The Construction of A Chinese Shallow Treebank

Ruifeng Xu
Dept. Computing,
The Hong Kong Polytechnic University,
Kowloon, Hong Kong
csrfxu@comp.polyu.edu.hk

Qin Lu
Dept. Computing,
The Hong Kong Polytechnic University,
Kowloon, Hong Kong
csluqin@comp.polyu.edu.hk

Yin Li
Dept. Computing,
The Hong Kong Polytechnic University,
Kowloon, Hong Kong
csyinli@comp.polyu.edu.hk

Wanyin Li
Dept. Computing,
The Hong Kong Polytechnic University,
Kowloon, Hong Kong
cswyli@comp.polyu.edu.hk

Abstract

This paper presents the construction of a manually annotated Chinese shallow Treebank, named *PolyU Treebank.* Different from traditional Chinese Treebank based on full parsing, the PolyU Treebank is based on shallow parsing in which only partial syntactical structures are annotated. This Treebank can be used to support shallow parser training, testing and other natural language applications. Phrase-based Grammar, proposed by Peking University, is used to guide the design and implementation of the PolyU Treebank. The design principles include good resource sharing, low structural complexity, sufficient syntactic information and large data scale. The design issues, including corpus material preparation, standard for word segmentation and POS tagging, and the guideline for phrase bracketing and annotation, are presented in this paper. Well-designed workflow and effective semiautomatic and automatic annotation checking are used to ensure annotation accuracy and consistency. Currently, the PolyU Treebank has completed the annotation of a 1-million-word corpus. The evaluation shows that the accuracy of annotation is higher than 98%.

1 Introduction

A Treebank can be defined as a syntactically processed corpus. It is a language resource containing annotations of information at various linguistic levels such as words, phrases, clauses and sentences to form a 'bank of linguistic trees'. There are many Treebanks built for different languages such as the Penn Treebank (Marcus 1993), ICE-GB (Wallis 2003), and so on. The Penn Chinese Treebank is an important resource (Xia et al. 2000; Xue et al. 2002). Its annotation is based on Head-driven Phrase Structure Grammar (HPSG). The corpus of 100,000 Chinese words has been manually annotated with a strict quality assurance process. Another important work is the Sinica Treebank at the Academic Sinica, Taiwan (Chen et al. 1999; Chen et al. 2003). Information-based Case Grammar (ICG) was selected as the language framework. A head-driven chart parser was performed to do phrase bracketing and annotating. Then, manual post-editing was conducted. According to the report, The Sinica Treebank contains 38,725 parsed trees with 329,532 words.

Most reported Chinese Treebanks, including the two above, are based on full parsing which requires complete syntactical analysis including determining syntactic categories of words, locating chunks that can be nested, finding relations between phrases and resolving the attachment ambiguities. The output of full parsing is a set of complete syntactic trees. Automatic full parsing, however, is difficult to achieve good performance. Shallow parsing (or partial parsing) is usually defined as a parsing process aiming to provide a limited amount of local syntactic information such as non-recursive noun phrases, V-O structures and S-V structures etc. Since shallow parsing can recognize the backbone of a sentence more effectively and accurately with lower cost, people has in recent years started to work using results from shallow parsing. A shallow parsed Treebank can be used to extract information for different applications especially for training shallow parsers.

Different from full parsing, annotation to a shallow Treebank is only targeted at certain local structures in a sentence. The depth of "shallowness" and the scope of annotation vary from different reported work. Thus, two issues in shallow Treebank annotation is (1) what information and (2) to what depths the syntactic information should be annotated. Generally speaking, the degree of "shallowness" and the syntactical labeling are determined by the requirement of the serving applications. The choice of full parsing or shallow parsing is dependent on the need of the application including resources and

the capability of system to be developed (Xia et al. 2000; Chen et al. 2000; Li et al. 2003). Currently, there is no large-scale shallow annotated Treebank available as a publicly resource for training and testing.

In this paper, we present a manually annotated shallow Treebank, called the *PolyU Treebank*. It is targeted to contain 1-million-word contemporary Chinese text. The whole work on the PolyU Treebank follows the Phrase-based Grammar proposed by Peking University (Yu et al. 1998). In this language framework, a phrase, lead by a lexical word(or sometimes called a content word) as a head, is considered the basic syntactical unit in a Chinese sentence. The building of the PolyU Treebank was originally designed as training data for a shallow parser used for Chinese collocation extraction. From linguistics viewpoint, a collocation occurs only in words within a phrase, or between the headwords of related phrases (Zhang and Lin 1992). Therefore, the use of syntactic information is naturally considered an effective way to improve the performance of collocation extraction systems. The typical problems like *doctor-nurse* (Church and Hanks 1990) could be avoided by using such information. When employing syntactical information in collocation extraction, we restrict ourselves to identify the stable phrases in the sentences with certain levels of nesting. Thus it has motivated us to produce a shallow Treebank.

A natural way to obtain a shallow Treebank is through extracting shallow structures from a fully parsed Treebank. Unfortunately, all the available fully parsed Treebank, such as the Penn Treebank and the Sinica Treebank, are annotated using different grammars than our chosen Phrase-based Grammar. Also, the sizes of these Treebank are much smaller in scale to be useful for training our shallow parser.

This paper presents the most important design issues of the PolyU Treebank and the quality control mechanisms. The rest of this paper is organized as follows. Section 2 introduces the overview and design principles. Section 3 to Section5, present the design issues on corpus material preparation, the standard for word segmentation and POS tagging, and the guideline for phrase bracketing and labeling, respectively. Section 6 discusses the quality assurance mechanisms including a carefully designed workflow, parallel annotation, and automatic and semi-automatic post-annotation checking. Section 7 gives the current progress and future work.

2 Overview and Design Principles

The objective of this project is to manually construct a large shallow Treebank with high accuracy and consistency.

The design principles of The PolyU Treebank are: high resource sharing ability, low structural complexity, sufficient syntactic information and large data scale. First of all, the design and construction of The PolyU Treebank aims to provide as much a general purpose Treebank as possible so that different applications can make use of it as a NLP resource. With this objective, we chose to follow the well-known Phrase-based Grammar as the framework for annotation as this grammar is widely accepted by Chinese language researchers, and thus our work can be easily understood and accepted.

Due to the lack of word delimitation in Chinese, word segmentation must be performed before any further syntactical annotation. High accuracy of word segmentation is very important for this project. In this project, we chose to use the segmented and tagged corpus of People Daily annotated by the Peking University. The annotated corpus contains articles appeared in the People Daily Newspaper in 1998. The segmentation is based on the guidelines, given in the Chinese national standard GB13715, (Liu et al. 1993) and the POS tagging specification was developed according to the 'Grammatical Knowledge-base of contemporary Chinese'. According to the report from Peking University, the accuracy of this annotated corpus in terms of segmentation and POS tagging are 99.9% and 99.5%, respectively (Yu et al. 2001). The use of such mature and widely adopted resource can effectively reduce our cost, ensure syntactical annotation quality. With consistency in segmentation, POS, and syntactic annotation, the resulting Treebank can be readily shared by other researchers as a public resource.

The second design principle is low structural complexity. That means, the annotation framework should be clear and simple, and the labeled syntactic and functional information should be commonly used and accepted. Considering the characteristics of shallow annotation, our project has focused on the annotation of phrases and headwords while the sentence level syntax are ignored.

Following the framework of Phrase-based Grammar, a *base-phrase* is regarded as the smallest unit where a base-phrase is defined as a 'stable' and 'simple' phrase without nesting components. Study on Chinese syntactical analysis suggests that phrases should be the fundamental unit instead of words in a sentence. This is because, firstly, the usage of Chinese words is very flexible. A word may have different POS tags serving for different functions in sentences. On the contrary, the use of Chinese phrases is much more stable. That is, a phrase has very limited functional use in a sentence. Secondly, the construction rules of Chinese phrases are nearly

the same as that of Chinese sentences. Therefore, the analysis of phrases can help identifying POS and grammatical functions of words. Naturally, it should be regarded as the basic syntactical unit. Usually, a base-phrase is driven by a lexical word as its headword. Examples of base-phrases include base NP, base VP and so on, such as the sample shown below.

[市场/n 经济/n]NP [团体/n 旅客/n]NP

Using base-phrases as the start point, nested levels of phrases are then identified, until the maximum phrases (will be defined later) are identified. Since we do not intend to provide full parsing information, there has to be a limit on the level of nesting. For practical reasons, we choose to limit the nesting of brackets to 3 levels. That means, the depth of our shallow parsed Treebank will be limited to 3. This restriction can limit the structural complexity to a manageable level.

Our nested bracketing is not strictly bottom up. That is we do not simply extend from base-phrase and move up until the 3rd level. Instead, we first identify the *maximal-phrase* which is used to identify the backbone of the sentence. The maximal-phrase provides the framework under which the base-phrases of up to 2 levels can be identified. The principles for the identification of scope and depth of phrase bracketing are briefly explained below and the operating procedure is indicated by the given order in which these principles are presented. More details is given in Section 5.

Step 1: Annotation of maximal-phrase which is the shortest word sequence of maximally spanning non-overlapping edges which plays a distinct semantic role of a predicate. A maximal-phrase contains two or more lexical words.

Step 2: Annotation of base-phrases within a maximal-phrase. In case a base-phrase and a maximal-phrase are identical and the maximal-phrase is already bracketed in Step 1, no bracketing is done in this step. For each identified base-phrase, its headword will be marked.

Step 3: Annotation of next level of bracketing, called mid-phrase which is expended from a base-phrase. A mid-phrase is annotated only if it is deemed necessary. The process starts from the identified base-phrase. One more level of syntactical structure is then bracketed if it exists within the maximal-phrase.

The third design principle is to provide sufficient syntactical information for natural language application even though shallow annotation does not necessarily contain complete syntactic information at sentence level. Some past research in Chinese shallow parsing were on single level base-phrases only (Sun 2001). However, for certain applications, such as for collocation extraction, identification of base-phrases only are not very useful. In this project, we have decided to annotate phrases within three levels of nesting within a sentence. For each phrase, a label is be given to indicate its syntactical information, and an optional semantic or structural label is given if applicable. Furthermore, the headword of a base-phrase is annotated. We believe these information are sufficient for many natural language processing research work and it is also manageable for this project within its working schedule.

Fourthly, aiming to support practical language processing, a reasonably large annotated Treebank is expected. Studies on English have shown that Treebank of word size 500K to 1M is reasonable for syntactical structure analysis (Leech and Garside 1996). In consideration of the resources available and the reference of studies on English, we have set out our Treebank size to be one million words. We hope such a reasonably large-scale data can effectively support some language research, such as collocation extraction.

We chose to use the XML format to record the annotated data. Other information such as original article related information (author, date, etc.), annotator name, and other useful information are also given through the meta-tags provided by XML. All the meta-tags can be removed by a program to recover the original data.

We have performed a small-scale experiment to compare the annotation cost of shallow annotation and full annotation (followed Penn Chinese Treebank specification) on 500 Chinese sentences by the same annotators. The time cost in shallow annotation is only 25% of that for full annotation. Meanwhile, due to the reduced structural complexity in shallow annotation, the accuracy of first pass shallow annotation is much higher than full annotation.

3 Corpus Materials Preparation

The People Daily corpus, developed by PKU, consists of more than 13k articles totaling 5M words. As we need one million words for our Treebank, we have selected articles covering different areas in different time span to avoid duplications due to short-lived events and news topics. Our selection takes each day's news as one single unit, and then several distant dates are randomly selected among the whole 182 days in the entire collection. We have also decided to keep the original articles' structures and topics indicators as they may be useful for some applications.

4 Word Segmentation and Part-of-Speech Tagging

The articles selected from PKU corpus are already segmented into words following the guidelines given in GB13715. The annotated corpus has a basic lexicon of over 60,000 words. We simply use this segmentation without any change and the accuracy is claimed to be 99.9%.

Each word in the PKU corpus is given a POS tag. In this tagging scheme, a total of 43 POS tags are listed (Yu et al. 2001). Our project takes the PKU POS tags with only notational changes explained as follows:

The morphemes tags including *Ag* (Adjectives morphemes), *Bg, Dg, Ng, Mg, Rg, Tg, Qg,* and *Ug* are re-labeled as lowercase letters, *ag, bg, dg, ng, mg, rg, tg, qg* and *ug*, respectively. This modification is to ensure consistent labeling in our system where the lower cases are used to indicate word-level tags and upper cases are used to indicate phrase-level labels.

5 Phrase Bracketing and Annotation

Phrase bracketing and annotation is the core part of this project. Not only all the original annotated files are converted to XML files, results of our annotations are also given in XML form. The meta tags provided by XML are very helpful for further processing and searching to the annotated text. .

Note that in our project, the basic phrasal analysis looks at the context of a clause, not a sentence. Here, the term *clause* refers the text string ended by some punctuations including comma (,), semicolon (;), colon (:), or period (.). Certain punctuation marks such as '、', '<', and '>' are not considered clause separators. For example,

经过严密侦查，警方锁定了2个嫌疑人。

is considered having two clauses and thus will be bracketed separately. It should be pointed out that he set of Chinese punctuation marks are different from that of English and their usage can also be different. Therefore, an English sentence and their Chinese translation may use different punctuation marks. For example, the sentence

汤姆、约翰、和杰克一起回学校

is the translation of the English '*Tom, John, and Jack go back to school together*' , which uses '、' rather than comma(,) to indicate parallel structures, and is thus considered one clause.

Each clause will then be processed according to the principles discussed in Section 2. The symbols '[' and ']' are used to indicate the left and right boundaries of a phrase. The right bracket is appended with syntactic labels as described in the general form of [Phrase]SS-FF, where SS is a mandatory syntactic label such as *NP*(noun phrase) and *AP*(adjective phrase), and FF is an optional label indicating internal structures and semantic functions such as *BL*(parallel), *SB*(a noun is the object of verb within a verb phrase). A total of 21 SS labels and 20 FF labels are given in our phrase annotation specification. For example, the functional label *BL* identifies parallel components in a phrase as indicated in the example [荣誉/n 与/c 尊严/n]NP-BL.

As in another example shown below,

[美国/ns 科学家/n]NP [绘制/v# 出/v]VP-SBU

[X染色体/n 的/u [高/a 精度/n 图谱/n]NP]NP

the phrase [绘制/v 出/v] is a verb phrase, thus it is labeled as *VP*. Furthermore, the verb phrase can be further classified as a verb-complement type. Thus an additional SBU function label is marked. We should point out that since the FF labels are not syntactical information and are thus not expected to be used by any shallow parsers. The FF labels carry structural and/or semantic information which are of help in annotation. We consider it useful for other applications and thus decide to keep them in the Treebank. **Appendix 1** lists all the FF labels used in the annotation.

5.1 Identification of Maximal-phrase:

The maximal-phrases are the main syntactical structures including subject, predicate, and objects in a clause. Again, maximal-phrase is defined as the phrase with the maximum spanning non-overlapping length, and it is a predicate playing a distinct semantic role and containing more than one lexical word. That means a maximal-phrase contains at least one base-phrase. As this is the first stage in the bracketing process, no nesting should occur. In the following annotated sentence,

[中国/ns 旅游年/n]NP 是/v [一/m 次/q
国家级/b 的/u 宣传/vn 促销/vn 活动/vn]NP (Eg.1)

there are two separate maximal-phrases, [中国/ns 旅游年/n]NP, and [一/m 次/q国家级/b 的/u 宣传/vn 促销/vn 活动/vn]NP. Note that 是/v is considered a base-phrase, but not a maximal-phrase because it contains only one lexical word. Unlike many annotations where the object of a sentence is included as a part of the verb phrase, we treat them as separate maximal-phrases both due to our requirement and also for reducing nesting.

If a clause is completely embedded in a larger clause, it is considered a special clause and given a special name called an *internal clause* . We will bracket such an internal clause as a maximal phrase with the tag *'IC'* as shown in the following example,

[极大/a 地/u 鼓舞/v]VP [全党/n 和/c 全国/n 各族/r
人民/n 更加/d 紧密/a 地/u 团结/v 起来/v]IC

5.2 Annotation of Base-phrases:

A base-phrase is the phrase with stable, close and simple structure without nesting components. Normally a base-phrase contains a lexical word as

headword. Taking the maximal-phrase [一/m 次/q 国家级/b 的/u 宣传/vn 促销/vn 活动/vn]NP in Eg.1 as an example, [一/m 次/q]QP and [宣传/vn 促销/vn 活动/vn]NP, are base-phrases in this maximal-phrase. Thus, the sentence is annotated as

[中国/ns 旅游年/n]NP 是/v

[[一/m 次/q]QP-ML 国家级/b 的/u [宣传/vn 促销/vn 活动/vn]NP]NP

In fact, [中国/ns 旅游年/n]NP and 是/v are also base-phrases. 是/v is not bracketed because it is a single lexical word as a base-phrase without any ambiguity and it is thus by default not being bracketed. [中国/ns 旅游年/n]NP is not further bracketed because it overlaps with a maximal-phrase. Our annotation principle here is that if a base-phrase overlaps with a maximal-phrase, it will not be bracketed twice.

The identification of base-phrase is done only within an already identified maximal-phrase. In other words, if a base-phrase is identified, it must be nested inside a maximal-phrase or at most overlaps with it. It should be pointed out that the identification of a base-phrase is the most fundamental and most important goal of Treebank annotation. The identification of maximal-phrases can be considered as parsing a clause using a top-down approach. On the other hand, the identification of a base-phrase is a bottom up approach to find the most basic units within a maximal-phrase.

5.3 Mid-Phrase Identification:

Due to the fact that sometimes there may be more syntactic structures between the base-phrases and maximal-phrases, this step uses base-phrase as the starting point to further identify one more level of the syntactical structure in a maximal-phrase. Takes Eg.1 as an example, it is further annotated as

[中国/ns 旅游年/n]NP 是/v [[一/m 次/q]QP-ML
[国家级/b 的/u [宣传/vn 促销/vn 活动/vn]NP]NP]NP

where the underlined text shows the additional annotation.

As we only limit our nesting to three levels, any further nested phrases will be ignored. The following sentence shows the result of our annotation with three levels of nesting:

[目前/t [企业/n 集团/n 发展/vn]NP [值得/v 注意/v
的/u [[几/m 个/q]QP 问题/n]NP]NP]NP

However, a full annotation should have 4 levels of nesting as shown below. The underlined text is the 4th level annotation skipped by our system.

[目前/t [企业/n 集团/n 发展/vn [值得/v 注意/v 的/u
[[几/m 个/q]QP 问题/n]NP]NP]NP]NP

5.4 Annotation of Headword

In our system, a '#' tag will be appended after a word to indicate that it is a headword of the base-phrase. Here, a headword must be a lexical word rather than a function word.

In most cases, a headword stays in a fixed position of a base-phrase. For example, the headword of a noun phrase is normally the last noun in this phrase. Thus, we call this position the default position. If a headword is in the default position, annotation is not needed. Otherwise, a '#' tag is used to indicate the headword.

For example, in a clause,

[美国/ns 科学家/n]NP [绘制/v 出/v]VP-SBU,

[绘制/v 出/v] is a verb phrase, and the headword of the phrase is 绘制/v, which is not in the default position of a verb phrase. Thus, this phrase is further annotated as:

[美国/ns 科学家/n]NP [绘制/v# 出/v]VP-SBU

Note that 科学家/n is also a headword, but since it is in the default position, no explicit annotation is needed.

6 Annotation and Quality Assurance

Our research team is formed by four people at the Hong Kong Polytechnic University, two linguists from Beijing Language and Culture University and some research collaborators from Peking University. Furthermore, the annotation work has been conducted by four post-graduate students in language studies and computational linguistics from the Beijing Language and Culture University.

The annotation work is conducted in 5 separate stages to ensure quality output of the annotation work. The preparation of annotation specification and corpus selection was done in the first stage. Researchers in Hong Kong invited two linguists from China to come to Hong Kong to prepare for the corpus collection and selection work. A thorough study on the reported work in this area was conducted. After the project scope was defined, the SS labels and the FF labels were then defined. A Treebank specification was then documented. The Treebank was given the name PolyU Treebank to indicate that it is produced at the Hong Kong Polytechnic University. In order to validate the specifications drafted, all the six members first manually annotated 10k-word material, separately. The outputs were then compared, and the problems and ambiguities occurred were discussed and consolidated and named Version 1.0. Stage 1 took about 5 months to complete. Details of the specification can be downloaded from the project website www.comp.polyu.edu.hk/~cclab.

In Stage 2, the annotators in Beijing were then involved. They had to first study the specification and understand the requirement of the annotation. Then, the annotators under the supervision of a team member in Stage 1 annotated 20k-word materials together and discussed the problems occurred.

During this two-month work, the annotators were trained to understand the specification. The emphasis at this stage was to train the annotators' good understanding of the specification as well as consistency by each annotator and consistency by different annotators. Further problems occurred in the actual annotation practice were then solved and the specification was also further refined or modified.

In Stage 3, which took about 2 months, each annotator was assigned 40k-word material each in which 5k-words material were duplicate annotated to all the annotators. Meanwhile, the team members in Hong Kong also developed a post-annotation checking tool to verify the annotation format, phrase bracketing, annotation tags, and phrase marks to remove ambiguities and mistakes. Furthermore, an evaluation tool was built to check the consistency of annotation output. The detected annotation errors were then sent back to the annotators for discussion and correction. Any further problems occurred were submitted for group discussion and minor modification on the specification was also done.

In stage 4, each annotator was dispatched with one set of 50k-word material each time. For each distribution, 15k-word data in each set were distributed to more than two annotators in duplicates so that for any three annotators, there would be 5K duplicated materials. When the annotators finished the first pass annotation, we used the post-annotation checking tool to do format checking in order to remove the obvious annotation errors such as wrong tag annotation and cross bracketing. However, it was quite difficult to check the difference in annotation due to different interpretation of a sentence. What we did was to make use of the annotations done on the duplicate materials to compare for consistency. When ambiguity or differences were identified, discussions were conducted and a result used by the majority would be chosen as the accepted result. The re-annotated results were regarded as the Golden Standard to evaluate the accuracy of annotation and consistency between different annotators. The annotators were required to study this Golden Standard and go back to remove similar mistakes. The annotated 50k data was accepted only after this. Then, a new 50k-word materials was distributed and repeated in the same way. During this stage, the ambiguous and out-of-tag-set phrase structures were marked as *OT* for further process. The annotation specification was not modified in order to avoid frequent revisit to already annotated data. About 4 months were spent on this stage.

In Stage 5, all the members and annotators were grouped and discuss the *OT* cases. Some typical new phrase structure and function types were appended in the specification and thus the final formal annotation specification was established. Using this final specification, the annotators had to go back to check their output, modify the mistakes and substitute the *OT* tags by the agreed tags. Currently, the project was already in Stage 5 with 2 months of work finished. A further 2 months was expected to complete this work.

Since it is impossible to do all the checking and analysis manually, a series of checking and evaluating tools are established. One of the tools is to check the consistency between text corpus files and annotated XML files including checking the XML format, the filled XML header, and whether the original txt material is being altered by accident. This program ensures that the XML header information is correctly filled and during annotation process, no additional mistakes are introduced due to typing errors.

Furthermore, we have developed and trained a shallow parser using the Golden Standard data. This shallow parser is performed on the original text data, and its output and manually annotated result are compared for verification to further remove errors

Now, we are in the process of developing an effective analyzer to evaluate the accuracy and consistency for the whole annotated corpus. For the exactly matched bracketed phrases, we check whether the same phrase labels are given. Abnormal cases will be manually checked and confirmed. Our final goal is to ensure the bracketing can reach 99% accuracy and consistency.

7 Current Progress and Future Work

As mentioned earlier, we are now in Stage 5 of the annotation. The resulting annotation contains 2,639 articles selected from PKU People Daily corpus. These articles contains 1, 035, 058 segmented Chinese words, with on average, around 394 words in each article. There are a total of 284, 665 bracketed phrases including nested phrases. A summary of the different SS labels used are given in Table 1.

NP	NT	NS	NR	NZ	TP	FP
117799	7899	1452	10137	339	5216	2431
S	V	VP	AP	DP	PP	QP
4910	4827	71000	16688	154	25672	9143
DE	SU	XD	IC	BA	BEI	RP
1252	125	178	3889	1010	373	171

Table 1. Statistics of annotated syntactical phrases

For each bracketed phrase, if its FF label does not fit into the corresponding default pattern, (like for the noun phrase(*NP*), the default grammatical structure is that the last noun in the phrase is the headword and other components are the modifiers, using *PZ* tags), its FF labels should then be explicitly labeled. The statistics of annotated FF tags

are listed in Table 2.

BL	FZ	ZZ	SBI	SBU	SD	DU
12341	5792	1253	14519	8805	5096	810
FJ	JY	DL	ML	SL	YY	DX
263	2699	26	160	352	507	10172
DD	FS	MD	GJ	SJ	OT	
583	1161	1091	102	2932	5255	

Table 2. Statistics of function and structure tags

For the material annotated by multiple annotators as duplicates, the evaluation program has reported that the accuracy of phrase annotation is higher than 99.5% and the consistency between different annotators is higher than 99.8%. As for other annotated materials, the quality evaluation program preliminarily reports the accuracy of phrase annotation is higher than 98%. Further checking and evaluation work are ongoing to ensure the final overall accuracy achieves 99%.

Up to now, the FF labels of 5,255 phrases are annotated as *OT*. That means about 1.8% (5,255 out of a total of 284,665) of them do not fit into any patterns listed in Table 2. Most of them are proper noun phrase, syntactically labeled as PP. We are investigating these cases and trying to identify whether some of them can be in new function and structure patterns and give a new label.

It is also our intention to further develop our tools to improve the automatic annotation analysis and evaluation program to find out the potential annotation error and inconsistency. Other visualization tools are also being developed to support keyword searching, context indexing, and annotation case searching. Once we complete Stage 5, we intend to make the PolyU Treebank data available for public access. Furthermore, we are developing a shallow parser and using The PolyU Treebank as training and testing data.

Acknowledgement

This project is partially supported by the Hong Kong Polytechnic University (Project Code A-P203) and CERG Grant (Project code 5087/01E)

References

Baoli Li, Qin Lu and Yin Li. 2003. Building a Chinese Shallow Parsed Treebank for Collocation Extraction, *Proceedings of CICLing 2003*: 402-405

Fei Xia, et al. 2000. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation *Proceedings of LREC-2000*, Greece

Feng-yi Chen, et al. 1999. Sinica Treebank, *Computational Linguistics and Chinese Language Processing,* 4(2):183-204

G. N. Leech, R.Garside. 1996. *Running a grammar factory: the production of syntactically analyzed corpora or "treebanks",* Johansson and Stenstron.

Honglin Sun, 2001. *A Content Chunk Parser for Unrestricted Chinese Text,* Ph.D Thesis, Peking University, 2001

Keh-jiann Chen et al. 2003. *Building and Using Parsed Corpora* (Anne Abeillé ed. s) KLUWER, Dordrecht

Kenneth Church, and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography, *Computational Linguistics*, 16(1): 22-29

Marcus, M. et al. 1993. Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, 19(1): 313-330.

Nianwen Xue, et al. 2002. Building a Large-Scale Annotated Chinese Corpus, *Proceedings of COLING 2002*, Taipei, Taiwan

Sean Wallis, 2003. *Building and Using Parsed Corpora* (Anne Abeillé eds) KLUWER, Dordrecht

Shiwen Yu, et al. 1998. *The Grammatical Knowledge- base of contemporary Chinese: a complete specification.* Tsinghua University Press, Beijing, China

Shiwen Yu, et al. 2001. Guideline of People Daily Corpus Annotation, Technical report, Beijing University

Shoukang Zhang and Xingguang Lin, 1992. *Collocation Dictionary of Modern Chinese Lexical Words*, Business Publisher, China

Yuan Liu, et al. 1993. *Segmentation standard for Modern Chinese Information Processing and automatic segmentation methodology*. Tsinghua University Press, Beijing, China

Appendix 1 The structural and semantic FF labels

BL	并列关系	[中国/ns 与/c 南非/ns]NP-BL
FZ	复指关系	[[祖国/n 的/u 心脏/n]NP [天安门/ns 广场/n]NS]NP-FZ
PZ	偏正关系	[稚嫩/a 的/u 娃娃脸/n]NP-PZ
ZZ	状中结构	[全力/n 组织/v]V-ZZ
SBI	述宾关系	[帮助/v 群众/n]V-SBI
SBU	述补关系	[负担/v [过/d 重/a]AP]VP-SBU
SD	顺递关系	[审核/v 发放/v]V-SD
JY	兼语关系	[[切实/ad 帮助/v]VP 群众/n [解决/v 各种/r 实际/a 困难/an]VP-SBI]VP-JY
PO	时间信息	[7 月/t 1 日/t]TP-PO
DU	时段信息	[今后/t 3 /m 年/q]TP-DU
FJ	附加信息	[代表/n 们/k]NP-FJ
DL	动量信息	[十/m 下/v 江南/ns]VP-DL
ML	名量信息	[5 /m 次/q]QP-ML
SL	时量信息	[[第二/m 天/q]QP-SL 4 时/t]TP
YY	原因信息	[因/p 冻/v]PP-YY [因/p 饿/a]PP-YY 死亡/v
DX	对象信息	[向/p [受灾/vn 地区/n]NP]PP-DX
DD	地点信息	[在/p 深圳/ns]PP-DD [参观/v 考察/v]V-BL
FS	方式信息	[通过/p [股票/n 上市/v]S]PP-FS
MD	目的信息	[为/p [发展/v 两岸/n 关系/n]VP-SBI]PP-MD
GJ	工具信息	[用/p 公款/n]PP-GJ 请客/v
SJ	时间信息	[于/p [3 1 日/t 上午/t]TP]PP-SJ
OT	其他	

Appendix 2 Example of an Annotated Article

XW_RM_19980402-07-009.xml

```
<?xml version="1.0" encoding="GB2312" ?>
<!DOCTYPE SPFILE SYSTEM "spfile.dtd">
<SPFILE><!--SPFILE stands for shallow parsed file. -->
    <metadata>
    <filename>XW_RM_19980402-07-009.xml</filename>
    <source>RENMINRIBAO_19980402</source>
    <annotators>
        <drafter name="Zhang Qian" date="2003-4-26" />
        <reviser name="Linda Li" date="2003-5-2" />
        <reviser name="Ruifeng Xu" date="2003-10-12" />
    </annotators>
    </metadata>
<head><title>[[四千/m 年/q 前/f]TP [尸体/n 防腐/vn 技术/n]NP]NP </title>
    <subtitle> </subtitle> <authors>
    <author name="作者姓名 1" affiliation="作者单位 1" />
    <author name="作者姓名 2" affiliation="作者单位 2" email="a@b.c" /> </authors></head>
<body>
<![CDATA[
19980402-07-009-001/m [[四千/m 年/q 前/f]TP [尸体/n 防腐/vn 技术/n]NP]NP
19980402-07-009-002/m 考古学家/n [[对/p 4 0 0 0/m 多/m 年/q 前/f 的/u 一/m 具/q 埃及/ns 木乃伊/n 进行/v 研究/vn]VP 后/f]TP 发现/v ，/w [[古/a 埃及/ns]NP 人/n]NP [早/a 在/p 那时/r]AP-SBU [就/d 利用/v 香料/n 进行/v [尸体/n 防腐/vn]VP]VP 。/w
19980402-07-009-003/m [据/p [《/w 华盛顿/ns 邮报/n 》/w 报道/v]S]PP-DX ，/w [[德国/ns 塔宾根/nz 大学/n]NT 的/u [乌里奇·韦塞尔/nr 等/u 人/n]NP-FZ]NP [对/p [1 9 1 4 年/t 在/p 埃及/ns 基扎/ns 出土/v 的/u [伊杜/nr 的/u 木乃伊/n]NP]NP]PP-DX 进行/v 了/u 检验/vn 。/w
19980402-07-009-004/m 他们/r 发现/v ，/w [其/r 骨骼/n]NP [已经/d [用/p 松脂/n 和/c 以/p 钠/n 为主/v 的/u 化合物/n]PP-GJ [处理/v 过/u 了/y]VP-SBU]VP 。/w [松脂/n 和/c [钠/n 化合物/n]NP]NP-BL 具有/v [[防腐/v 和/c 保存/v 尸体/n]VP-BL 的/u 功效/n]NP 。/w 另外/c ，/w 他们/r [还/d 发现/v]VP [[伊杜/nr 遗体/n 的/u 骨骼/n]NP 浸透/v 了/u [具有/v 保护/vn 作用/n 的/u [粘性/n 物质/n]NP]NP]IC 。/w
19980402-07-009-005/m 伊杜/nr 是/v [[大约/d 公元前/t 2 1 5 0 年/t 时/Ng 古/a 埃及王国/ns 时期/n]TP 的/u [松木/n 贸易/vn 官员/n]NP]NP 。/w
]]></body>
</SPFILE>
```