworks and established theoretical constructs that have been proven to be solid whenever possible. While the influence of Government and Binding (GB) theory (Chomsky 1981) and X-bar theory (Jackendoff 1977) is obvious in our corpus, we do not adopt the whole package. Instead we try to identify and adopt those assumptions of GB that are the least controversial, such as the assumption that every phrase has a head that determines its categorical status. We did not consider the implications of case theory for Chinese. Case theory for a language like Chinese that lacks overt case markers is so subtle that at this point in time, we doubt our abilities to apply it to a non-trivial corpus on a consistent basis.
- *Convertability*. To make the treebank useful for the whole community, we study annotation guidelines and word segmentation standards used by other sites. While it is impossible to make our treebank compatible with all other standards, we design the guidelines in such a way that the treebank can be readily mapped to many of the alternative plausible analyses.

In this section, we discuss each set of guidelines in detail.

3.1 Word segmentation

As just mentioned, adequate consistency can be achieved only if well-designed guidelines are provided. In order to come up with linguistically justifiable specifications for word segmentation, it is necessary to have a clear understanding of what a word is or at least decide on a workable definition of a word. It has long been noted in the linguistic literature that phonological, morphological, syntactic and semantic criteria do not necessarily converge on a single notion of *word* in all cases. For example, Sciullo and Williams (1987) discuss four different notions of *word*; namely, morphological object, syntactic atom, phonological word and listeme. According to them, *syntactic atoms* are the primitives of syntax. In the X-bar theoretic framework, syntactic atoms are X^0. They are atomic in the sense that the syntactic rules cannot analyze their contents.[11] Specific to Chinese, Packard (2000) defines eight notions of

[11] Whether morphology and syntax are truly independent is still an open question (Sciullo and Williams 1987; Halle and Marantz to appear). We shall not go into details in this paper.

word; namely, orthographic word, sociological word, lexical word, semantic word, phonological word, morphological word, syntactic word, and psycholinguistic word. Because our goal was to build syntactic structures for sentences, we adopted the notion of *syntactic word* (or *syntactic atom*) for our word segmentation task. Based on this definition, we considered appropriate wordhood tests. Various wordhood tests, mostly heuristic, have been proposed in the Chinese linguistic literature and Dai (1992) provides an excellent summary[12]. The following are the most important ones. (We assume the string that we are trying to segment is X-Y, where X and Y are two morphemes. The morphemes are relatively easy to determine and most of the time they correspond to single Chinese characters.)

- Bound morpheme: a bound morpheme should be attached to its neighboring morpheme to form a word when possible.
- Productivity: if a rule that combines the expression X-Y does not apply generally (i.e., it is not productive), then X-Y is likely to be a word.
- Frequency of co-occurrence: if the expression X-Y occurs very often, it is likely to be a word.
- Complex internal structure: strings with complex internal structures should be segmented when possible.
- Compositionality: if the meaning of X-Y is not compositional, it is likely to be a word.
- Insertion: if another morpheme can be inserted between X and Y, then X-Y is unlikely to be a word.
- XP-substitution: if a morpheme cannot be replaced by a phrase of the same type, then it is likely to be part of a word.

All of these tests are very helpful. However, none of them is sufficient in itself for covering the entire range of difficult cases, most of which do not appear in dictionaries. Either the test is applicable only to limited cases (e.g. the XP-substitution test) or there is no objective way to perform the test as the test refers to vaguely defined properties (e.g., in the productive test, it is not clear where to draw the line between a *productive* rule and a *non-productive* rule). For more discussion on this topic, please refer to Sciullo and Williams (1987), Dai (1992), Packard (1998), Packard (2000) and Xue (2001).

Since no single test is sufficient, we adopted all of the above except for the productivity test and the frequency test. Rather than have the annotators memorize the entire set and make their own decisions using these diagnostic tests, in the guidelines we spell out what the results of applying the tests would be for all of the relevant phenomena. For example, for the treatment of verb-resultative compounds, we select the relevant tests, in this case the insertion test and the XP-substitution test, and give several examples of the results of applying these tests to verb-resultative compounds. This makes it straightforward, and thus efficient, for the annotators to follow the guidelines.

[12] Note that these tests are specific to Chinese and would not readily apply to English.

The guidelines are organized according to the internal structure of the corresponding expressions (e.g. a verb-resultative compound is represented as V+V, while a verb-object expression is as V+N), so it is easy for the annotators to search the guidelines for needed references. The differences between our segmentation standards and other well-known standards are small and generally have to do with granularity. For example, the PKU corpus treats a Chinese person name as having two segments, with the last name and the first name each being a segment. In contrast, the Penn Chinese Treebank would treat a person name as one segment. For example, "胡锦涛/Hu-Jintao" would be segmented as 胡锦涛 in our treebank but as 胡 锦涛 in the PKU corpus. There is little linguistic justification as to what the correct segmentation should be. The "right" segmentation can only be determined in the context of an actual application. For word compounds that have linguistically justifiable internal structures, our approach is to assign an hierarchical structure to them. For instance, for a resultative verb compound like "走/walk 过来/over", we treat it as having two segments, but in the bracketing phase we assign a label to the verb compound as a whole. This approach would allow the user to choose their desired level of granularity. For the complete segmentation guidelines and the comparisons between our guidelines and other well-known word segmentation standards (Liu, Tan and Shen 1993; Chinese Knowledge Information Processing Group 1996), please see Xia (2000b).

3.2 POS tagging

Since Chinese words are not marked with respect to tense, case and number, the central issue in POS tagging is whether the definitions of POS tags should be based on meaning or on syntactic distribution. This issue has been debated since the 1950s (Gong 1997) and there are still two opposing points of view. For example, a word such as 毁灭 in Chinese can be translated into *destroy/destroys/destroyed/destroying/destruction* in English and it is used in similar syntactic environments as its translations in English. One view holds that the part-of-speech for a word should be based solely on the meaning of this word. According to this view, a verb is equated with semantic notions such as action or activity and a noun generally describes an entity or state. Since the meaning of 毁灭 remains roughly the same across all of these usages, it should always be tagged as a verb, regardless of its syntactic environment. The opposite view says the part-of-speech of a word should be determined by its syntactic distribution. According to this view, when 毁灭 occurs as the head of a noun phrase, it should be tagged as a noun in that context. Similarly, when it serves as the head of a verb phrase, it should be tagged as a verb. In this way, the part-of-speech of a word encodes its syntactic function.

We adopted the second view for the following reasons. First, since our purpose is to annotate the syntactic structure of the sentences in the corpus, it makes sense for us to use the POS tags to encode syntactic information rather than semantic information. The syntactic categories of phrases in the bracketing phase, which we will describe in the next section, are direct projections of the part-of-speech information of the constituent words. If 毁灭 is annotated as a verb regardless of its syntactic function in a particular context, we will have a situation in which a verb is the head of a noun phrase, which is an implausible outcome. The second reason, which is perhaps more important, is that if the POS tags are assigned based on meaning regardless of the syntactic contexts, then the POS tagging information in a corpus is no more useful than a dictionary that simply lists the part-of-speech of each word. The advantage of a corpus over a dictionary should be that the POS tags in a corpus, just like phrasal categories, encode context.

A key part of POS guideline development is determining the POS tagset. There is not a "right" tagset or a "right" number of tags. In general, a larger tagset is more informative, but is more likely to cause annotation errors. It is a matter of finding the right tradeoff. A new tag is warranted if there exists a class of words that can be consistently tagged. In practice, different standards find different compromises. ASBC (Yu *et al.* 1998) uses 46 tags, while the PKU corpus uses 26 tags. Our POS tagset has 33 tags, as shown in Table 3.

3.3 Syntactic bracketing

The most important decision to make in recent treebank development efforts has been whether to annotate phrase structure, following the Penn English Treebank, or dependency relations as in the Prague Dependency Treebank (Böhmová *et al.* 2003). In a phrase structure formalism, the most important grammatical relations are between linearly adjacent constituents. The relative positions of the constituents are meaningful and long-distance (or non-adjacent) dependencies are captured with traces and indices. This kind of formalism is suitable for languages like English where the word order is rigid and long-distance dependencies occur in a constrained manner. For a language like Czech, which has free word order, it is difficult to capture the grammatical relations using a representation scheme in which relative position plays an important role. Thus, dependency treebanks such as the Prague Dependency Treebank (Böhmová *et al.* 2003), allow *non-projectivity* or crossing edges, in which dependency relations may exist between phrases that are not linearly adjacent. Since in Chinese, despite its many differences from Western languages like English, rigid word order is the rule rather than the exception, we chose to annotate phrase structure rather than dependency structure. Even for phrase structure annotation, there are still a variety of formalisms. However, the decision here may be less crucial than often assumed. Recent work has also shown that one grammatical formalism can be algorithmically converted to another if proper distinctions are consistently made. Hockenmaier (2003) shows that the main difficulty in converting the Penn English Treebank to a CCG formalism (Steedman 1996, 2000) is the lack of argument/adjunct distinctions in the Penn English Treebank and the flat structure in the noun phrases. This means that as long as the grammatical relations are captured and the important distinctions are made, it is possible for the user to convert the treebank to an alternative desired representation scheme. We accepted the challenge of providing a full phrase structure analysis that would facilitate semantic annotation, and would allow us to apply established tools and techniques. We adopted the Penn English Treebank methodology, with some extensions. In

Table 3. *Our POS tagset in alphabetical order*

Tag	Description	Example
AD	adverb	还
AS	aspect marker	着
BA	把 in ba-construction	把, 将
CC	coordinating conjunction	和
CD	cardinal number	一百
CS	subordinating conjunction	虽然
DEC	的 in a relative-clause	的
DEG	associative 的	的
DER	得 in V-de const. and V-de-R	得
DEV	地 before VP	地
DT	determiner	这
ETC	for words 等, 等等	等, 等等
FW	foreign words	ＩＳＯ
IJ	interjection	啊
JJ	other noun-modifier	男, 共同
LB	被 in long bei-const	被, 给
LC	localizer	里
M	measure word	个
MSP	other particle	所
NN	common noun	书
NR	proper noun	美国
NT	temporal noun	今天
OD	ordinal number	第一
ON	onomatopoeia	哈哈, 哗哗
P	preposition excl. 被 and 把	从
PN	pronoun	他
PU	punctuation	、？。
SB	被 in short bei-const	被, 给
SP	sentence-final particle	吗
VA	predicative adjective	红
VC	是	是
VE	有 as the main verb	有
VV	other verb	走

addition to making explicit argument/adjunct distinctions, we also marked dropped subjects and added some NP-internal structures, all of which are detailed below.

Our representation scheme for bracketing is a combination of an hierarchical organization of constituent structures and functional tags. Following the Penn English Treebank II (Marcus *et al.* 1994), we use four types of notational devices: labeled brackets, functional tags, null elements and indices. We will explain each of these below.

Labeled brackets

As we have briefly mentioned in previous sections, the bracketing phase of this project focuses on the syntactic relationships between constituents. In our guidelines,

Table 4. *A schematic representation of the three major grammatical relations*

Complementation			
head-initial	*head-final*	*Adjunction*	*Coordination*
(XP X (YP) (ZP) ...)	(XP (YP) (ZP) ... X)	(XP (YP) ... (XP) ... (ZP))	(XP {CONJ} (XP) {CONJ} (XP) ...)

we select three grammatical relations as the most basic: *complementation*, *adjunction* and *coordination*. We enforce the rule that one labeled pair of brackets represents only one grammatical relation. Each of these three grammatical relations is assigned a unique hierarchical structure, as is illustrated in Table 4. *X* and *CONJ* are terminal nodes while *XP*, *YP* and *ZP* are non-terminal nodes.

Example 3 illustrates the three major structural relations. The noun phrase "法规性/policy 文件/document" is a *complement* to the verb "推出/promulgate", and this is represented by attaching the non-terminal NP node at the same level as the terminal VV node. The *coordination* relation is illustrated by the relation between the two *VP*s "制定/legislate" and "推出/promulgate 法规性/policy 文件/document", conjoined by "和/and". The adverbial phrase "积极/actively 及时/timely 地/DE" is *adjoined* to the verb phrase "制定/legislate 和/and 推出/promulgate 法规性/policy 文件/document". Each pair of brackets has one label and this label indicates the grammatical category of the constituent. Table 5 lists the seventeen labels used in our bracketing guidelines. The hierarchical organization of the constituents is represented by the relative position of these constituents within a phrase.

```
(IP (NP-PN-SBJ (NR 浦东/Pudong))
    (VP (DVP (ADVP (AD 积极/actively)
                   (PU 、)
                   (AD 及时/timely))
             (DEV 地/DE))
        (VP (VP (VV 制定/legislate)
                (NP-OBJ (-NONE- *RNR*-1)))
            (CC 和/and)
            (VP (VV 推出/promulgate)
                (NP-OBJ-1 (NN 法规性/policy)
                          (NN 文件/document))))))
```

Pudong legislated and promulgated policy documents actively and in a timely manner.

Example 3. The three structural relations.

Table 5. *Tags for syntactic phrases*

ADJP	adjective phrase
ADVP	adverbial phrase headed by AD (adverb)
CLP	classifier phrase
CP	clause headed by C (complementizer)
DNP	phrase formed by "XP + DEG"
DP	determiner phrase
DVP	phrase formed by "XP + DEV"
FRAG	fragment
IP	simple clause headed by I (INFL)
LCP	phrase formed by "XP + LC"
LST	list marker
NP	noun phrase
PP	preposition phrase
PRN	parenthetical
QP	quantifier phrase
UCP	unidentical coordination phrase
VP	verb phrase

```
(IP (NP-PN-TPC (NR 海尔/Haier)
               (NN 集团/group))
    (NP-TMP (NT 九十年代/1990s))
    (PP-LOC (P 在/in)
            (NP (NN 国内外/inside and outside the country)))
    (NP-SBJ (NN 知名度/name recognition))
    (VP (ADVP (AD 很/very))
        (VP (VA 高/high)))
    (PU 。))
```

Haier Group was well-known inside and outside the country during the 1990s.

Example 4. The use of functional tags.

Functional Tags

Besides the hierarchical representations, functional tags are used to provide further information. These functional tags can be regarded as secondary and are used to complement hierarchical representations. For example, in Chinese, multiple noun phrases (labeled *NP* in the Penn Chinese Treebank) can occur before the verb within a clause (or above the verb if seen hierarchically). Structurally, they are all above the verb even though it is obvious that they have different syntactic functions. Therefore, they are further differentiated by functional tags. Example 4 illustrates the use of functional tags. Functional tags are used to annotate the different grammatical roles

Table 6. *Functional tags and null categories used in CTB*

Functional tags				*Null Categories*	
ADV	adverbial	MNR	manner	*pro*	dropped argument
APP	appositive	OBJ	direct object	*PRO*	used in non-finite constructions
BNF	beneficiary	PN	proper noun		
CND	condition	PRD	predicate	*T*	trace of A'-movement
DIR	direction	PRP	purpose or reason	*	trace of A-movement
EXT	extent	Q	question	*RNR*	right node raising
FOC	focus	SBJ	subject	*OP*	operator
HLN	headline	SHORT	short form	*?*	other unknown empty categories
IJ	interjective	TMP	temporal		
IMP	imperative	TPC	topic		
IO	indirect object	TTL	title		
LGS	logical subject	VOC	vocative		
LOC	locative	WH	wh-phrase		

such as topic (*TPC*) and subject (*SBJ*).[13] In addition, functional tags are also used to distinguish different types of adverbial elements such as location (*LOC*) and temporal phrases (*TMP*).

Null categories

In addition to labeled brackets and functional tags, we also use special symbols to annotate the phonologically null elements. The seven null categories and their uses are listed in Table 6.

Some of the null categories are coindexed with lexical material within the sentence. The use of null categories and indices is illustrated in Example 5. The sentence does not have an overt subject and thus it is marked with *pro*. Since it is not co-referential with any lexical material within the sentence, it does not receive an index. In contrast, the topicalized noun phrase "各/every 种/kind 收费/fee", labeled "NP-TPC-2", is co-referential. It is moved from the object position and therefore an empty category coindexed with the topic, "(NP-OBJ (-NONE- *T*-2))", is posited in the object position.

Our representational scheme allows the identification of basic grammatical functions such as subjects, objects and adjuncts in the corpus, which can be used to train and test syntactic parsers. A detailed description of our representation scheme is available in (Xue and Xia 2000).

[13] A subject is an argument of a verb, but it is not a complement of a verb. It is treated as if it is an adjunct because according to GB theory the subject is moved from its base position to the [Spec, IP] position, a movement that is not marked in our treebank. The [Spec, IP] position, unlike the position of the object, is not a sibling of the verb in the phrase structure.

```
(IP (NP-TPC-2 (DP (DT 各/every)
                  (CLP (M 种/kind)))
              (NP (NN 收费/fee)))
    (NP-SBJ (-NONE- *pro*))
    (VP (ADVP (AD 已/already))
        (PP-TMP (P 在/in)
                (LCP (NP (QP (CD 一/one)
                             (CLP (M 个/CL)))
                         (NP (NN 月/month)))
                     (LC 前/before))
        (VP (VV 宣布/publish)
            (NP-OBJ (-NONE- *T*-2)))))
    (PU 。))

All kinds of fees were already published one month ago.
```

Example 5. The use of null categories and indices.

4 Comparison with other Treebanks

In this section we briefly compare the design of the Penn Chinese Treebank with the Penn English Treebank (Marcus *et al.* 1994) and the Sinica Chinese Treebank (Chen *et al.* 2004). The Penn English Treebank serves as a good reference for comparison because the two corpora use very similar notational devices. However, as demonstrated below, we also made several significant enhancements to make it easier to extract predicate-argument structure. The Sinica Chinese Treebank, on the other hand, is currently the only other Chinese treebank that is publicly available. In fact, its first version was released at about the same time as the release of CTB-I.

4.1 Comparison with the Penn English Treebank

When we designed the Chinese Treebank, we started with the annotation style adopted by the Penn English Treebank. However, our representation scheme differs from the one used in the Penn English Treebank in one important aspect. Our decision to enforce the rule that *one pair of labeled brackets only represents one hierarchical grammatical relation* leads to the annotation of more structures. Example 6 shows how the English Treebank allows heterogeneous relations between the constituents of a phrase. For example, the *PP* "on Aug. 1, 1988" is an adjunct to the *VP* while the *NP* "a dividend" is a complement, but both of the them are attached at the same level.

In the Chinese Treebank, the requirement that one pair of labeled brackets represents one structural grammatical relation forces constituents with different

```
(S (NP-SBJ (DT The)
           (NN mortgage)
           (CC and)
           (NN equity)
           (JJ real)
           (NN estate)
           (NN investment)
           (NN trust))
   (ADVP (RB last))
   (VP (VBD paid)
       (NP (DT a)
           (NN dividend))
       (PP-TMP (IN on)
               (NP (NP (NNP Aug.)
                       (CD 1))
                   (, ,)
                   (NP (CD 1988)))))
   (. .))
```

Example 6. Complements and adjuncts are attached at the same level in the Penn English Treebank II.

grammatical relations to be attached at different levels. For the VP-internal structure, this means that the complements and the adjuncts are annotated with different structural configurations. Example 7 shows how the adjunct "首/first 次/time" and the complement "二十亿/two billion 元 /Yuan 大关/milestone" are attached at different VP nodes. We believe this revised design will make it easier to provide predicate-argument structure labeling, as is currently being done for the Penn English Treebank (Palmer, Gildea and Kingsbury to appear).

For NP-internal structure, this means that the coordinated constituents and their shared modifiers are attached at different levels, among other things. In Example 8, the modifier "重庆/Chongqing 的/DE" and the coordinated *NP* "经济/economics 和/and 政治/politics" are attached at different NP nodes. This is in contrast with the English example in Example 6, where the coordinated "mortgage and equity" are attached at the same level as other modifiers of the head noun "trust".

It must be pointed out that the complement/adjunct distinction can only be made when the facts of the language warrant it. One of the reasons why the argument/adjunct distinction is not made in the Penn English Treebank is that it was decided that this distinction cannot be made consistently. However, in Chinese it is generally agreed that this distinction can be made (Huang 1982).

From our experience in creating the Chinese Treebank and our knowledge about the Korean Penn Treebank (Han, Han, Ko and Palmer 2002), and the Penn Arabic

```
(IP (NP-TMP (NT 去年/last year))
    (PU ，)
    (NP-SBJ (NP (DP (DT 全/whole))
                (NP (NN 区/district)))
            (DP (DT 各/every)
                (CLP (M 项/item)))
            (NP (NN 存款/deposit)))
    (VP (QP-ADV (OD 首/first)
                (CLP (M 次/time)))
        (VP (VV 突破/exceed)
            (AS 了/ASP)
            (NP-OBJ (QP (CD 二十亿/two billion)
                        (CLP (M 元/Yuan)))
                    (NP (NN 大关/milestone))))))
```

Last year, the district's deposit of all categories exceeded the level of two billion Yuan for the first time.

Example 7. Complements and adjuncts are attached at different levels in the Penn Chinese Treebank.

```
(NP (DNP (NP-PN (NR 重庆/Chongqing))
         (DEG 的/DE))
    (NP (NN 经济/economy)
        (CC 和/and)
        (NN 政治/politics)))
```

Chongqing's economy and politics.

Example 8. Coordinated structure is attached at a separate level.

Treebank[14], we conclude that the general annotation style used by the Penn English Treebank works very well for other languages. Language-dependent items include the linguistic phenomena that have to be studied and included in the guidelines, the diagnostic tests for resolving certain difficult cases, and some of the tools for annotating and preprocessing the data.

4.2 Comparison with the Sinica Treebank

The Sinica Chinese Treebank (Chen *et al.* 2004) was created by researchers at Academia Sinica. Its first release has 239,532 words. Compared with the Penn Chinese

[14] More information can be found at `http://www.ircs.upenn.edu/arabic`

Treebank's clearly syntactic orientation, the Sinica Treebank primarily focuses on annotating semantic structure. For instance, its tagset contains 54 thematic role labels for the arguments of verbs, 12 for nominalized verbs and six for nouns. While the function tags in our treebank cover most of the important relations such as *subject*, *object* and *topic*, our tagset is not as rich as theirs.

On the other hand, the Sinica Treebank only annotates minimal syntactic structures, using just 6 phrasal categories (S = a complete tree headed by a predicate, VP = a phrasal headed by a predicate, NP = a phrase headed by a noun, GP = a phrase headed by locational noun, PP = a phrase headed by a preposition, XP = a conjunctive phrase), compared with the 17 phrasal labels used in the Penn Chinese Treebank. The average sentence length in their Treebank is 7.6 words, compared to 28.9 words in CTB-II. The reason that the *sentences* in their treebank are so short is because this treebank uses commas as sentence delimiters along with a few other punctuation marks. Furthermore, the Sinica Treebank does not mark phonologically null elements such as traces and it does not use indices to mark co-referentiality. The two treebanks also differ in word segmentation standards, POS tagsets, and linguistic analyses for many syntactic constructions. A full evaluation of the relative merits and weaknesses of the two Chinese treebanks is difficult, and it is very likely that users' preferences for one treebank over the other largely depends on intended NLP tasks, their particular algorithms, and other factors. We believe that both treebanks are valuable resources and will help to advance the Chinese NLP field.

5 Treebank engineering issues

In general, our annotation can be described as a semi-automatic process: the data were first automatically processed with NLP tools and then manually checked. In the segmentation/part-of-speech tagging phase, the data were first segmented with a word segmenter and then tagged with a POS tagger. The outputs of the segmenter and tagger were then manually corrected by our annotators. In the bracketing phase, the segmented and part-of-speech tagged data were first parsed by a parser and then the output was manually corrected by our annotators.

This is not entirely the case in the development of CTB-I because at that time, we did not have a parser, and we had only an integrated segmenter/part-of-speech tagger trained on a different dataset. As a result, the bracketing was done manually from scratch. As CTB-I data became available, we trained our own segmenter (Xue and Converse 2002), an off-the-shelf part-of-speech tagger (Ratnaparkhi 1996) and statistical parsers (Chiang 2000; Bikel and Chiang 2000). The fact that these tools are all implemented with machine-learning algorithms makes it possible for us to train increasingly more accurate tools as more data become available. The tools in turn help us improve our annotation throughput, in a bootstrapping cycle that helps both the tools and the annotation.

In this section, we first describe how we trained tools to speed up our annotation. Since bracketing is known to be a difficult task (Marcus *et al.* 1994), we implemented procedures to ensure high accuracy and inter-annotator agreement in the bracketing phase. We will describe our quality control procedure and report results of our accuracy and consistency evaluation.

Table 7. *One character can occur in as many as four different positions*

position	tag	example
Left	LL	产生 'to come up with'
Word by itself	LR	产 小麦 'to grow wheat'
Middle	MM	生产线 'assembly line'
Right	RR	生产 'to produce'

5.1 Speeding up annotation with automatic tools

When we built CTB-I, we did not have our own set of tools (segmenters, pos-taggers and parsers) for preprocessing. We used an integrated stochastic segmenter and part-of-speech tagger provided by BBN, which was trained on the Academia Sinica Balanced Corpus (ASBC). Since the ASBC and our treebank use different tagsets, we mapped the ASBC tags to our tags automatically. Although the mapping was not one-to-one and introduced some errors, this process greatly accelerated annotation. For bracketing, we did not have a parser so the bracketing was done from scratch.

Upon the completion of CTB-I, we were able to train a new set of NLP tools using our own data. These tools, when used as preprocessors, substantially accelerated our annotation.

5.1.1 A machine learning approach to segmentation

Using the data from the Penn Chinese Treebank, we trained a statistical segmenter, using a maximum entropy approach (Ratnaparkhi 1998). In training our maximum entropy segmenter, we reformulated the segmentation problem as a tagging problem. Specifically we tagged characters as LL (left), RR (right), MM (middle) and LR (single-character word), based on their positions within words. The example in Table 7 shows that a Chinese character 产 can have multiple tags if it occurs in different positions within different words. Similarly, a sentence can be assigned different tag sequences if there is an ambiguity in the segmentation, as the example (adapted from Sproat *et al.* 1996) in Example 9 shows.

The segmentation task is to resolve these ambiguities and find the correct tag sequence that yields the correct interpretation. The ambiguities are resolved by examining the context in which the character occurs and using features to encode this context. It should be pointed out that sometimes the ambiguity cannot be completely resolved just by looking at neighboring words and a larger context is needed (Gan 1995; Xue 2001). As a preliminary step, in this experiment our features only use the information in the local context, which includes the current character, the previous two and the next two characters, and the previous tag.

Training data can be trivially derived from a manually segmented corpus, as is illustrated in Example 9. Using 80,000 words from CTB-I as training data and the remaining 20,000 words as test data, the maximum entropy segmenter achieved an accuracy of 91% (F-measure). When the second installment of the Penn

(a) Segmentation I
日文 章鱼 怎么 说？
Japanese octopus how say
"How to say octopus in Japanese?"

(b) Segmentation II
日 文章 鱼 怎么 说？
Japan article fish how say

(c) The tag sequence for segmentation I
日/LL 文/RR 章/LL 鱼/RR 怎/LL 么/RR 说/LR ？

(d) The tag sequence for segmentation II
日/LR 文/LL 章/RR 鱼/LR 怎/LL 么/RR 说/LR ？

Example 9. Ambiguous segmentation for the same sentence.

Chinese Treebank (CTB-II) was released, we optimized the features and retrained this segmenter on roughly 238,000 words and tested it on the remaining 13,000 words. The accuracy improved to 94.89% (F-measure). For details of this segmenter the reader is referred to Xue and Converse (2002).

5.1.2 Training a maximum entropy POS tagger for Chinese

Unlike segmenters, a POS tagger is a standard tool for the processing of Indo-European languages where words are trivially identified by white spaces in text form. Once Chinese sentences are segmented into words, Chinese POS taggers can be trained in a similar fashion as POS taggers for English. The contexts that are used to predict the POS tags are roughly the same in both Chinese and English. These are the surrounding words, the previous tags and word components. One notable difference is that Chinese has fewer affixes than Western languages, and affixes are generally good predictors for the part-of-speech of a word. Nevertheless, some Chinese characters, even though they are not affixes, are still good predictors for the part-of-speech of the words they are components of. Another difference is that words in Chinese are much shorter than words in English when we count the number of characters in a word.

Our POS tagger is essentially the maximum entropy tagger developed by Ratnaparkhi (1996) which has been retrained on our Chinese data. We used the same 80,000-word chunk that was used to train the segmenter, with the remaining 20,000 words for testing. Our results show that the accuracy of this tagger is about 93%. When CTB-II became available, we retrained the tagger on 238,000 words and tested it on the remaining 13,000 words. The tagging accuracy improved to 94.47%.[15]

[15] In both experiments, the input sentences are already segmented into words according to the treebank.

In contrast, the baselines for the two experiments are 86% and 88%, respectively, if we simply tag each known word with its most frequent tag, and tag unknown words with the tag *NN*. Considering that our corpus is still relatively small, the performance of our tagger is very promising. We expect that better accuracy will be achieved as more data become available.

The availability of the Chinese segmenter and tagger speeds up the annotation, and at the same time as more data are annotated we are able to train more accurate preprocessing tools. The value of preprocessing in segmentation and POS tagging is self-evident and these automatic tools turn annotation into a much easier error-correction activity rather than annotation from scratch. On average, the speed of correcting the output of a segmenter and a POS-tagger is about 2500 words per hour, nearly twice as fast as annotating the same data from scratch.

The value of a parser as a preprocessing tool is less obvious, because the errors made by a parser are not as local as the errors made by a segmenter or a POS tagger, and an annotator has to do considerable backtracking to undo some of the incorrect analyses produced by the parser. Our experimental results show that even with the apparent drawback of having to backtrack from the parses produced by the parser, the parser is still a useful preprocessing tool that helps annotation substantially. We will discuss our experiment next.

5.1.3 Training a statistical parser

To determine the usefulness of the parser as a preprocessing tool, we used a statistical parser (Chiang 2000) based on Tree-Insertion Grammar. The parser used 80,000 words of fully bracketed data for training and 10,000 words for testing and obtained 73.9% labeled precision and 72.2% labeled recall.[16] We then conducted an experiment to determine whether the use of a parser as a preprocessor improves annotation speed. We randomly selected a 13,469-word chunk of data from the corpus. The data was blindly divided into two portions of equal size (6731 words for portion 1, and 6738 words for portion 2). The first portion was annotated from scratch. The second portion was first preprocessed by this parser and then an annotator corrected its output. The throughput rate was carefully recorded. It took the same annotator 28.0 hours to finish the first portion, and 16.4 hours to finish the second portion. In other words, despite the need of backtracking, using the parser as a preprocessor increases the annotation speed from 240 words/hour to 410 words/hour.[17] For more details about this experiment, please see (Chiou, Chiang and Palmer 2001).

5.2 Quality control

A major challenge in providing syntactic annotation of corpora is ensuring consistency and accuracy. There are many factors that affect annotation consistency and

[16] Precision is the number of correctly bracketed constituents divided by the total number of constituents in the parse output and recall is the number of correctly bracketed constituents divided by the total number of constituents in the gold standard.

[17] The character/word ratio in the CTB is 1.7 character/word. This amounts to an increase from 408 characters/hour to 697 characters/hour.

accuracy: guideline design, annotator background and tools for annotation support. We have already described our guideline design in section 3. A detailed discussion about the importance of linguistic training in the annotation of this corpus can be found in (Xue, Chiou and Palmer, 2002). The bracketing interface which was originally created for the Penn English Treebank and later ported to Chinese for our use also proved to be invaluable in facilitating annotation and preventing certain types of annotator errors. In this section, we describe an annotation procedure we implemented to ensure consistency and accuracy, and report the evaluation results. We also briefly describe two tools that we used for the final cleanup.

5.2.1 Double-blind annotation and evaluation with Parseval

To monitor our annotation accuracy and inter-annotator agreement, we randomly select 20 percent of the treebank data for double-blind annotation. That is, for these data, each annotator annotates them independently. The annotators meet weekly to compare the doubly annotated data. This is done in three steps: first, we calculate inter-annotator agreement by running the Parseval software[18] that produces three metrics – precision, recall and numbers of crossing brackets (Black *et al.* 1991). Second, the annotators examine the inconsistencies detected by the evaluation tool, and decide on the correct annotation. In most cases, the inconsistencies are caused by some obvious mistakes, and the annotators can agree on the correct annotation. In rare occasions, the inconsistencies are due to a misinterpretation of the guidelines or the lack of clear specification in the guidelines, in which case the guidelines will be revised. This was especially true in the development of CTB-I. The comparison provides an opportunity both for the continued training of the annotators and for identifying gaps or inconsistencies in the guidelines. After the inconsistencies are corrected or adjudicated, the corrected and adjudicated data are designated as the Gold Standard. The final step is to compare the Gold Standard with each annotator's annotation and determine each annotator's accuracy. Our results show that both measures for CTB-I are well above 90%. During the bracketing phase of CTB-II, the average annotator accuracy was 96.7% (F-score) and the average inter-annotator consistency was 93.8% (F-score).

5.2.2 Post-annotation checking with automatic tools

As a final quality control step, we ran two tools. The first one is called *LexTract* (Xia *et al.* 2001). Given a treebank, LexTract automatically extracts a grammar – a grammar can be either a lexicalized tree-adjoining grammar (LTAG) or a context-free grammar (CFG). There is a none-to-one mapping from the nodes in treebank trees to the nodes in the extracted "rules" in the grammar.[19] Annotation errors in a treebank very often result in linguistically implausible rules in the grammar. Once

[18] The tool was developed by Satoshi Sekine and Michael Collins (`www.cs.nyu.edu/cs/projects/proteus/evalb`).

[19] A "rule" in an LTAG is a tree called *an elementary tree*, and a rule in a CFG is a context-free rule.

we identify such implausible rules, we can trace the errors back to the treebank trees that generate these rules, and make the necessary corrections.

We used LexTract for the final cleanup of CTB-I. Before running LexTract, the trees in the Treebank had been manually checked at least twice and the annotation accuracy was above 95%. It took one person about 10 hours to manually identify implausible rules in the extracted grammar built by LexTract, and another 20 hours to correct the treebank trees that generated these implausible rules. In total, we found about 580 annotation errors in the CTB-I. The error types included unbalanced brackets, illegal tags, wrong or incompatible syntactic labels, wrong or missing function tags, missing tree nodes, extra tree nodes, and so on. This process improves both the treebank and the extracted grammar. For details about LexTract and its application to treebank error detection, see Xia (2001).

The second tool for the final cleanup is called *CorpusSearch*, which was developed by Beth Randall.[20] This tool is similar to a tree search tool called *tgrep*, which was used by the Penn English Treebank project.[21] Given a set of patterns for common annotation errors, CorpusSearch can quickly find all the treebank trees that match those patterns. The annotators will then manually fix the errors in the trees. Once we have applied LexTract and CorpusSearch to our treebank and fixed the errors found by these tools, the treebank is ready to be released.

6 Conclusion and future work

The goal of the Penn Chinese Treebank Project is to create a large-scale high-quality treebank in a reasonable amount of time, which is useful for the entire Chinese NLP community. So far we have completed CTB-I and CTB-II, 250,000 words in total. We are planning another release in the fall of 2004, which will include an additional 250,000 words of treebanked Sinorama data with a corresponding parallel English corpus. For CTB-II, the average annotator accuracy was 96.7% and the average inter-annotator agreement was 93.8%. The high accuracy and inter-annotator consistency demonstrated the feasibility of providing full phrase structure analysis for Chinese. With the help of NLP tools (word segmenters and POS taggers, and statistical parsers), our annotators can segment and POS tag about 2500 words per hour. The speed of bracketing approaches 410 words (697 characters) per hour in first pass annotation. The treebank has been used for various NLP tasks such as word segmentation, POS tagging, parsing, and grammar extraction. The treebank data were also used in the first SIGHAN Word Segmentation Bakeoff (Sproat and Emerson 2003).

In addition to the 500,000 word corpus being released to the public, we also developed a set of tools that assist in various aspects of annotation. Researchers may also find the three sets of annotation guidelines valuable as they cover a

[20] The details of this tool can be found at `www.ling.upenn.edu/dringe/CorpStuff/Manual/Contents.html`.

[21] Because the source code for *tgrep* is not available, we cannot modify the tool to handle Chinese.

wide range of topics in Chinese morphology and syntax. Moreover, the creation of our Treebank demonstrates the feasibility of applying the Penn English Treebank annotation scheme to other languages. Furthermore, several meetings and workshops in the first two years of the project created strong ties among researchers in the field, which led to the formation of SIGHAN, a special interest group of ACL. Since its formation in 2001, SIGHAN has sponsored two workshops and one word segmentation contest.

The Penn Chinese Treebank provides a solid basis upon which other linguistic information and relationships can be represented and annotated. At the discourse level, for example, the annotation of the syntactic constituents makes it possible to represent inter-sentential co-referentiality between constituents. One subtask of this kind of annotation, mapping the anaphorical expressions to their antecedents, is already under way (Converse 2002). The treebank also makes it possible to make generalizations across different instances of the same predicate in the form of predicate-argument structures. For example, in "这/this 条/CL 法案/bill 通过/pass 了/AS" and "议 会/Congress 通过/pass 了/AS 这/this 条/CL 法案/bill", "这/this 条/CL 法案/bill" plays the same role with regard to the verb "通过/pass" even though it occurs in different syntactic positions (subject and object respectively). The availability of the constituents makes it possible to capture this regularity by assigning the same argument label (say *arg1* of 通过) to "这/this 条 /CL 法案/bill" in both sentences. We are currently adding such labels to the treebank in an effort to build a Chinese Proposition Bank, similar to the English Proposition Bank (Palmer, Gildea and Kingsbury to appear). The details of the Chinese Proposition Bank can be found in Xue and Palmer (2003). We are also beginning to sense-tag words in the Chinese Treebank (Dang, Chia, Palmer and Chiou 2002).

Acknowledgements

We gratefully acknowledge the invaluable assistance of Ralph Weischedel at BBN in providing the word segmenter and POS tagger used for the initial data preprocessing. We thank Zhibiao Wu for porting our treebanking and Parseval tools to Chinese and assisting with encoding issues, and Ann Taylor and Mitch Marcus for extremely pertinent treebanking advice. Special thanks to Meiya Chang and Tsạn-Kuang Lee for annotation. We thank LDC for making the Chinese Treebank available to the public. We also thank Andi Wu, Zixin Jiang and Changning Huang from Microsoft, Jin Yang from Systran, Shiwen Yu and Hui Wang from Peking University, DeKai Wu from Hong Kong Science and Technology University, Chu-Fen Huang and Keh-Jiann Chen from Academia Sinica, and our linguistics experts including Shizhe Huang, Tony Kroch, Peter Cole, James Huang and Aravind Joshi. Finally, we thank all of the participants at our meetings and workshops, and all the people who use our treebank. This research was funded by DOD MDA902-97-C-0307 and DARPA N66001-00-1-8915.

Appendix: Project inception and timeline

Our first step in assessing community interest in a standard reference corpus for Chinese was a three-day workshop on issues in Chinese language processing which

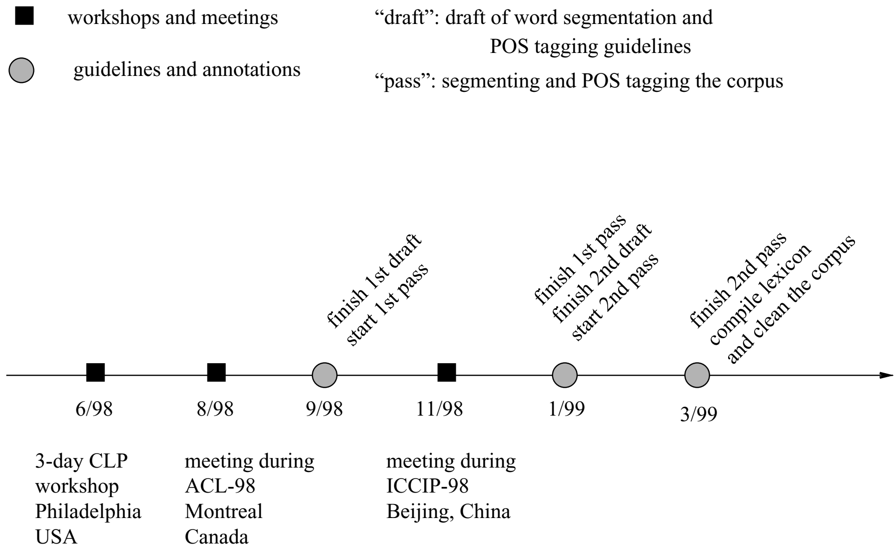

Fig. 1. The first phase of CTB-I: segmentation and POS tagging.

was held at the University of Pennsylvania. The aim of this workshop was to bring together influential researchers from Taiwan, Singapore, Hong Kong, China and the United States in a move towards consensus building with respect to word segmentation, part-of-speech (POS) tagging, syntactic bracketing and other areas. The American groups included the Institute for Research in Cognitive Science and the Linguistics Data Consortium (which distributes the English Treebank) at the University of Pennsylvania, the University of Maryland, Queens College, the University of Kansas, the University of Delaware, Johns Hopkins University, Systran, BBN, AT&T, Xerox, West Group, Unisys and the US Department of Defense. We also invited representatives of Academia Sinica in Taiwan and Hong Kong Science and Technology University. The workshop included presentations of guidelines being used in mainland China, Taiwan, and Hong Kong, as well as talks on existing Chinese segmenters and part-of-speech taggers. We divided participants into several working groups to discuss specific issues in segmentation, POS tagging and the syntactic annotation of newswire text.

There was general consensus at this workshop that a large-scale effort to create a Chinese treebank would be well received, and that linguistics expertise was a necessary prerequisite to successful completion of such a project. The workshop made considerable progress in defining criteria for segmentation guidelines as well as addressing the issues of part-of-speech tagging and syntactic bracketing. The Penn Chinese Treebank project began shortly after the workshop was held.[22] Since then, we have organized two workshops on Chinese language processing and held

[22] See our Penn Chinese Treebank website (www.cis.upenn.edu/chinese) for guidelines and publications as well as sample files.

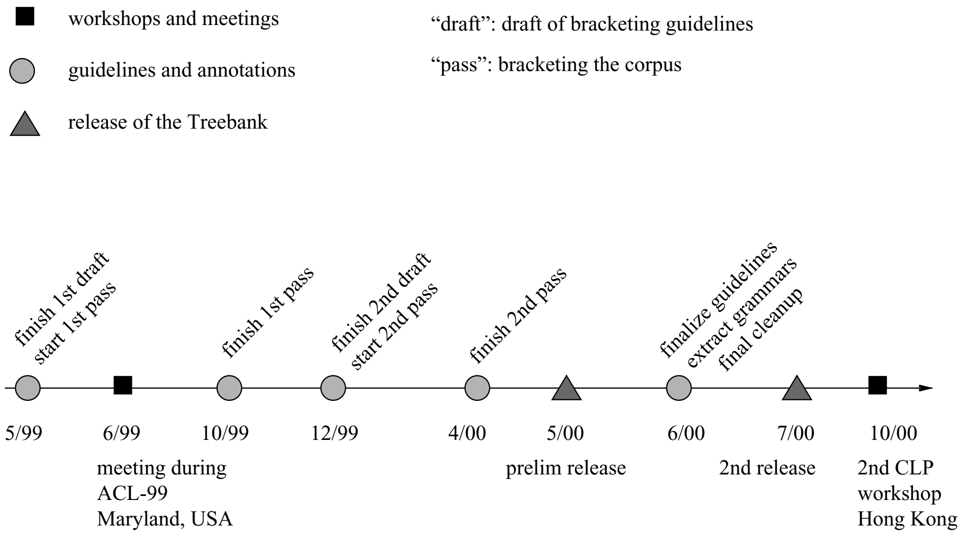

Fig. 2. The second phase of CTB-I: bracketing and data release.

several meetings in the USA and abroad to give updates and get feedback on our treebank.

CTB-I was developed almost concurrently with the creation of the three sets of annotation guidelines. This initial stage of the project was finished within about two years. Figures 1 and 2 summarize the milestones of this stage of the project.

References

Abeillé, A. (2003) *Building and Using Syntactically Annotated Corpora.* Kluwer.

Böhmová, A., Hajič, J., Hajicová, E. and Hladká, B. (2003) The Prague Dependency Treebank: A three level annotation scenario. In: Abeillé, A., editor, *Treebanks: Building and Using Annotated Corpora.* Kluwer.

Bender, E. (2000) The Syntax of Mandarin -basb. *J. East Asian Linguistics*, **9**(2).

Bikel, D. M. and Chiang, D. (2000) Two atatistical parsing models applied to the Chinese Treebank. *Proceedings of the 2nd Chinese Language Processing Workshop*, Hong Kong, China.

Black, E., Abney, S., Flickinger, D., Gdaniec, C. *et al.* (1991) A procedure for quantitatively comparing the syntactic coverage of English grammars. *Proc. DARPA Speech and Natural Language Workshop.*

Cahill, A., McCarthy, M., Van Genabith, J. and Way, A. (2002) Automatic annotation of the Penn Treebank with LFG-structure information. *Proceedings LREC 2002 Workshop on Linguistic Knowledge Representation – Bootstrapping Annotated Language Data*, pp. 8–15.

Charniak, E. (2000) A Maximum-entropy-inspired parser. *Proc. NAACL-2000.*

Chen, K.-J., Huang, C.-R., Chen, F.-Y., Luo, C.-C., Chang, M.-C. and Chen, C.-J. (2004) Sinica Treebank: Design criteria, representational issues and implementation. In: Abeillé, A., editor, *Building and Using Parsed Corpora.* Kluwer.

Chiang, D. (2000) Statisitical parsing with an automatically-extracted tree adjoining grammar. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 456–463, Hong Kong, P. R. China.

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

Chiou, F.-D., Chiang, D. and Palmer, M. (2001) Facilitating Treebank annotating using a statistical parser. *Proc. of the Human Language Technology Conference (HLT-2001)*, San Diego, CA.

Chomsky, N. (1981) *Lectures on Government and Binding*. Dordrecht: Foris.

CKIP (1995) An Introduction to the Academia Sinica Balanced Corpus (in Chinese). Technical Report 95-02, Taipei: Academia Sinica.

Collins, M. (1997) Three generative, lexicalised models for statistical parsing. In *Proc. of ACL-1997*.

Collins, M. (2000) Discriminative reranking for natural language parsing. *Proc. of ICML-2000*.

Converse, S. P. (2002) Developing guidelines for the annotation of anaphors in the Chinese Treebank. *Proceedings 1st SIGHAN Workshop on Chinese Language Processing*, Taipei, Taiwan.

Dai, X.-L. (1992) *Chinese Morphology and its Interface with the Syntax*. PhD thesis, Ohio State University.

Dang, H. T., Chia, C., Palmer, M. and Chiou, F.-D. (2002) Simple features for Chinese word sense disambiguation. *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Gan, K.-W. (1995) *Integrating Word Boundary Disambiguation with Sentence Understanding*. PhD thesis, National University of Singapore.

Gong, Q. (1997) *Zhongguo Yufaxue Shi (The History of Chinese Syntax)*. *Yuwen* Press.

Hajič, J. (1998) Building a syntactically annotated corpus: the Prague Dependency Treebank. *Issues of Valency and Meaning*. Karolinum, Praha.

Halle, M. and Marantz, A. (1993) Distributed Morphology and Pieces of Inflection. In: Hale K. and S. J. Keyser, editors, The View from Building 20, MIT Press. Cambridge.

Han, C., Han, N., Ko, E. and Palmer, M. (2002) Korean treebank: development and evaluation. *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*, Las Palmas, Spain.

Hockenmaier, J. (2003) *Data and Models for Statistical Parsing with Combinatory Categorial Grammar*. PhD thesis, University of Edinburgh.

Huang, J. C.-T. (1982) *Logic Relations in Chinese and the Theory of Grammar*. PhD thesis, MIT.

Jackendoff, R. S. (1977) *X-bar Syntax*. MIT Press.

Levy, R. and Manning, C. (2003) Is it harder to parse Chinese, or the Chinese Treebank. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan.

Li, M., Li, J., Dong, Z., Wang, Z. and Lu, D. (2003) Building a large Chinese corpus annotated with semantic dependency. *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.

Liu, Y., Tan, Q. and Shen, X. (1993) Segmentation Standard for Modern Chinese Information Processing and Automatic Segmentation Methodology.

Luo, X. (2003) A maximum entropy Chinese character-based parser. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, Sapporo, Japan.

Marcus, M., Santorini, B. and Marcinkiewicz, M. A. (1993) Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.

Marcus, M., Kim, G., Marcinkiewicz, M. A. *et al.* (1994) The Penn Treebank: Annotating, predicate argument structure. *Proc. of the ARPA Speech and Natural Language Workshop*.

Packard, J. L., editor. (1998) *New Approaches to Chinese Word Formation: Morphology, Phonology and the Lexicon in Modern and Ancient Chinese*. Mouton de Gruyter.

Packard, J. L. (2000) *The Morphology of Chinese: A linguistic and Cognitive Approach.* Cambridge University Press.

Palmer, M., Gildea, D. and Kingsbury, P. (To appear) The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics.*

Ratnaparkhi, A. (1996) A maximum entropy part-of-speech tagger. *Proceedings of the Empirical Methods in Natural Language Processing Conference*, Pennsylvania.

Ratnaparkhi, A. (1998) *Maximum Entropy Models for Natural Language Ambiguity Resolution.* PhD thesis, University of Pennsylvania.

Santos, D., Costa, L. and Rocha, P. 2003. Cooperatively evaluating Portuguese morphology. In: Mamede, N. J., Baptista, J., Trancoso, I. and das Gragas Volpe Nunes, M., editors, *Computational Processing of the Portuguese Language, Proceedings of the 6th International Workshop, (PROPOR 2003).* pp. 259–266. Faro, Portugal. Springer-Verlag.

Sciullo, A. M. Di and Williams, E. (1987) *On the Definition of Word*. MIT Press.

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.

Sproat, R. and Emerson, T. (2003) The First International Chinese Word Segmentation Bakeoff. *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.

Steedman, M. (1996) *Surface Structure and Interpretation.* MIT Press.

Steedman, M. (2000) *The Syntactic Process.* MIT Press.

Xia, F. (2000a) The Part-of-Speech Guidelines for Chinese Treebank Project. Technical Report IRCS 00-07, University of Pennsylvania.

Xia, F. (2000b) The Segmentation Guidelines for Chinese Treebank Project. Technical Report IRCS 00-06, University of Pennsylvania.

Xia, F. (2001) *Automatic Grammar Generation from Two Different Perspectives.* PhD thesis, University of Pennsylvania.

Xia, F., Han, C., Palmer, M. and Joshi, A. (2001) Extracting and comparing lexicalized grammars for different languages. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence, (IJCAI-2001)*, pp. 1321–1326. Seattle, WA.

Xia, F. and Palmer, M. (2001) Converting dependency structures to phrase structures. In: Allan, J., editor, *First International Conference on Human Language Technology Research*, pp. 61–65. Morgan Kaufmann.

Xue, N. (2001) *Defining and Automatically Identifying Words in Chinese*. PhD thesis, University of Delaware.

Xue, N., Chiou, F.-D. and Palmer, M. (2002) Building a large annotated Chinese corpus. *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Xue, N. and Converse, S. (2002) Combining classifiers for Chinese word segmentation. *Proceedings of the 1st SIGHAN Workshop on Chinese Language Processing*, Taipei, Taiwan.

Xue, N. and Palmer, M. (2003) Annotating the propositions in the Penn Chinese Treebank. *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.

Xue, N. and Xia, F. (2000) The Bracketing Guidelines for Penn Chinese Treebank Project. Technical Report IRCS 00-08, University of Pennsylvania.

Yu, S., Zhu, X., Wang, H. and Zhang, Y. (1998) *The Grammatical Knowledge-base of Contemporary Chinese — A Complete Specification (in Chinese).* Tsinghua University Press.