Computational Linguistics and Chinese Language Processing
Vol. 10, No. 3, September 2005, pp. 397-416

The Design and Construction of the PolyU Shallow Treebank

Ruifeng Xu*, Qin Lu*, Yin Li* and Wanyin Li*

Abstract

This paper presents the design and construction of the PolyU Treebank, a manually annotated Chinese shallow treebank. The PolyU Treebank is based on shallow annotation where only partial syntactical structures within sentences are annotated. Guided by the Phrase-Standard Grammar proposed by Peking University, the PolyU Treebank has been designed and constructed to provide a large amount of annotated data containing shallow syntactical information and limited semantic information for use in natural language processing (NLP) research. This paper describes the relevant design principles, annotation guidelines, and implementation issues, including the achievement of high quality annotation through the use of well-designed annotation workflow and effective post-annotation checking tools. Currently, the PolyU Treebank consists of a one-million-word annotated corpus and has been used in a number of NLP research projects with promising results.

Keywords: Shallow Treebank, Shallow Parsing, Corpus Annotation, Natural Language Processing

1. Introduction

A treebank can be defined as a syntactically processed corpus. It is a language resource with linguistic information annotated at, variously, the word, phrase, clause, and sentence levels, in order to form a bank of linguistic trees. Many treebanks have been constructed for different languages, including Penn Treebank [Marcus *et al.* 1993] and the ICE-GB [Wallis *et al.* 2003] for English, and the Penn Chinese Treebank [Xia *et al*. 2000; Xue *et al.* 2002] and the Sinica Treebank [Chen *et al*. 1999; Chen *et al.* 2003] for Chinese.

Most of the reported Chinese treebanks, including the Penn Chinese Treebank and Sinica Treebank, are based on full parsing, where complete syntactical analysis is performed. This includes determining the syntactic categories of words, locating chunks that can be nested,

* Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong
Tel: +852-27667326; +852-27667247; +852-27667325 Fax: +852-27740842
E-mail: {csrfxu, csluqin, csyinli, cswyli}@comp.polyu.edu.hk

[Received October 13, 2004; Revised March 18, 2005; Accepted March 21, 2005]

finding relations between phrases, and resolving attachment ambiguities. Thus, the output of full parsing is a set of complete syntactic trees. Due to the complexity of natural languages, automatic full parsing is still quite challenging. An alternative to automatic full parsing is to adopt a divide-and-conquer strategy, i.e., to divide full parsing into several independent sub-tasks which can be applied relatively easily. One of these sub-tasks is shallow (or partial) parsing. The purpose of shallow parsing is to identify local syntactical structures that are relatively simple and easy to identify while ignoring the complicated task of analyzing how these phrases are syntactically used to construct sentences. Thus, shallow parsing only identifies local structures in sentences. These local structures form the sub-trees of a full syntactic tree. Because shallow parsing does not involve complex and ambiguous attachment analysis, it can find some local structures at much lower cost and with a much higher accuracy. For these reasons, shallow parsing has in recent years been the focus of more research, and it has been applied in many NLP applications. However, the lack of a large-scale Chinese shallow treebank has been an impediment to research in this area. This has motivated us to construct a Chinese shallow treebank for Chinese natural language processing applications. This treebank, referred as the PolyU Treebank, is named after the University where it is being developed.

One problem with shallow parsing is that, unlike full parsing, it seeks to identify only certain local structures in a sentence. Furthermore, at present, there is no widely-accepted common standard for the determining scope and depth of local structures, and different reported works vary in how they define what local structures are [Dalemans *et al.* 1999; Sun 2001; Li *et al.* 2003]. Therefore, in this work, we will first discuss the objectives of shallow parsing based on our needs and those of other NLP researchers and define the scope of shallow parsing. In accordance with this defined scope, we will then show how the PolyU Treebank has been constructed by manually annotating shallow syntactic structures from a selected corpus.

Obviously, the scope and the depth of shallow annotation should be determined based on the requirements of the applications using the treebank. Based on the typical requirements of NLP research tasks such as Chinese collocation extraction, terminology extraction, and the acquisition of descriptions of terminologies conducted at the authors' research institution, we restrict shallow syntactic structures to the *maximal phrases* that play various roles as subjects, predicates, complement clauses and other syntactic components in sentences. Within the scope of the present work, our aim is to identify *base-phrases*, that is minimum syntactic unit in a maximal phrase. We also identify those nested phrases between base-phrases and maximal phrases which we call *mid-phrases*. Maximal phrase, Base-phrase, Mid-phrase will be defined in detail in Section 3. Each identified phrase is given a mandatory syntactic label and an optional semantic label. Its header is also identified. An important feature of our treebank is

that the identified phrases are augmented with semantic information. This kind of information is useful in many areas of NLP research but is difficult to identify automatically and sometimes not annotated in the other existing treebanks.

For guidance in syntactic annotation, we choose to use the Phrase-Standard Grammar (PSG) as proposed by Peking University [Yu *et al.* 1998]. There are two reasons for this choice. First, the PSG grammar framework is widely accepted in mainland China. Second, in order to reduce the cost of annotation and to ensure the maximum sharing of our output, we perform shallow syntactic annotation on the segmented and tagged People's Daily corpus, developed in Peking University [Yu *et al.* 2001].

The process of constructing our treebank, which has taken more than 15 months, has included guideline design, the development of annotation specifications, and annotation and quality assurance checking. The one-million-word annotated shallow treebank is more than 98.8% accurate in terms of phrase bracketing and more than 98% accurate in phrase labeling. Such a large-scale treebank can be used to support a variety of NLP research. Currently, it has been used to train and to test a shallow parser [Lu *et al.* 2003]. Furthermore, other research conducted in authors' institution, including Chinese collocation extraction, Chinese terminologies extraction, and information retrieval, have also benefited from the PolyU Treebank. We are currently optimizing the treebank and making it available to other researchers as a public resource.

This paper presents the major issues involved in the design and construction of the PolyU Treebank and its quality control mechanisms. The rest of this paper is organized as follows. Section 2 introduces the design principles. Section 3 describes the annotation guidelines. Section 4 describes the tasks involved in annotating the PolyU Treebank, including corpus data preparation, word segmentation, POS tagging, phrase bracketing, and phrase labeling specifications. Section 5 discusses the quality assurance mechanisms and the post-annotation checking tools developed for this project. Section 6 gives some examples to illustrate how this shallow treebank can be used in NLP. Section 7 gives conclusions.

2. Design Principles

Due to the fact that currently, no large-scale shallow-annotated Chinese treebanks are available, in the course of designing PolyU Treebank, we referenced two important fully-annotated Chinese treebank: the Penn Chinese Treebank and the Sinica Treebank. The Penn Chinese Treebank was annotated based on the Government and Bind framework and contains about 500,000 Chinese words, most of which were mainly manually annotated according to a strict quality assurance process [Xue *et al.* 2002]. The Sinica Treebank was developed by the Academic Sinica, Taiwan. Phrase bracketing and annotation were carried out using a head-driven chart parser guided by Information-based Case Grammar (ICG), and

followed by manual post-editing. The Sinica Treebank contains 39,000 parsed trees and 329,000 words [Chen *et al*. 1999; Chen *et al.* 2003]. A natural way to obtain a shallow treebank is to extract shallow structures from a fully annotated treebank. Unfortunately, the Penn Treebank and Sinica Treebank were annotated using different grammar frameworks as well as different word segmentation/POS tagging strategies, making them unsuitable for our annotation scheme.

To ensure that the PolyU Treebank would be high in quality and widely accepted, it was designed and constructed based on four basic principles:

Principle 1: High resource-sharing capability

The PolyU Treebank was designed to sever as a general purpose treebank for use in as wide a range of applications as possible. This called for the selection of an effective and well-accepted grammatical framework for representing syntactical information as well as for a well-accepted word segmentation/POS tagging scheme.

We chose to use the Phrase-Standard Grammar (PSG), proposed by Peking University. PSG is widely accepted by Chinese NLP researchers. In the PSG framework, phrases rather than words are treated as basic Chinese syntactical units. The reason is that while an individual word can be used in different ways and may have different part-of-speech (POS) tags representing its different functions in sentences, a phrase is made up of a number of words normally driven by a headword, and consequently, has a stable internal structure and order. Based on this framework, syntactical analysis should be performed in a cascaded fashion, and a linear character string can finally be syntactically analyzed to form a cascaded tree.

In the absence of an orthographic device for delimiting words in Chinese, it is necessary to segment words before performing POS tagging. We used a segmented and tagged corpus consisting sentences from the People's Daily, annotated by Peking University. This corpus was accurately segmented and tagged in accordance with the PSG framework, and contains articles from the People's Daily published in 1998. The claimed accuracy of word segmentation and POS tagging is 99.9% and 99.5%, respectively [Yu *et al.* 2001]. Using this popular and accurate resource significantly reduced the cost of annotation in our research and ensured the maximum sharing of our output.

Principle 2: Low structural complexity

The second design principle was that the PolyU Treebank should not be structurally very complex; its annotation framework should be clear and simple and its syntactic and functional information should be labeled according to commonly used and widely accepted standards.

To ensure that our shallow annotation approach satisfied the requirements typical language applications in terms of syntactical information, we chose to focus on the annotation

of phrases and the identification of headwords while ignoring sentence-level syntax. More specifically, we wanted to identify three types of information: (1) *base-phrases,* that is, non-nesting phrases with at least one headword; (2) *maximal phrases,* that is, phrases that marked the boundary of our scope of examination, inclosing the base-phrases and plays the role of subject, predicate, complement clause, embedded clause, or other syntactic components of sentences; and (3) *mid-phrases,* that is the intermediate nesting phrases between base-phrases and maximal phrases if they existed. Maximal phrases and base-phrases will be defined and discussed in detail in Section 3. As for mid-phrases, a limit was imposed on the level of nesting since we did not intend to provide full parsing information. In order to limit the structural complexity, we limited nesting brackets to only three levels. In other words, mid-phrases were limited to only at most one level.

Principle 3: Sufficient and useful syntactic information

The third design principle was to provide syntactic information at a low level of complexity that would be useful for and effective in a wide variety of NLP applications. Earlier works in Chinese shallow annotation had annotated only non-nesting base-phrases [Sun 2001]. However, base-phrase annotation alone is not adequate for many applications. Our annotation scheme permits three levels of nesting, and this has a number of advantages. First, maximal phrases indicate the essential syntactic elements of a sentence, such as the subject and predicate, and the availability of this information makes it is possible in many applications to refine the search context window. Secondly, base-phrases are the simplest and most stable structural elements of a sentence. Thus, they are regarded as the smallest syntactic units. Lastly, nested mid-phrases are useful for describing distant modifier relations within maximal phrases, which is helpful in certain applications.

The PolyU Treebank provides not only adequate syntactical information but also some semantic information. To achieve this, each phrase is given a syntactic label and sometimes also a label providing semantic information. For example, *"国家航空和宇宙航行局"(NASA)* is a noun phrase and is assigned the label *NP*. Furthermore, in terms of semantics, it is a noun phrase that indicates the name of an organization, so it is given the appropriate additional label, *NT*. The fact that the PolyU Treebank is a "Not-So-Shallow" treebank makes it substantially different from and more useful than other base-phrase only shallow treebanks. The information it provides can be used in language applications to remove ambiguities. Finally, we should point out that in our treebank, the headword of a base-phrase is also annotated.

Principle 4: Large quantities of annotated data with great accuracy

The sizes of existing Chinese treebanks range from 100,000 to 500,000 words. It is an acceptable size for full parsing [Leech and Garside 1996] but not sufficient for lexical-level analysis. With reference to work on the English language, it is our goal to create a treebank of

one million words. A treebank of this size can support the design and training of a shallow parser and be directly used in the collocation extraction and named entity identification work being conducted by authors' research group.

A well-developed treebank must be very accurately annotated. With the goal of reducing annotation errors, we have designed clear and simple annotation guidelines. To avoid inaccuracies arising from automatic parsing, we have performed annotation manually, and post-annotation error and consistency checking have been performed with tools developed by us. Finally, to avoid human errors, some texts are double- and triple-annotated and then compared. This allows makes it easy to identify and correct errors.

3. Annotation Guideline Design

The establishment of annotation guidelines is the first step in treebank development. To ensure high quality output, the guidelines must follow the design principles and must be clear, unambiguous, easy to understand, and easy to follow. The PolyU Treebank guidelines include definitions of (1) syntactical phrase categories, (2) categories of semantic information, and (3) different phrase levels, including maximal phrases, mid-phrases and base-phrases. Because the PolyU Treebank is based on a segmented and POS tagged corpus, the part-of-speech tags in the corpus are used (with only minor modifications for the sake of annotation consistency). Appendix 1 provides a complete list and explanations of the POS tags. These tags will be used in the examples provided in this paper.

Brackets, [and] are used to indicate the left and right boundaries of phrases. The right bracket is appended with syntactic labels in the 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 semantic information, such as *BL*(parallel). For example, a noun phrase with parallel components will be annotated as *[荣誉/n 与/c 尊严/n]NP-BL (honor and dignity)*.

3.1 Defining the syntactical phrase categories

The first level of information for describing phrases is that in the syntactical phrase category. With reference to the works of Penn Chinese Treebank and Sinica Treebank, our guidelines define a total of eight syntactical phrase categories:

NP — Noun phrase. An *NP* is headed by a noun and the header is normally the last noun in the phrase, e.g., *[市场/n 经济/n#]NP (market economy).*

TP — Time phrase. A *TP* consists of continuous time words and is used to indicate a time, e.g., *[早上/t 8 时/t]TP (8:00 in the morning).*

FP — Position phrase. A *FP* is headed by a position word, *f*, and is used to indicate position information, e.g., *[内蒙古/ns东北部/f#]FP (North-east of Inner Mongolia).*

VP — Verb phrase. A *VP* is a phrase headed by a predicate and containing no subject, e.g., *[顺利/a启动/v#]VP-ZZ (successfully start), and [分析/v# 问题/n]VP-SBI (analyze the problem).*

AP — Adjective phrase. The header of an *AP* is an adjective and the whole phrase acts as an adjective in the sentence, e.g.,*[公正/a合理/a#]AP (fair and reasonable).*

DP — Adverb phrase. The header of a *DP* is an adverb, and the whole phrase plays the role of an adverbial role in a sentence, e.g., *[已/d 不再/d#]DP (no longer).*

PP — Preposition phrase. A *PP* is the phrase which begins with a preposition, e.g., *[在/p贵州/ns农村/n]PP (In the countryside of Guizhou Province).*

QP — Quantifier phrase. A *QP* consists of a number and a quantifier. The quantifier acts as the header. Normally, a *QP* is used as the modifier of an *NP* or a *VP*, e.g., *[[数千/m名/q#]QP 士兵/n (several thousand soldiers).*

3.2 Defining semantic information categories

The PolyU Treebank is unique in that it is annotated with semantic labels. A annotation of the *FF* labels is not mandatory. Only those phrases with pre-defined semantic phrase categories are labeled. Semantic information is very useful for some language applications. For example, *山东/ns 烟台/ns 市/n (Yantai City, Shan Dong Province)* and *烟台/ns 大学/n (Yantai University)* are both noun phrases, but the first one is the name of a place and the second that of an organization. Using the semantic information labels *NS* (Name of a place) and *NT* (Name of an organization) allows one to distinguish between these two NPs. This is highly useful in named entity extraction and automatic summarization. The additional semantic labels can be considered a natural byproduct of manual annotation since annotators naturally need to go through the mental process of identifying them. We simply making them available so that such used knowledge are not wasted during annotation.

In the following, we listed the semantic categories.

Semantic information categories for Noun Phrases

NT — Name of an organization, e.g., *[烟台/ns 大学/n]NP-NT (Yantai University).*

NS — Name of a place, e.g., *[江苏省/ns铜山县/ns]NP-NS (Jiangsu Province, Tongshan Country).*

NR — Name of a person, e.g., *[胡/nr 锦涛/nr]NP-NR (Hu Jintao).*

NZ — Other proper noun phrase, e.g., *[诺贝尔/nr奖/n]NP-NZ (The Nobel Prize).*

BL — Juxtaposition structure. A *BL* label indicates that the phrase is made up of two or more parallel components, e.g., *[中国/ns 与/c 南非/ns]NP-BL (China and South Africa).*

FZ — Appositive. An *NP* with *FZ* labels normally has two equivalents, e.g., *[[国家/n主席/n]NP [江/nr 泽民/nr]NR]NP-FZ (the president of China, Jiang Zemin).*

PZ — Noun modifier. A *PZ* is the default semantic structure of an *NP*, e.g., *[美丽/a 的/u 花/n#]NP-PZ (beautiful flower).*

FS — Noun plurals. A *FS* indicates that the last word in a noun phrase is a suffix for noun plurals, e.g., *[朋友/j# 们/k]NP-FS (friends).*

DE — A *DE* construction is a special kind of an *NP* structure in Chinese. It ends with "的"(*DE*) and indicates the absence of the complementation, e.g., *比/v[原先/d预料/v的/u]NP-DE低/a (lower than originally expected).*

SU — A *SU* construction is a special kind of *NP* structure in Chinese. The typical pattern is *所(SUO)+VP+NP*, e.g., *[所/u画/v禽鸟/n#]NP-SU (the birds painted by).*

Semantic information categories for Verb Phrases

SBI — Predicate and its object. A *VP* with the label *SBI* contains of a predicate and an object, e.g., *[打/v# 篮球/n]VP-SBI 是/v 我/r 的/u 爱好/n (playing basketball is my hobby).*

SBU — Complement. The label *SBU* indicates that the second part of the *VP* phrase is the complement modifying the first part of the *VP*, e.g.*[医治/v# 无效/v]VP-SBU (ineffectively treat).*

ZZ — When a *VP* has the label *ZZ*, the verb is the header and other words are its modifiers, e.g., *[[有效/ad 打击/v#]VP-ZZ了/u 敌人/n]VP-SBI (effectively strike the enemy).*

SD — Serial verb constructions. A *SD* indicates that there are serial actions in a *VP* phrase, where the last action is the cardinal action, e.g., *[[审核/v 发放/v]VP-SD 护照/n]VP-SBI (verify and issue the passport).*

BA — A *BA* construction is a special kind of *VP* structure in Chinese. The typical pattern is*把(BA)+NP1 +VP*, e.g., *[把/p[扶贫/vn开发/vn工作/vn]NP-PZ 作为/v#]VP-BA (place the work of poverty reduction and social development as).*

BEI — A *BEI*-construction is a special kind of a *VP* structure in Chinese. The typical patterns are*被(BEI)+ NP+VP and NP+被+VP*, e.g., *商店/n [被/p[责令/v# 停业/vn]VP-SBI]VP-BEI (the shop was ordered to close).*

Semantic information categories for Time Phrases

PO — A point-of-time indicator. The label *PO* indicates that the *TP* carries point-of-time information, e.g., *[7月/t 1日/t]TP-PO (July 1).*

DU — A period-of-time indicator. A *DU* indicates a period of time, e.g., *[今后/t 3/m年/q]TP-DU (following three years).*

Semantic information categories for Prepositional Phrases

YY — Causation information. A *YY* label is used only to modify a *PP* to indicate that the *PP* carries causation information, e.g., *[因/p 饿/a]PP-YY 死亡/v (starved to death).*

DX — Object information. The label *DX* is used to modify a *PP* to indicate object information, e.g., *[向/p [受灾/vn 地区/n]NP]PP-DX (to the disaster area).*

DD — Place information. This is the place indicator of a *PP*, e.g., *[在/p 深圳/ns]PP-DD (in Shenzhen).*

FM — Method information. A *PP* with an *FS* label signals the existence of method information, e.g., *[通过/p [股票/n 上市/v]S]PP-FM (Through the stock market).*

MD — Motivation information. A *PP* with an *MD* label signals the existence of motivation information, e.g., *[为/p 动武/v]PP-MD [找/v 借口/n]VP-SBI (looking for an excuse for war).*

GJ — Tool information. A *GJ* label indicates that a *PP* carries tool information, e.g., *[用/p 公车/n]PP-GJ (using a public-bus).*

SJ — Time information. A *SJ* label indicates that a *PP* carries time information, e.g., *[到/v 目前/t 为止/v]PP-SJ (up to now).*

3.3 Phrase bracketing

Phrases in the PolyU Treebank are divided into three levels: maximal phrases, mid-phrases and base-phrases. The syntactical analysis and annotation of the PolyU Treebank begins with the identification of maximal phrases which define the scope of examination for bracketing.

A *maximal phrase* is a predicate that plays the role a distinct syntactic component of a sentence, realized by the maximum span of its non-overlapping length. Maximal phrases form the backbone of a sentence. The identification of maximal phrases is one of the most difficult steps in the whole process in that annotators have to syntactically analyze sentences and understand their syntactic components even though they have not yet been labeled. The objective of identifying maximal phrases is to separate a sentence into several syntactic components for examination. After maximal phrases are identified, the base-phrases can then be identified within the scope of examination, that is, within each maximal phrase.

A *base-phrase* is defined as a minimum non-nesting phrase with a stable internal structure and independent semantic role. Normally, a base-phrase has a lexical word as its headword. Essentially, a base-phrase must consist of continuous words and contain no nesting components. It never overlaps with other phrases and must be contained within a maximal

phrase. Base-phrases normally conform to a number of typical patterns, such as *[a+n]->NP*, *[a+a]->AP*.

A *mid-phrase* is a nested phrase within a maximal phrase and has a base-phrase as its header. A mid-phrase may contain more than one base-phrase, but only one will be its header. A mid-phrase may have nested components, but none of them may overlap.

The headword of each phrase is also annotated. Further details and examples of phrase bracketing will be provided in Section 4.

4. Implementation of the PolyU Treebank

4.1 Corpus data preparation

The People's Daily corpus, developed by Peking University, consists of more than 13,000 articles and a total of five million words. Since only one million words are required in the PolyU Treebank, we carried out a data selection process. To avoid the duplication of short-lived events and topics, we treated each day's news as a single unit, and we picked six random days in each month from among the six months of data in the entire collection as the raw treebank data.

4.2 Word Segmentation and Part-of-Speech Tagging

In the tasks of the word segmentation and POS tagging of the People's Daily corpus, we were guided by the PSG grammar and "The Grammatical Knowledge-base of Contemporary Chinese" [Yu *et al.* 1998]. The specifications include a total of 43 POS tags. Peking University claimed that the accuracy of word segmentation and POS tagging was higher than 99.9% and 99.5%, respectively [Yu *et al.* 2001].

In this project, we directly used the PKU POS tagging results and made only some notational changes. These changes were made to ensure consistent labeling in our system, where lower cases are used to in word-level tags and upper cases are used in phrase-level labels.

4.3 Phrase Bracketing and Annotation

Identification of Maximal-phrases:

A maximal phrase contains at least one base-phrase and plays a syntactic role in the sentence. Consider the following example sentence:

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

(China Tourism Year is a national-level promotion and marketing activity)

We find that the above sentence has a S-V-O structure. *中国/ns 旅游年/n* is the subject, *是/v* is the predicate, and *一/m 次/q 国家级/b 的/u 宣传/vn 促销/vn 活动/vn* is the object. Clearly there are three syntactic components in this sentence, thus, two separate maximal-phrases, *[中国/ns 旅游年/n]NP (China Tourism Year)* and *[一/m 次/q 国家级/b 的/u 宣传/vn 促销/vn 活动/vn]NP (a national-level promotion and marketing activity)* are annotated. Note that *是/v* is also considered a maximal phrase because it acts as a predicate. However, since it has only one lexical word and is structurally unambiguous, by default, it is not bracketed. Admittedly, *是/v* and *一/m 次/q 国家级/b 的/u 宣传/vn 促销/vn 活动/vn* can be constructed as a *VP*, but we regard this kind of bracketing is more useful for indicating how phrases may be used to construct a sentence. That is to say, this kind of bracketing would take us into the realm of full parsing, which is not our objective. Thus, we choose to bracket them as separate phrases. As a result, the maximal phrase annotation result is

> *[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-PZ.*

Consider another example,

> *富裕/v 起来/v 的/u 当地/a 农民/n 自发/d 地/u 组织/v 了/u 多个/a 业余/a 乐团/n*
>
> (the rich farmers took the initiative to organize several amateur bands) (Example 2)

We can separate this sentence into three components, *富裕/v 起来/v 的/u 当地/a 农民/n* is the subject, *自发/d 地/u 组织/v 了/u* is the predicate, and *多个/a 业余/a 乐团/n* is the object. Thus, this sentence is annotated with three maximal phrases, bracketed and labeled as follows:

> *[富裕/v 起来/v 的/u 当地/a 农民/n#]NP [自发/d 地/u 组织/v# 了/u]VP-ZZ [多个/a 业余/a 乐团/n]NP-PZ*

Most syntactical labels can be used in maximal phrases, except for *AP* (adjective phrase), *DP* (adverb phrase), and *QP* (quantifier phrase). Meanwhile, *NP-NT*, *NT-NS*, *NP-NZ* may only be used to label maximal phrases. These types of phrases do not normally contain nesting components or header words.

Base-phrases Identification:

Base-phrases are identified only within an already-identified maximal phrase, either nesting inside it or overlapping it. Normally a base-phrase contains two-to-four words with one lexical word as its header.

Take the maximal phrase *[一/m 次/q 国家级/b 的/u 宣传/vn 促销/vn 活动/vn]NP-PZ* in **Example 1** as an example, *[一/m 次/q]QP (a) and [宣传/vn 促销/vn 活动/vn#]NP-PZ (promotion and marketing activity)* are base-phrases in this maximal phrase. Thus, the sentence is annotated as follows:

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

As it happens, *[中国/ns 旅游年/n]NP* and *是/v* are also base-phrases, but because they overlap with maximal phrases, they are not further bracketed. Our annotation principle here is that if a base-phrase overlaps with a maximal phrase, it will not be bracketed twice.

It should be pointed out that the identification of base-phrase is the most fundamental and important goal of treebank annotation. The identification of maximal phrases can be thought as the parsing of a clause using a top-down approach. The identification of base-phrase is however, follows bottom-up approach, the object of which is to identify the most basic units within maximal phrases.

Mid-Phrases Identification:

Because other syntactic structures may sometimes exist between base-phrases and maximal phrases, it is useful to identify one more level of syntactic structure within a maximal-phrase, the mid-phrase. This step begins with the examination of a base-phrase. Thus, **Example 1** is further annotated as follows:

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

where, the underlined text contains the additional annotations.

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

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

(*several issues which are worthy of consideration in the development of current enterprise*).

Full annotation would identify four levels of nesting, as shown below, but our system does not include the additional level of bracketing indicated by the underlined annotations as this is beyond our limit of 3 levels.

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

Annotation of Headwords

In our system, a '#' tag is appended to a word to indicate that it is a headword. Here, a headword must be a lexical word (sometimes also called a content word) rather than a function word. In most cases, a headword stays in a fixed position in a base-phrase. For example, the headword of a noun phrase is normally the last noun in the phrase. Thus, it is considered to be in the default position and to need no explicit annotation. For example, in the clause

[美国/ns 科学家/n]NP [绘制/v 出/v]VP-SBU (*the American scientists drafted*),

[绘制/v 出/v] (*drafted*) is a verb phrase, and the headword of the phrase is *绘制/v*, which is not in the default position for a verb phrase headword. 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 in *[美国/ns 科学家/n] (the American scientists*), but since it is in the default position (for the noun phrase *NP*, according to the default grammatical structure, the last noun in the phrase is the headword, and the other components are the modifiers taking the *PZ* label), no explicit annotation is needed.

5. Quality Assurance and Annotation Progress

Our research team is made up of four people from the Hong Kong Polytechnic University (HKPU), two linguists from Beijing Language and Culture University (BLCU), and some research collaborators from Peking University. The annotation work has been carried out by four post-graduate students of languages and computational linguistics from BLCU.

5.1 Quality Assurance

To achieve high quality annotation, guidelines and annotation specifications must be carefully prepared. In the first stage, two linguists from China worked with the team in Hong Kong to prepare annotation guidelines. At this stage, the annotation range of syntactic categories and semantic information categories were also determined. Then, sample annotation was performed in Hong Kong, and the results were summarized to identify some typical patterns for constructing phrases. After that, all the members annotated in duplicates a 60,000-word sample according to the draft specifications. Based on analysis of the results and feedback, the specifications were revised.

In the annotation stage, about 25% of the materials were distributed in identical form to the annotators. When the first pass annotation was finished, the duplicate annotations were compared. Inconsistencies were discussed to identify the most appropriate annotation results. This result was then taken as the ultimate standard (the so called *Gold Standard)* for evaluating inter-annotator accuracy and consistency. The annotators were required to study this Gold Standard and to use it as the basis for correcting mistakes in their own annotations.

Furthermore, a group of checking and evaluating tools were developed. The first tool performs post-annotation checking to ensure that (1) all Part-of-Speech tags are valid, (2) all phrase boundary marks are matched, (3) there are no cross-bracketed phrases, and (4) all the phrase syntactical labels and semantic labels are annotated in the correct format. This tool is effective for removing obvious annotation mistakes.

The most difficult task is to maintain inter-annotator consistency. To assist this work, we developed two tools. A multiple annotation checking tool was developed to compare and evaluate duplicate annotation results. Any mismatches in phrase brackets and labels were detected and manually verified using the tool. Such annotation error cases were used to train the annotators so that they could then manually remove similar annotation errors from their own annotated data. For individual annotated results, we developed a consistency checking tool. This tool first collects all the annotated phrases and their statistics in the treebank, and it then checks in all of the material for annotation consistency. That is, for any word string forming a phrase, the tool checks the whole treebank to see whether the same word string appearing in different places is bracketed and labeled in the same way. Differences that are detected are verified manually. This tool was found to be useful for checking frequently-used phrases.

5.2 Current Project Status

The corpus currently contains 2,639 articles and a total of 1,035,058 segmented Chinese words. The annotators have identified a total of 282,119 bracketed phrases, including nested phrases. **Table 1** provides statistics about the annotated phrases with different *SS* labels

(mandatory syntactic labels). The annotators have also annotated 98,779 phrases for semantic information.

Table 1. Statistic for annotated phrases with different SS labels

NP	VP	AP	DP	TP	FP	PP	QP
138,785	81,846	16,688	2,812	5,216	2,431	25,198	9,143

All of the annotated material in duplicates has been evaluated against the Gold Standard. On average, the precision of phrase bracketing reached 99.5% and that of recall, 99%. The accuracy achieved in the syntactic labeling of correctly bracketed phrases was, on average, 99.8%, while that of semantic labeling was 98.5%. It was more difficult to determine the accuracy of individually annotated data, that is, of data that was only annotated by one person. Our approach was to randomly select a sample consisting of 5% of the material individually annotated by each annotator. We then annotated these samples in duplicates to evaluate the accuracy of the original annotations. The evaluation results showed that the precision achieved in the phrase bracketing of individually annotated data was 98.8%, while that of recall was 98.2%. The accuracy of syntactic labeling was 99.5% and that of semantic labeling was 98.0%.

6. Applications of The PolyU Treebank

The fact that the PolyU Treebank provides not only syntactic but also semantic information of phrases means that it can be applied to a variety of NLP applications. Of course, the most obvious candidate is the training and testing of an automatic shallow parser [Lu *et al.* 2003]. Other applications in which it can be used are Chinese collocation extraction and research on the acquisition of temporal expressions.

In 2003, our team developed an effective window-based statistical algorithm for extracting Chinese collocation which the precision rate of extracted bigram collocation reached 61% [Xu 2003]. The extraction results included some pseudo-collocations, that is, word combinations that frequently co-occurred but were in fact irrelevant, like the typical ‘doctor-nurse’ combination in English [Church and Hanks 1990]. The fact that these pseudo-collocations were statistically significant made it difficult to remove them individually using any statistic-based extraction method. However, given that a Chinese collocation normally occurs only within a phrase or between the headwords of relevant phrases [Zhang and Lin 1992], we were able to use the syntactic information, i.e., the boundaries and headword of phrases, recorded in the PolyU Treebank to refine the searching context window, eliminate some pseudo-collocations, and also retrieve some low-frequency collocations.

The PolyU Treebank is currently being used to acquire temporal expressions. The annotated time phrases (*TP*) and the additional annotation with more finely-tuned

point-of-time (*TP-PO*) and period-of-time (*TP-DU*), are very helpful to acquire and classify temporal expressions.

7. Conclusions and Future Work

This paper has described the design and construction of a manually annotated one-million-word Chinese shallow treebank. This is the first attempt to not only construct a large-scale shallow Treebank for use in practical applications but also provide a treebank for a public use.

The PolyU Treebank has four main advantages:

1. It offers a set of practical, shallow annotation specifications with low ambiguity. These specifications can be used to guide both treebank annotation and the development of an automatic shallow parser.
2. The PolyU Treebank provides useful syntactic information, including the boundaries and syntactic categories of base-phrases, nested phrases, and maximal-phrases. Because it adopts a widely accepted grammar framework and makes use of a widely accepted phrase categories, other researchers can readily use the PolyU Treebank.
3. The PolyU Treebank provides useful semantic information, which is unavailable in other syntactic treebanks.
4. The PolyU Treebank offers a large amount of high-quality data.

Presently, we are developing visualization tools that will support user-friendly keyword searching, context indexing, and annotation case searching. We are also keen to include the annotation of semantic information labels for phrases so as to make the PolyU Treebank more useful in a wider range of research applications. Currently, the PolyU Treebank is being used in research on Chinese collocation extraction, Chinese terminology extraction and summarization, and the acquisition of temporal expressions. In these tasks, the syntactic and semantic knowledge obtained from the PolyU Treebank has been found to improve performance. Finally, we intend to make the PolyU Treebank data available for public access in the hope that the availability of, such a large-scale Chinese shallow Treebank will facilitate NLP research.

Acknowledgement

This project was partially supported by The Hong Kong Polytechnic University (Project Code A-P203) and a CERG Grant (Project code 5087/01E). Special thanks go to Mr. Wei Yan for leading the annotation team at Beijing Language and Culture University and to the anonymous reviewers for their valuable comments, which improves the quality and readability of this paper.

References

Chen, F. Y., P. F. Tsai, K. J. Chen, and C. R. Hunag, "Sinica Treebank," *International Journal of Computational Linguistics and Chinese Language Processing*, 4(2), 1999, pp. 183-204.

Chen K. J., C. R. Huang, F. Y. Chen, C. C. Luo, M. C. Chang, C. J. Chen, and Z. M. Gao, "Sinica Treebank: Design Criteria, Representational Issues and Implementation," *Building and Using Parsed Corpora*, ed. by A. Abeillé, Dordrecht: Kluwer, 2003, pp.231-248.

Church, K., and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, 16(1), 1990, pp. 22-29.

Dalemans W. B. Sabine, and V. Jorn, "Memory-based Shallow Parsing," In *Proceedings of Conference on Computational Natural Language Learning* , 1999, Bergen, pp.53-60.

Leech, G. N., and R. Garside, *Running a Grammar Factory: the Production of Syntactically Analyzed Corpora or "Teebanks"*, Johansson and Stenstron, 1996.

Li, B. L., Q. Lu, and Y. Li., "Building a Chinese Shallow Parsed Treebank for Collocation Extraction," In *Proceedings of Conference on Intelligent Text Processing and Computational Linguistics*, 2003, Mexico City, pp. 402-405.

Lu, Q., J. Zhou, and R. F. Xu, "Machine Learning Approaches for Chinese Shallow Parsing," In *Proceedings of IEEE International Conference on Machine Learning and Cybernetics*, 2003, Xi'an, China, pp.2309-2314.

Sun, H. L., "A Content Chunk Parser for Unrestricted Chinese Text, " PhD Thesis, Peking University, 2001.

Marcus, M. B. Santorini, and M. A. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank," *Computational Linguistics*, 19(1), 1993, pp. 313-330.

Wallis, S., "Completing Parsed Corpora: from Correction to Evolution," *Building and Using Parsed Corpora*, ed. by A. Abeillé , Dordrecht: Kluwer, 2003, pp.61-71.

Xia, F., M. Palmer, N. W. Xue, M. E. Okurowski, J. Kovarik, F. D. Chiou, S. Z. Huang, T. Kroch and M. Marcus, "Developing Guidelines and Ensuring Consistency for Chinese Text Annotation," In *Proceedings of second International Conference on Language Resources and Evaluation*, 2000, Athens, Greece

Xu, R. F., Q. Lu, and Y. Li, "An Automatic Chinese Collocation Extraction Algorithm based on Lexical Statistics," In *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering*, 2003, Beijing, pp.321-326.

Xue, N. W., F. D. Chiou, and M. Palmer,"Building a Large-Scale Annotated Chinese Corpus," In *Proceedings of 17th International Conference on Computational Linguistics*, 2002, Taipei, Taiwan, pp.336-343

Yu, S. W., X. F. Zhu, H. Wang, and Y. Y. Zhang, *The Grammatical Knowledge- base of Contemporary Chinese: A Complete Specification*. Tsinghua University Press, Beijing, China, 1998.

Yu, S. W., et al. “Guideline of People’s Daily Corpus Annotation,” Technical report, Beijing University, 2001.

Zhang, S. K. and X. G. Lin，*Collocation Dictionary of Modern Chinese Lexical Words*, 1st ed., Business Publisher, Beijing, China, 1992.

Appendix 1. Part-of-Speech Tag Set

ag	形容词语素 adjective morpheme	a	形容词 adjective	ad	副形词 adverb-adjective	an	名形词 adnoun
bg	区别语素 distinguish morpheme	b	区别词 distinguish word	c	连词 conjunction	dg	副语素 adverb morpheme
d	副词 adverb	e	叹词 exclamation	f	方位词 position word	h	前缀 heading element
i	成语 Idiom	j	简略语 abbreviation	k	后缀 tail element	l	惯用语 habitual word
mg	数语素 numeral morpheme	m	数词 numeral	ng	名语素 noun morpheme	n	名词 noun
nr	人名 person's name	ns	地名 toponym	nt	组织名 organization noun	nx	外文 foreign character
nz	专有名词 other proper noun	o	拟声词 onomatopoeia	p	介词 preposition	q	量词 quantifier
rg	代语素 pronoun morpheme	r	代词 pronoun	s	方位词 Location word	tg	时语素 time morpheme
T	时间词 time	u	助词 Auxiliary	vg	动语素 verb morpheme	v	动词 verb
vd	副动词 adverb-verb	vn	动名词 gerund	w	符号 punctuation	yg	语气词素 modal morpheme
y	语气词 modal word	z	状态词 state word				