THE COMPILATION OF A SAMPLE PFR CHINESE CORPUS OF SKELETON-PARSED SENTENCES

May Lai-Yin Wong

Department of Linguistics and Modern English Grammar (Lancaster University)

Abstract

The approach taken in this paper for the construction of a treebank is inspired by the skeleton parsing approach. From the PFR Chinese Corpus, a sample text of some 100,000 word tokens was chosen for the production of the treebank. A clear account of the 17 non terminal constituents that are defined and instantiated in the corpus texts will be provided in a parsing scheme. A set of parsing guidelines on practical issues related to map any parses on to sentences in the application of the parsing scheme will also be considered. It is noteworthy also to discuss the major difficulties encountered in the course of skeleton parsing, as this illuminates some of the peculiarities of the Chinese language. The conclusion is an evaluation of the success of the treebank compilation.*

1. Introduction

Treebanks are simply corpora in which syntactic constituent structure is made explicit by a process of corpus annotation (Leech and Garside 1991: 15; Abeillé 2003: xiv). My major concern here is not with software to achieve this annotation automatically (as at the time of writing, there are no effective available parsers designed for the Chinese language), but with the establishment of a parsing scheme and its manual application to written Chinese corpus data. More specifically, the approach taken here is inspired by the skeleton parsing approach (Eyes and Leech 1993; Garside 1993; Black et al. 1996; Leech and Eyes 1997). Skeleton parsing seeks to produce simplified constituent-structure annotations.

2. PFR sample skeleton treebank: text selection

From the PFR Chinese Corpus (Yu 1999), a sample text of some 100,000 word tokens, yielding approximately 2,500 sentences was chosen for the production of my treebank. I contend that a standard block of about one hundred thousand

words is a unit of about the right size for skeleton parsing and handling by a human treebanker. Furthermore, my choice of text was governed by the need to produce a parsed sample corpus of reasonable length which could not only be manageable for hand-parsing but also represent a typical chuck of the data that would eventually be treebanked by me.

3. PFR sample skeleton treebank: parsing scheme

As Sampson (1995: 2ff) puts it, the process of parsing refers to the ability to extract from a linear sequence of words the underlying hierarchical grammatical structure, and thus a parsing scheme "is a set of categories and notational conventions allowing the grammatical properties of a text to be made explicit". In other words, it is a guideline document which helps the human analyst parse sentences (Leech and Garside 1991: 15-16). A clearly defined parsing scheme is essential for the production of a satisfactorily parsed text.

3.1. UCREL skeleton parsing annotation scheme

As most of the existing treebanks are primarily based upon English texts, it does not come as a surprise that the annotation schemes used on those treebanks chiefly reflect the syntactic categories which are directly relevant to English grammar. A case in point is the UCREL skeleton parsing scheme, as illustrated in Table 1.[1]

TABLE 1: **The UCREL skeleton parsing annotation scheme**

UCREL Skeleton Parsing Annotation Scheme	
Fa	Adverbial Clause
Fc	Comparative Clause
Fn	Noun Clause
Fr	Relative Clause
G	Genitive
J	Adjective Phrase (predicative)
N	Noun Phrase
Nr	Adverbial Noun Phrase (temporal)
Nv	Adverbial Noun Phrase (non-temporal) (not in AP or SEC corpora)
P	Prepositional Phrase
S	Sentence (used eg in quoted speech, also with + and & as co-ordinates)
Tg	*-ing* Clause
Ti	Infinitive Clause
Tn	Past Participle Clause
V	Verb Phrase
(null)	Unlabelled Constituent

----

[1] The table was adapted from UCREL's website http://www.comp.lancs.ac.uk/computing/research/ucrel/skeletontags.html.

As stated, some of the syntactic constituent labels in the UCREL skeleton parsing scheme are specially designed to suit English grammar. When I attempted to adopt this scheme wholesale for my own research, I found that some modifications were needed to accommodate the syntactic properties of the Chinese language. Since there is no conclusive morphological evidence that motivates the postulation of infinitival clauses in Chinese (Xue et al. 2000: 32), the non-finite clauses, including the *-ing* clause (Tg), infinitive clause (Ti) and past participle clause (Tn) were wnot taken over from the UCREL parsing scheme to my parsing scheme. Similarly, I did not include noun clause (Fn)[2] and relative clause (Fr)[3] as they are also not compatible with Chinese syntax. Furthermore, owing to the fact that different languages tend to employ different strategies in signalling the same grammatical relations, the parsing label of comparative clause (Fc) was also not adopted in my parsing scheme. While English makes use of a clause to give comparisons, the comparative constructions in Chinese are expressed by means of a prepositional phrase.[4] Lastly, I did not take the genitive (G) as one of the constituent labels in my parsing scheme so as to avoid terminological controversy.[5]

In view of the differences between the English and Chinese grammatical systems, new constituent labels that are not used in the UCREL skeleton parsing scheme had to be invented for the purposes of this research. These are: adverb phrase, correlative clause, adverbial idiom/set phrase, adverbial adjective phrase, adverbial prepositional phrase, adverbial verb phrase and verbal object.

---

[2] Nominal clauses are also difficult in the Chinese linguistics. The notion of the nominal clause may be useful in English as the clauses used as sentential subject or object are formally distinctive from those used independently: they are typically introduced by the conjunction *that* as in "That John stole my book was totally incredible" (Quirk and Greenbaum 1973: 316-322). Chinese clauses or sentences, however, do not vary morphosyntactically when they are used as subject or direct object (Liu 1996: 245, 253). Hence, there is no justification for adopting this grammatical category in a parsing scheme designed for the Chinese language. The lack of previous accounts of nominal clauses in Chinese tends to serve as counterevidence to the notion that nominal clauses exist in Chinese.

[3] Relative clauses have long been a source of controversy in Chinese linguistics. Some scholars (e.g. Li and Thompson 1989: 579ff; Aoun and Li 1993; Chiu 1993; Wu 2000) believe that a nominalisation (whereby a verb, verb phrase, or sentence, followed by the particle *de,* functions as a noun phrase) can be called a relative clause in Chinese if the head noun that it modifies refers to an unspecified element involved in the situation described by the nominalisation. On the other hand, Chao (1968), among others, does not adopt the notion of relative clause in his descriptive grammar of Chinese.

[4] The comparative marker 比 bi "than" and the phrase that immediately follows it form a prepositional phrase which serves as a preverbal adjunct (Zhao 1989; Liu 1999: 204ff). Compare the following contrived sentences from English and Chinese respectively, both of which express the same meaning:

(a) He does the assignment [**Fc** better than I do **Fc**].
(b) 他做的功课做得[**P** 比我 **P**]好。 *Ta zuo de gongke zuo de bi wo hao*

[5] *The genitive constructions in English roughly* correspond to those constructions marked by the particle 的 *de* in Chinese. However, this apparent correspondence is complicated by two issues. Firstly, possessive constructions in Chinese do not necessarily take the particle 的 *de* (Li and Thompson 1989: 115) as in 他爸爸 *ta baba* and 他的爸爸 *ta de baba* which both indicate the same meaning "his father". Secondly, apart from marking possessions, the particle 的 *de* can be an adjectival marker and a marker of nominalisation and explicit modification (Zhu 1982 and 2000). Since the term "genitive" is confusing as far as Chinese linguistics is concerned, I decided to exclude this constituent from my parsing scheme.

3.2. PFR skeleton parsing labels

As Leech and Eyes (1997: 37) note, Sampson's (1995) annotation scheme demonstrates three key components of a clearly specified parsing scheme:

(a) A list of symbols used in the annotation: non-terminals, terminals, and other symbols;
(b) A basic definition of the symbols: e.g. N = noun phrase;
(c) A description, which is as detailed as possible, of how the symbols are actually applied to text sentences. For example, how do annotators recognise a noun phrase when they see one, and how do they distinguish noun phrase tokens from words or word sequences which are not noun phrases?

With respect to the first and second points, in keeping with Sampson, I intend to provide these 3 sets of data for my treebank: (a) the non-terminal labels and (b) their definitions with illustrative examples are given for the PFR treebank in Table 2.[6]

TABLE 2: **The list of constituent labels for the PFR Sample Skeleton Treebank parsing scheme**

Nonterminal Category	Symbol	Example
Adverbial Clause	Fa	<Fa>只要_c <N>我们_r</N> <R>进一步_d</R> <解放思想_i ，_w 实事求是_i> ，_w <V>抓住_v <N>机遇_n</N></V> ，_w <开拓进取_l></Fa> ，_w <Fa>*zhiyao*_c <N>*women*_r</N> <R>*jinyibu*_d</R> <*jiefangsixiang*_i ，_w *shishiqiushi*_i> ，_w <V>*zhuazhu*_v <N>*jiyu*_n</N></V> ，_w <*kaituojinqu*_l></Fa> ，_w "If we become more open-minded and down-to-earth and make every effort to explore new possibilities, … "
Correlative Clause	Fc	<Fc>越_d 走_v 越_d 宽广_a</Fc> <Fc>*yue*_d *zou*_v *yue*_d *kuanguang*_a</Fc> "the more we walk, the broader (the road) will be"
Main Clause (to which the adverbial clause is subordinated)	Fm	<Fm><Fa>只要_c <N>我们_r</N> <R>进一步_d</R> <解放思想_i ，_w 实事求是_i> ，_w <V>抓住_v <N>机遇_n</N></V> ，_w <开拓进取_l></Fa> ，_w <N><V>建设_v <V>有_v <N>中国_ns 特色_n 社会主义_n</N></V></V> 的_u 道路_n</N> 就_c <V>会_v <Fc>越_d 走_v 越_d 宽广_a</Fc></V> 。_w</Fm> <Fm><Fa>*zhiyao*_c <N>*women*_r</N> <R>*jinyibu*_d</R> <*jiefangsixiang*_i ，_w *shishiqiushi*_i> ，_w <V>*zhuazhu*_v <N>*jiyu*_n</N></V> ，_w <*kaituojinqu*_l></Fa> ，_w <N><V>*jianshe*_v <V>*you*_v <N>*Zhongguo*_ns *tese*_n *shehuizhuyi*_n</N></V></V> *de*_u *daolu*_n</N> *jiu*_c <V>*hui*_v <Fc>*yue*_d *zou*_v *yue*_d *kuanguang*_a</Fc></V> 。_w</Fm>

---

[6] All of the examples in this paper are given in the Chinese characters, followed by pinyin romanisations and English translations.

Nonterminal Category	Symbol	Example
		"If we become more open-minded and down-to-earth and make every effort to explore new possibilities, we are in a better position to set up an ideology which can fully represent Chinese characteristics."
Adverbial Idiom/ Set Phrase	Ia	<Ia>坚定不移_i 地_u</Ia> <Ia>*jiandingbuyi* _i *de*_u</Ia> "persistently" <Ia>满怀信心_l 地_u</Ia> <Ia>*manhuaixinxin*_l *de*_u</Ia> "confidently"
Adjective Phrase	J	<J>非常_d 重要_a</J> <J>*feichang*_d *zhongyao*_a</J> "very important"
Adverbial Adjective Phrase	Ja	<Ja>成功_a 地_u</Ja> <Ja>*chenggong*_a *de*_u</Ja> "successfully"
Noun Phrase	N	<N>百年_m 历史_n</N> <N>*bainian*_m *lishi*_n</N> "a hundred years' history"
Adverbial Noun Phrase	Na	<Na>今天_t 上午_t</Na> <Na>*jintian*_t *shangwu*_t</Na> "this morning" <Na>现在_t</Na> <Na>*xianzai*_t</Na> "at present"
Prepositional Phrase	P	<P>对_p <N>香港_ns</N></P> <P>*dui*_p <N>*Xianggang*_ns</N></P> "to Hong Kong"
Adverbial Prepositional Phrase	Pa	<Pa>在_p <N>这_r 一_m 年_q</N></Pa> <Pa>*zai* _p <N>*zhe*_r *yi*_m *nian*_q</N></Pa> "in this year"
Adverb Phrase	R	<R>还_d 不_d</R> <R>*hai*_d *bu*_d</R> "not…though"
Sentence (including direct speech quotation, also with & and + as co-ordinates)	S	<S N=»1»><Pa>在_p <N>这_r 一_m 年_q 中</N>_f</Pa>，_w <N>中国_ns 的_u 外交_n 工作_vn</N> <V>取得_v 了_u <N>重要_a 成果_n</N></V> 。_w</S> <S N="1"><Pa>*zai*_p <N>*zhe*_r *yi*_m *nian*_q *zhong*</N>_f</Pa> ，_w <N>*Zhongguo*_ns *de*_u *waijiao*_n *gongzuo*_vn</N> <V>*qude*_v *le*_u <N>*zhongyao*_a *chengguo*_n</N></V> 。_w</S> "In this year, the Chinese government has gained great success in its diplomacy."

Nonterminal Category	Symbol	Example
Verb Phrase	V	<V>展望_v <N>新_a 的_u 世纪_n</N></V> <V>*zhanwang*_v <N>*xin*_a *de*_u *sheji*_n</N></V> "have hope in the new era"
Adverbial Verb Phrase	Va	<Va>迈向_v <N><充满_v 希望_n 的_u> １９９８年_t 之际_f</N></Va> <Va>*maixiang*_v <N><*chongman*_v *xiwang*_n *de*_u> １９９８ *nian*_t *zhiji*_f</N></Va> "at the moment we are looking forward to the prosperous year of 1998" <Va>退休_v 前_f</Va> <Va>*tuixiu*_v *qian*_f</Va> "before retirement"
Verbal Object	Vo	<V>希望_v <Vo>依靠_v <N>大家_r</N></Vo></V> <V>*xiwang*_v <Vo>*yikao*_v <N>*dajia*_r</N></Vo></V> "wish to rely on you"
Initial Conjunct	&	<N&>中国_ns 改革_v</N&> 和_c <N&>*Zhongguo*_ns *gaige*_v</N&> *he*_c "China's revolution"
Non-initial Conjunct	+	和_c <N+>发展_v 的_u 全局_n</N+> *he*_c <N+>*fazhan*_v *de*_u *quanju*_n</N+> "and the entire development"

4. Guidelines of skeleton parsing

Having reviewed my annotation scheme in some depth, I can now present my guidelines for annotation, in keeping with my desire to match the advantages of Sampson's susanne scheme (see section 4.2). The parsing scheme matches features (a) and (b) of a clear and explicit parsing scheme (Leech and Eyes 1997: 37). Nonetheless, an annotation scheme is more than (a) and (b) above. Feature (c), a set of parsing guidelines should also be provided in order to explain how the parsing symbols are actually applied to text sentences to avoid undesirable inconsistency.

During the course of annotation, as more data was analysed, the guidelines took shape and were recorded and updated. Hence, after I finished the task of parsing a sample text taken from the PFR Chinese Corpus, I had a set of guidelines ready to be consolidated into a document to be available to users of the treebank and to future annotators who might want to adopt the same scheme. It is advisable, as Kahrel et al. (1997: 241ff) note, to document explicitly all of the decisions taken in the development of an annotation scheme, as well as its application so that future users can apply the scheme in a manner consistent with that of the originators of the scheme. The decisions were then adhered to consistently in the annotation of similar cases thereafter. In the following subsections, I will discuss the issues that arose and illustrate them with examples drawn from the sample treebank.

4.1. Underspecification — Use of unlabelled bracketings

Brackets may be left unlabelled in cases where a particular grouped sequence of words cannot fit into any of the existing phrase or clause categories. Examples of constituents enclosed in unlabelled brackets are given below from (a) to (e).

(a) Multi-word premodifiers of noun phrases marked by the particle 的 *de* (see also section 5.2.3):

e.g. <N><全国_n 各族_r 的_u> 人民_n</N> <N><*quanguo*_n *gezu*_r *de*_u> *renmin*_n</N> "people from different ethnic groups throughout the country";

e.g. <N><同_p 联合国_nt 和_c 其他_r 国际_n 组织_n 的_u> 协调_vn</N> <N><*tong*_p *Lianheguo*_nt *he*_c *qita*_r *guoji*_n *zuzhi*_n *de*_u> *xietiao*_vn</N> "the compromise between the United Nations and other international organisations";

e.g. <N><最近_t 一个_m 时期_n 一些_m 国家_n 和_c 地区_n 发生_v 的_u> 金融_n 风波_n</N> <N><*zuijin*_t *yige*_m *shiqi*_n *yixie*_m *guojia*_n *he*_c *diqu*_n *fasheng*_v *de*_u> *jinyong*_n *fengbo*_n</N> "the recent financial crises happened in some countries and districts";

(b) Serial verb constructions which are used as if they were compound verbs (see also section 5.2.4):

e.g. <坚持_v 奉行_v> <*jianchi*_v *fengxing*_v> "insist on following";
e.g. <指挥_v 演奏_v> <*zhihui*_v *yanzou*_v> "lead and perform";
e.g. <看望_v 慰问_v> <*kanwang*_v *weiwen*_v> "visit and send regards to …".

(c) Serial adjective constructions:

e.g. <团结_a 一致_a> <*tuanjie*_a *yizhi*_a> "be united together";
e.g. <圆满_a 成功_a> <*yuanman*_a *chenggong*_a> "perfectly successful".

(d) Idioms/set phrases which are used idiosyncratically as if they were single-word nouns or verbs (see also section 5.2.2):

e.g. <大势所趋_i ，_w 民心所向_l> <*dashisuoqu*_i ，_w *minxinsuoxiang*_l> "urged by the trend, supported by general public";

e.g. <大气磅礴_i ，_w 波澜壮阔_i> <*diqibangbo*_i ，_w *bolanzhuangkuo*_i> "powerful wind, fierce waves";

e.g. <流光溢彩_l ，_w 火树银花_i> <*liuguangyicai*_l ，_w *huoshuyinhua*_i> "filled with colourful lights, magnificent".

(e) Coordinated verbs with shared direct object:

e.g. <V><学习_v 和_c 掌握_v> <N>党_n 的_u 十五大_j 精神_n</N></V> <V><*xuexi*_v *he*_c *zhangwo*_v> <N>*dang*_n *de*_u *shiwuda*_j *jingshen*_n</N></V> "learn and master the Communist Party's 15 principles";

e.g. <V><尊重_v 、_w 认识_v 和_c 掌握_v> <N>客观_a 规律_n</N></V> <V><*zhuanzhong*_v 、_w *renshi*_v *he*_c *zhangwo*_v> <N>*keguan*_a *guilü*_n</N></V> "respect, understand and master what we learn in our daily life".

4.2. Bracketing of multi-word constituents

The unlabelled bracketing facility evidently has its uses in skeleton parsing as it allows analysis to proceed where labelling decisions are not obvious or straightforward. Nevertheless, for some multi-word adverb phrases containing two adverbs (e.g. <R>还_d 不_d</R> <R>*hai*_d *bu*_d</R> "not…though"; <R>永远_d 不再_d</R> <R>*yongyuan*_d *buzai*_d</R> "never forever"; <R>一直_d 都_d</R> <R>*yizhi*_d *dou*_d</R> "constantly"), and multi-word attributive adjectival phrases containing an adjective premodified by at least one adverb (e.g. <J>非常_d 重要_a 的_u</J> <J>*feichang*_d *zhongyao*_a *de*_u</J> "very important"; <J>很_d 不_d 平凡_a 的_u</J> <J>*hen*_d *bu*_d *pingfan*_a *de*_u</J> "very extraordinary; <J>十分_m 高兴_a</J> <J>*shifen*_m *gaoxing*_a</J> "very happy"), though Eyes and Leech (1993: 53) chose to put them into unlabelled brackets, they were labelled in my treebank. The reason for this is that their internal structure is clear, having a head (adjective or adverb) being modified by another adverb.

4.3. Bracketing of single-word constituents

As suggested in the EAGLES Recommendations for the Syntactic Annotation of Corpora, Version of 11th March 1996 (Leech et al. 1996), it is considered preferable to bracket single-word constituents where they show their phrasal status by the possibility of adding modifiers or replacing them by a multi-word phrase as in example 1, or where they are in coordination with other multi-word constituents as in example 2.

(1)　<N>人民_n 生活_vn</N> **<R>**进一步_d**</R>** **<V>**改善_v**</V>**
　　<N>*renmin*_n *shenghuo*_vn</N> **<R>***jinyibu*_d**</R>** **<V>***gaishan*_v**</V>**
　　"the life of the citizens is further improved"

(2)　**<N><N&>**全党_n**</N&>** 和_c **<N+>**全国_n 各族_r 人民_n**</N+></N>**
　　<N>**<N&>***quandang*_n**</N&>** *he*_c **<N+>***quanguo*_n *gezu*_r *renmin*_n**</N+></N>**
　　"the Communist Party and the citizens of varied ethnic groups throughout the country"

4.4. Punctuation

Generally speaking, I included punctuation within the bracketing. As for phrase/sentence-initial and phrase/sentence-final punctuations, I enclosed them within the parsing bracketing, as in example 3:

(3)　<P>为_p **<N>**"_w 两手抓_l 、_w 两手_m 都_d 要_v 硬_a "_w**</N>**</P> <V>提供_v 了_u <N>新_a 的_u 理论_n 根据_n</N></V>
　　<P>*wei*_p **<N>**"_w *liangshouzhua*_l 、_w *liangshou*_m *dou*_d *yao*_v *ying*_a "_w**</N>**</P> <V>*tigong*_v *le*_u <N>*xin*_a *de*_u *lilun*_n *genju*_n</N></V>
　　"provide new theoretical evidence to the principle of perseverance in an undertaking"

As regards medial punctuation marks, typically commas, I attached them to the highest available node in the parse tree, thus these punctuation marks can be used as delimiters of major constituents, as in example 4:

(4) <S N=”5”>**<S&>**<N>[中国_ns 政府_n]nt</N> <Ja>顺利_ad</Ja> <V>恢复_v <P>对_p <N>香港_ns</N></P> <Vo>行使_v <N>主权_n</N></Vo></V>**</S&>** ，_w 并_c **<S+>**<P>按照_p <N><“_w 一国两制_j ”_w 、_w “_w 港人治港_l ”_w 、_w 高度_d 自治_v 的_u> 方针_n</N></P> <V>保持_v <N>香港_ns 的_u 繁荣_an 稳定_an</N></V>**</S+>** 。_w</S>

<S N=”5”>**<S&>**<N>[*Zhongguo*_ns *zhengfu*_n]nt</N> <Ja>*shunli*_ad</Ja> <V>*huifu*_v <P>*dui*_p <N>*Xianggang*_ns</N></P> <Vo>*xingshi*_v <N>*zhuquan*_n</N></Vo></V>**</S&>** ，_w *bing*_c **<S+>**<P>*anzhao*_p <N><“_w *yiguoliangzhi*_j ”_w 、_w “_w *gangrenzhigang*_l ”_w 、_w *gaodu*_d *zizhi*_v *de*_u> *fangzhen*_n</N></P> <V>*baochi*_v <N>*Xianggang*_ns *de*_u *fanrong*_an *wending*_an</N></V>**</S+>** 。_w</S>

"The Chinese government has succeeded in resuming its sovereignty over Hong Kong and maintaining the prosperity and stability of Hong Kong, according to the objectives on 'one country two systems', 'Hong Kong people ruling themselvels' and 'high degree of independence'."

4.5. Ambiguity

Linguistic forms are often ambiguous. My annotation scheme, however, did not contain any notation for representing ambiguity explicitly with which the human analyst selects one possible sense for a form and represents it. I decided not to explicitly mark an ambiguous form because even if a given item has more than one reading, the human analyst will not recognise this in the course of parsing and just annotate the item with the interpretation that seems initially most plausible.

With such a detailed and carefully articulated parsing scheme and guidelines, I can now proceed to describe the actual process of skeleton parsing and difficulties encountered in the process of parsing.

5. The process of skeleton parsing
5.1. The basic concept of skeleton parsing

The basic idea of skeleton parsing, as Garside and McEnery (1993: 19) demonstrate, is that the treebanker marks only those syntactic structures which seem "intuitively obvious", rather than keeping track of a particular reference grammar. In the course of skeleton parsing, I inserted a nested set of brackets around a sequence of word tokens which appeared to be intuitively correct to group as a single unit. I then assigned to each of these units (i.e. sentence constituents) a label from the set of categories specified in my parsing scheme. An excerpt of the PFR Skeleton Treebank is provided in Figure 1.

<ORIGINALMP>19980101-01-001-013_m</ORIGINALMP> <S N=" 28" ><V>环顾_v <N>全
球_n</N></V> ，_w <N><N>日益_d 密切_a 的_u> 世界_n 经济_n 联系_vn</N> ，_w <N>日新
月异_i 的_u 科技_n 进步_vn</N> ，_w <R>正在_d</R><P>为_p <N><各国_r 经济_n 的_u> 发
展_vn</N></P> <V>提供_v <N>历史_n 机遇_n</N></V> 。_w</S> <S N=" 29" >但是_c ，_w
<N>世界_n</N> <R>还_d 不_d</R> <J>安宁_a</J> 。_w</S> <S N=" 30" ><N><南北_n 之
间_f 的_u> 贫富_n 差距_n</N> <V>继续_v <Vo>扩大_v</Vo></V> ；_w <N>局部_b 冲突_
vn</N> <时有发生_l> ；_w <N><<不_d 公正_a 不_d 合理_a 的_u> 旧_a 的_u> 国际_n 政治_n
经济_n 秩序_n</N> <R>还_d</R> <V>没有_v <N>根本_a 改变_vn</N></V> ；_w <N>发展中
国家_l</N> <P>在_p <N>激烈_a 的_u 国际_n 经济_n 竞争_vn 中_f</N></P> <R>仍_d</R>
<V>处于_v <N>弱势_n 地位_n</N></V> ；_w <N>人类_n 的_u <N><N&>生存_vn</N&> 与_
c <N+>发展_vn</N+></N></N> <R>还_d</R> <V>面临_v <N>种种_q <N><N&>威胁_vn</
N&> 和_c <N+>挑战_vn</N+></N></N></V> 。_w</S> <S N=" 31" ><N>和平_n 与_c 发展_
vn 的_u> 前景_n</N> <V>是_v <J>光明_a 的_u</J></V> ，_w <N>２１_m 世纪_n</N>
<R>将_d</R> <V>是_v <N>充满_v 希望_n 的_u> 世纪_n</N></V> 。_w</S> <S
N=" 32" >但_c <N>前进_v 的_u 道路_n</N> <V><V><V&>不会_v</V&> <R>也_d</R>
<V+><R>不_d</R> 可能_v</V+></V> <一帆风顺_i></V> ，_w <N>关键_n</N> <V>是_v
<S><N>世界_n 各国_r 人民_n</N> <V>要_v <J>进一步_d 团结_a 起来_v</J></V></S></V>
，_w <V><V><V&><R>共同_d</R> 推动_v</V&> <V+><R>早日_d</R> 建立_v</V+></V>
<N><公正_a 合理_a 的_u> 国际_n 政治_n 经济_n 新_a 秩序_n</N></V> 。_w</S>

FIGURE 1: **An excerpt of the PFR Sample Skeleton Treebank**

5.2. Difficulties in skeleton parsing Chinese text

It is noteworthy here to discuss the major difficulties that I encountered in the course of skeleton parsing a sample text taken from my corpus, as this illuminates some of the peculiarities of the Chinese language.

5.2.1. Ba *constructions*

Firstly, 把 *ba* constructions make the parse of a verb phrase incomplete. The *ba* construction is a widely discussed topic in the grammar of Chinese (see, for instance, Li and Thompson 1989: 463-491; Chen 1990; Kit 1992; Zou 1993; Xia and Wu 1996; Li 1997; Xue et al. 2000; Li 2001). In general, the structure of the *ba* construction is expressed and underlined in example 5: a *ba* sentence has a subject, followed by *ba* and the *ba* noun phrase (i.e. the NP directly following *ba*) followed by a verb.

(5)   *subject ba NP verb*

In my PFR treebank, 104 instances of the *ba* construction were found, which fall into three types of this construction. The general pattern of *ba* sentences is to place the direct object of the following verb immediately after *ba* as in example 6.

(6)   <P>把_p **<N$_i$>**电厂_n **</N$_i$>**</P> <V>建_v 好_a</V> **Ø$_i$**
      <P>*ba*_p **<N$_i$>**dianchang_n **</N$_i$>**</P> <V>*jian*_v *hao*_a</V> **Ø$_i$**
      "build an electricity supply station"

Typically, a transitive verb should follow the pattern "<V>…<N>…</N></V>". However, the verb, 建_v 好_a, *jian-hao*, "build well, build in good shape" that

follows the *ba* construction lacks a direct object, which is equivalent to the *ba* noun phrase, i.e. *ba* NP$_i$ verb $\emptyset_i$, where $\emptyset_i$ denotes the empty position of the preposed verb object that shares the same reference as the prepositional complement of *ba*.[7] In other words, the structure of the verb phrase following the *ba* construction is in the form of <V>...</V> rather than the canonical form mentioned before.

More complicated *ba* constructions involve the occurrence of two following verbs and a passivised verb. In cases like example 7 where two different transitive verbs follow the *ba* construction, it is not immediately obvious whether that the prepositional complement of *ba* co-refers to the object of the first verb (改编 *gaibian* "be adapted for") or that of the second one (为 *wei* "be changed as"). Since the second verb already takes a direct object (器乐曲 *qiyuequ* "acoustics of musical instrument"), the *ba* noun phrase must be co-referential with the object of the first verb.

(7) <P>把_p **<N$_i$>**<广大_b 听众_n 耳熟能详_i 的_u> 歌曲_n**</N$_i$>**</P>
<V>改编_v **$\emptyset_i$** <V>为_v <N>器乐曲_n</N></V></V>
<P>*ba*_p **<N$_i$>**<*guangda*_b *tingzhong*_n *ershunengxiang*_i 的_u> *gequ*_n**</N$_i$>**</P> <V>*gaibian*_v **$\emptyset_i$** <V>*wei*_v <N>*qiyuequ*_n</N></V></V>
"change those popular songs into acoustics of musical instrument"

In less obvious cases like example 8, however, it is impossible to locate any empty position that co-refers to the *ba* complement. The verb 带入 *dairu* "bring to" that follows the *ba* construction is used causatively without any visible passivisation. As Norman (1988: 164) notes, Chinese verbs do not make any distinction between the active (or unaccusative) and passive (or causative). The *ba* noun phrase 人们 *renmen* "people" therefore actually refers to the logical subject of the verb.

(8) <P>把_p <N>人们_n</N></P> <V>带入_v <N>迷人_a 的_u 艺术_n 境地_n</N></V>
<P>*ba*_p <N>*renmen*_n</N></P> <V>*dairu*_v <N>*miren*_a *de*_u *yishu*_n *jingdi*_n</N></V>
"bring people into a fascinating imaginary place"

5.2.2. Idioms or set phrases

The use of idioms (tagged "i") or set phrases (tagged "l") as if they were nouns and verbs is also problematic. Noun-like idioms and set phrases are illustrated in example 9 and verb-like set phrases in example 10. To my knowledge, the grammatical categories of this kind of idiomatic expressions have not been documented so far.

(9) <N>今晚_t 的_u 长安街_ns</N> <流光溢彩_l ，_w 火树银花_i>
<N>*jinwan*_t *de*_u *Changanjie*_ns</N> <*liuguangyicai*_l ，_w *huoshuyinhua*_i>
"Tonight the Changan Street was filled with colourful lights and really looked magnificent."

---

[7] 建_*v* 好_*a, jian-hao* "build well, build in good shape" is a compound verb. More specifically, it is a verb-complement (V-R) compound (Chao 1968: 435ff). The resultative complement 好 *hao* "good" is bound to and follows the verb 建 *jian* "build" and expresses the result of the action of the verb.

(10) <N>国民经济_n</N> <稳中求进_l>
      <N>*guominjingji*_n</N> <*wenzhongqiujin*_l>
      "The national economy is progressing steadily."

That they can be used rather idiosyncratically as a noun or a verb makes it almost impossible for even a human analyst to determine the phrasal category of a given idiomatic expression: whether it is a noun phrase or a verb phrase. As in the above two examples, it is unclear whether the idiom/set phrase placed after the subject noun phrase is intended to function as a nominal expression or a verbal one. Unlike English, in which the subject must be followed by a verbal predicate, a Chinese predicate can be a verbal predicate, an adjectival predicate or a nominal predicate (Chao 1968: 90). In the absence of further evidence of the categorial status of such segments, those idioms and set phrases occurring in the predicate position were left unlabelled in my treebank.

5.2.3. *Lengthy premodifiers of a noun phrase*

Unlike English, which favours the use of postmodification if a modifier of a noun phrase is long (Quirk and Greenbaum 1973: 425; de Haan 1991), Chinese prefers premodification to postmodification, regardless of the length of the modifier (cf. Liu 1996: 265-274). It is thus common in the PFR treebank that a noun is qualified by a grammatical unit of over six words which is marked by the particle 的 *de* at the end, as in example 11. The particle 的 *de* is traditionally treated as a marker of modification (Chao 1968: 285).

(11) <N><中国_ns 与_p 周边_n 国家_n 和_c 广大_b 发展中国家_l 的_u>
      友好_a 合作_vn</N>
      <N><*Zhongguo*_ns *yu*_p *zhoubian*_n *guojia*_n *he*_c *guangda*_b *fazhan-zhongguojia*_l *de*_u> *youhao*_a *hezuo*_vn</N>
      "the co-operation between China and her surrounding countries and developing countries"

These lengthy premodifiers make the structure of the noun phrase in which they occur extremely difficult to interpret. Some premodifiers of this sort are complicated by the fact that they are further modified by another element marked by *de* in their internal structure, as in example 12.

(12) <N>党_n 的_u <<基本_a 路线_n 提出_v 的_u> 党_n 在_p 社会主义_
      n 初级_b 阶段_n 经济_n 、_w 政治_n 、_w 文化_n 的_u> 基本_a 纲
      领_n</N>
      <N>*dang*_n *de*_u <<*jiben*_a *luxian*_n *tichu*_v *de*_u> *dang*_n *zai*_p
      *shehuizhuyi*_n *chuji*_b *jieduan*_n *jingji*_n 、_w *zhengzhi*_n 、_w *wenhua*_n
      *de*_u> *jiben*_a *gangling*_n</N>
      "the primary principles of the Communist Party on economy, politics and culture, which are also on a par with the Party's basic directions"

5.2.4. *Serial verb constructions*

Serial verb constructions in Chinese also increase the complexity of parsing. There is an immense literature on Chinese serial verb constructions (see, for

instance, Li and Thompson 1989: 594ff; Lin and Soo 1994; Liu 1996). Generally speaking, a serial verb construction refers to a succession of two or more actions that share the same subject, as illustrated in the following concocted example.

(13) <N>我</N> <V>去 <N>朋友 家</N></V> <V>吃 <N>晚饭</N></V>
      <N>*wo*</N> <V>*qu* <N>*pengyou jia*</N></V> <V>*chi* <N>*wanfan*</N></V>
      "I went to my friend's house to have dinner."

However, some of the serial verb constructions in my treebank do not conform to this general pattern of two successive verbs, each of which has a different direct object. Unlike ordinary serial verbs, the serial verbs, as shown in examples 14 and 15, do not take a direct object separately. They are more like compound verbs than serial verbs, though it is not clear that they can be fully assimilated to the former category. Evidence in support of this analysis comes from the fact that these verbs (i.e. 指挥_v 演奏_v *zhihui yanzou* "lead and perform" as in example 14, and坚持_v 奉行_v *jianchi fengxing* "insist and follow" as in example 15), functioning as if they were a single unit, take the same object, i.e. the following noun phrase.

(14) <V><指挥_v 演奏_v 了_u <N>一_m 批_q 中外_j 名曲_n</N></V>
      <V><*zhihui*_v *yanzou*_v> *le*_u <N>*yi*_m *pi*_q *zhongwai*_j *mingqu*_n</N></V>
      "led and performed a variety of Chinese and western popular songs"
(15) <坚持_v 奉行_v> <N>独立自主_l 的_u 和平_n 外交_n 政策_n</N>
      <*jianchi*_v *fengxing*_v> <N>*dulizizhu*_l *de*_u *heping*_n *waijiao*_n *zhengce*_n</N>
      "insist on adopting an independent diplomatic policy in maintaining peace"

Besides sharing the same direct object, another clue that tends to prove that the two verbs are actually used as a compound verb is the suffixation of the morpheme 了 -*le*, as highlighted in example 14. The verbal -*le* has generally been taken as an aspect marker, indicating completion (Norman 1988: 163; Xiao 2002), and it is attached to verbs and not to the objects of verbs (Chao 1968: 247), excluding the possibility that the first verb takes the second verb (and the following noun phrase) as its object. Further research on clarifying their subcategorisation (whether they are serial or compound verbs) ought to be done in order to give a more precise parse.

6. Conclusion: Quality control of the skeleton parsing process

In evaluating the success of an annotation project, Eyes and Leech (1993: 37-42) provide six essential criteria that can be used for evaluating my skeleton parsing scheme.

1. Consensual categories: The linguistic categories that were employed in my parsing scheme have been demonstrated, by comparison to seven syntactic theories, to represent grammatical features largely agreed upon by linguists, rather than features which are theory-specific or deeply controversial.
2. Overall coverage: My sample treebank represents a reasonable length of text (comprising about 100,000 word tokens or 2,500 sentences) to be manually parsed and could be re-used in future research.

3. Productivity: Productivity was satisfactory with the simplified syntactic analysis provided by skeleton parsing.

4. Accuracy: The output of the parsed sentences was cross-checked by several posteditors with a background in linguistics. While one can never guarantee 100% accuracy, I believe the sample treebank to be highly accurate.

5. Uniformity of analysis: To demonstrate consistency of analysis, a concordance of the verb 要 *yao* "need" was drawn from my skeleton treebank. This verb always takes a verbal object, i.e. a verb functioning as the direct object of another verb, which is represented as *Vo* in my parsing scheme and is distinct from *V*, which stands for an independent verb phrase (see Table 2 for a description of the symbols *Vo* and *V*). There are 252 instances of the verb *yao* in my treebank. In each case, it is followed by a verbal object consistently marked as *Vo* not *V*, as highlighted in Figure 2.[8]

<V>注定_v <Vo>要_v <Vo>失败_v</Vo></Vo></V>

<V>要_v <J>进一步_d 团结_a 起来_v</J></V>

<V>要_v <Ja>更_d 好_a 地_u</Ja> <Vo>坚持_v <N><解放思想_i 、_w 实事求是_i 的_u> 思想_n 路线_n</N></Vo></V>

<V>要_v <Ja>更_d 好_a 地_u</Ja> <Vo>坚持_v <P以_p <N>经济_n 建设_vn</N></P> <Vo>为_v <N>中心_n</N></Vo></Vo></V>

<V>要_v <Vo>看_v <Vo>能否_v <P>把_p <N>经济_n 工作_vn</N></P> 搞_v 上去_v</Vo></Vo></V>

<V>要_v <Ja>切实_ad</Ja> <P>把_p <N>精力_n</N></P> <Vo>集中_v <Vo>到_v <N><贯彻_v 落实_v 好_a 中央_n 关于_p 今年_t 经济_n 工作_vn 的_u> <N><N&>总体_n 要求_n</N&> 和_c <N+>各项_r 重要_a 任务_n</N+> 上_f 来_v</N></N></Vo></Vo></V>

<V>要_v <Ja>更_d 好_a 地_u</Ja> <Vo>坚持_v <N> "_w 两手抓_l 、_w 两手_m 都_d 要_v 硬_a "_w 的_u> 方针_n</N></Vo></V>

<V>要_v <Ja>更_d 好_a 地_u</Ja> <Vo>发扬_v <N>求真务实_l 、_w <密切_ad 联系_v 群众_n 的_u> 作风_n</N></Vo></V>

<V>要_v <Vo>发扬_v <N>这样_r 的_u 好_a 作风_n</N></Vo></V>

<V>要_v <Vo>尊重_v <N>群众_n 的_u 意愿_n</N></Vo></V>

<V>要_v <R>大力_d</R> <Vo>倡导_v <N><说实话_l 、_w 办_v 实事_n 、_w 鼓_v 实劲_n 、_w 讲_v 实效_n 的_u> 作风_n</N></Vo></V>

<V>要_v <Vo>坚守_v <N>岗位_</N></Vo></V>

<V>要_v <Vo>让_v <S><N>他们_r</N> <Ja>健康_ad</Ja> <V>成长_v</V></S></Vo></V>

FIGURE 2: **A concordance of the verb *yao***

---

[8] The verb *yao* can take an adjectival direct object, as shown in one instance of the concordance extracted. A vast majority of adjectives may function as verbs in Mandarin Chinese by taking aspect markers (e.g. -*guo* "experiential aspect", -*le* "perfective aspect", etc.) or directional complements (e.g. *qilai* "inchoative"). See Li and Thompson (1989: 141-147).

6. Linguistic validity: One of the aims of carrying out a skeleton parsing on a sample text of the PFR Chinese Corpus is to gain a better understanding of how to precisely locate adverbial clauses in a piece of POS tagged text. Further research will be conducted into adverbial clauses in written Chinese.

References

Abeillé, A. (ed.), 2003, *Treebanks: Building and Using Parsed Corpora*. (Text, Speech and Language Technology, vol. 20.) Dordrecht, Boston and London: Kluwer Academic Publishers.

Aoun, J. and A. Y.-h. Li, 1993, *Syntax of Scope*. Cambridge: MIT Press.

Black, E., S. Eubank, H. Kashioka, D. Magerman, R. Garside and G. Leech, 1996, "Beyond skeleton parsing: producing a comprehensive large-scale general-English treebank with full grammatical analysis". *Proceedings of the 16th International Conference on Computational Linguistics (COLING '96),* 107-12. Copenhagen.

Chao, Y., 1968, *A Grammar of Spoken Chinese*. Berkeley and Los Angeles: University of California Press.

Chen, H.-H., 1990, "A logic-based government-binding parser for Mandarin Chinese". *Proceedings of the 13th International Conference on Computational Linguistics (COLING '90),* ed. H. Karlgren, 48-53. Helsinki, Finland.

Chiu, H-ch., 1993, *The inflectional structure of Mandarin Chinese*. Unpublished Ph.D. thesis, University of Los Angeles.

de Haan, P., 1991, "On the exploration of corpus data by means of problem-oriented tagging: postmodifying clauses in the English noun phrase". *English Computer Corpora: Selected Papers and Research Guide,* ed. S. Johansson and A.-B. Stenström, 52-65. Berlin: Mouton de Gruyter.

Eyes, E. and G. Leech, 1993, "Syntactic annotation: linguistic aspects of grammatical tagging and skeleton parsing". *Statistically-driven Computer Grammars of English: The IBM/Lancaster Approach,* ed. E. Black, R. Garside and G. Leech, 36-61. Amsterdam: Rodopi.

Garside, R., 1993, "The large-scale production of syntactically analysed corpora", *Literary and Linguistic Computing* 8: 39-46.

— and A. McEnery, 1993, "Treebanking: the compilation of a corpus of skeleton-parsed sentences". *Statistically-driven Computer Grammars of English: The IBM/Lancaster Approach,* ed. E. Black, R. Garside and G. Leech, 17-35. Amsterdam: Rodopi.

Kahrel, P., R. Barnett and G. Leech, 1997, "Towards cross-linguistic standards or guidelines for the annotation of corpora". *Corpus Annotation: Linguistics Information from Computer Text Corpora,* ed. R. Garside, G. Leech and A. McEnery, 231-42. London and New York: Longman.

Kit, Ch., 1992, "Parsing Chinese *ba* and *bei* constructions - an LFG approach". *Proceedings of the 3rd International Conference on Chinese Information Processing,* 12-6. Beijing, China.

Leech, G. and E. Eyes, 1997, "Syntactic annotation: treebanks". *Corpus Annotation: Linguistics Information from Computer Text Corpora,* ed. R. Garside, G. Leech and A. McEnery, 34-52. London and New York: Longman.

Leech, G. and R. Garside, 1991, "Running a grammar factory: the production of syn-
    tactically analysed corpora or 'treebanks'". *English Computer Corpora: Selected Papers
    and Research Guide,* ed. S. Johansson and A.-B. Stenström, 15-32. Berlin: Mouton de
    Gruyter.
—, R. Barnett, and P. Kahrel, 1996, "Recommendations for the Syntactic Annotation of
    Corpora". EAGLES Document EAG-TCWG-SASG/1.8 Version of 11th March 1996.
    Available on-line at http://www.ilc.pi.cnr.it/EAGLES96/segsasg1/html.
Li, A. Y.-h., 2001, "The *ba* construction". University of Southern California, ms. Available
    on-line at http://www.usc.edu/dept/LAS/ealc/chinling/audreyli.htm.
Li, Ch. and S. Thompson, 1989, *Mandarin Chinese: A Functional Reference Grammar.*
    Berkeley and Los Angeles: University of California Press.
Li, J., 1997, "The *ba* construction in Mandarin Chinese: a serial verb analysis". *Clitics,
    Pronouns and Movement,* ed. J. Black and V. Motapanyane, 175-216. Amsterdam and
    Philadelphia: John Benjamins Publishing Company.
Lin, K. H. C. and V.-w., Soo, 1994, "Hypothesis scoring over theta grids information
    in parsing Chinese sentences with serial verb constructions". *Proceedings of the 15th
    International Conference on Computational Linguistics (COLING '94),* 942-948. Kyoto,
    Japan.
Liu, Ch.-sh., 1999, *Anaphora in Mandarin Chinese and binding at the interface.* Unpublished
    Ph.D. thesis, University of California, Irvine.
Liu, G., 1996, «On serial verbs in Chinese and their representation in HPSG". *Proceedings
    of International Chinese Computing Conference (ICCC-96),* 129-35. Singapore.
Liu, Y.-h., 1996, *Modern Chinese Grammar for Teachers of Chinese as a Second Language and
    Advanced Learners of Modern Chinese.* Taiwan: Shi Da Shu Yuan.
Norman, J., 1988, *Chinese.* Cambridge and New York: Cambridge U.P.
Quirk, R. and S. Greenbaum, 1973, *A University Grammar of English.* London: Long-
    man.
Sampson, G., 1995, *English for the Computer: The SUSANNE Corpus and Analytic Scheme.*
    Oxford: Clarendon Press.
Wu, Z. X.-z., 2000, *Grammaticalisation and the development of functional categories in
    Chinese.* Unpublished Ph.D. thesis, University of Southern California.
Xia, X.-y. and De-kai Wu, 1996, "Parsing Chinese with an almost-context-free grammar".
    Paper presented at Conference on Empirical Methods in Natural Language Processing
    (EMNLP-96), May 17-18. University of Pennsylvania, Philadelphia.
Xiao, Z.-h., 2002, *A corpus-based study of aspect in Mandarin Chinese.* Unpublished Ph.D.
    thesis, Lancaster University.
Xue, N. Fei Xia, Shi-zhe Huang and A. Kroch, 2000, *The Bracketing Guidelines for
    the Penn Chinese Treebank (3.0).* Available on-line at http://www.cis.upenn.edu/
    ~chinese/.
Yu, S.-w., 1999, *Xiandai Hanyu Yuliaoku Jiagong: Ciyu Qiefen Yu Cixing Biaozhu Guifen Yu
    Shouce* [Corpus annotation of Modern Chinese Corpora: Guidelines and Handbook on
    Word Segmentation and Part-of-Speech Tagging]. Beijing: Institute of Computational
    Linguistics, Peking University. Available on-line at http://www.icl.pku.edu.cn/research/
    corpus/coprus-annotation.htm.

Zhao, Er-pin, 1989, *The categorical status of THAN and BI in English and Chinese comparative constructions.* Unpublished Ph.D. thesis. University of Texas, Austin.

Zhu, De-xi, 1982, *Yufa Xiangyi* [Lecture Notes on Chinese Grammar]. Beijing: Commercial Press.

—, 2000, *Xiandai Hanyu Yufa Yanjiu* [A Grammatical Study of Modern Chinese]. Beijing: Commercial Press.

Zou, Ke, 1993, "The syntax of the Chinese *ba* construction", *Linguistics* 31: 715-736.