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

* I am indebted to the audience in the Bilbao-Deusto Student Conference 2004, held in Bilbao, Spain, from July 8 to 10 for their insightful comments. All the errors are, of course, mine.

[ASJU 39-2, 2005, 271-287]
http://www.ehu.es/ojs/index.php/asju
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>
<thead>
<tr>
<th>UCREL Skeleton Parsing Annotation Scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fa</td>
</tr>
<tr>
<td>Fc</td>
</tr>
<tr>
<td>Fn</td>
</tr>
<tr>
<td>Fr</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>J</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>Nr</td>
</tr>
<tr>
<td>Nv</td>
</tr>
<tr>
<td>P</td>
</tr>
<tr>
<td>S</td>
</tr>
<tr>
<td>Tg</td>
</tr>
<tr>
<td>Ti</td>
</tr>
<tr>
<td>Tn</td>
</tr>
<tr>
<td>V</td>
</tr>
<tr>
<td>(null)</td>
</tr>
</tbody>
</table>

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) and relative clause (Fr) 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. Lastly, I did not take the genitive (G) as one of the constituent labels in my parsing scheme so as to avoid terminological controversy.

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 “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 of de in Chinese. However, this apparent correspondence is complicated by two issues. Firstly, possessive constructions in Chinese do not necessarily take the particle of de (Li and Thompson 1989: 115) as in 他的爸爸 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 of 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>
<thead>
<tr>
<th>Nonterminal Category</th>
<th>Symbol</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adverbial Clause</td>
<td>Fa</td>
<td>&lt;Fa&gt;&lt;N&gt;我们_人&lt;/N&gt;&lt;R&gt;前进_的&lt;/R&gt;“解放思想_i’，_w 实事求是_i’，_w &lt;V&gt;抓住_y &lt;N&gt;机遇_n&lt;/N&gt;’，_w &lt;V&gt;开拓进取_b&lt;/V&gt;’，_w &lt;V&gt;zhiyao_c &lt;N&gt;妇女&lt;/N&gt; &lt;R&gt;jinyibu_d&lt;/R&gt; &lt;i&gt;shishiqiushi_i&lt;/i&gt;’，_w &lt;V&gt;zhuazhu_v &lt;N&gt;jiu_n&lt;/N&gt;’，_w &lt;V&gt;kaituojinqu_b&lt;/V&gt;’，_w “If we become more open-minded and down-to-earth and make every effort to explore new possibilities, …”</td>
</tr>
<tr>
<td>Correlative Clause</td>
<td>Fc</td>
<td>&lt;Fc&gt;&lt;V&gt;学_的&lt;/V&gt; &lt;V&gt;刚_n&lt;/V&gt;“the more we walk, the broader (the road) will be”</td>
</tr>
<tr>
<td>Main Clause (to which the adverbial clause is subordinated)</td>
<td>Fm</td>
<td>&lt;Fm&gt;&lt;Fa&gt;&lt;N&gt;我们_人&lt;/N&gt;&lt;R&gt;前进_的&lt;/R&gt;“解放思想_i’，_w 实事求是_i’，_w &lt;V&gt;抓住_y &lt;N&gt;机遇_n&lt;/N&gt;’，_w &lt;V&gt;开拓进取_b&lt;/V&gt;’，_w &lt;V&gt;jiefangsixiang_i&lt;/V&gt;’，_w &lt;V&gt;shishiqiushi_i&lt;/V&gt;’，_w &lt;V&gt;zhuazhu_v &lt;N&gt;jiu_n&lt;/N&gt;’，_w &lt;V&gt;kaituojinqu_b&lt;/V&gt;’，_w “If we become more open-minded and down-to-earth and make every effort to explore new possibilities, …”</td>
</tr>
</tbody>
</table>

6 All of the examples in this paper are given in the Chinese characters, followed by pinyin romanisations and English translations.
<table>
<thead>
<tr>
<th>Nonterminal Category</th>
<th>Symbol</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adverbial Idiom/Set Phrase</td>
<td>Ia</td>
<td>“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.”</td>
</tr>
<tr>
<td>Adjective Phrase</td>
<td>J</td>
<td>“very important”</td>
</tr>
<tr>
<td>Adverbial Adjective Phrase</td>
<td>Ja</td>
<td>“successfully”</td>
</tr>
<tr>
<td>Noun Phrase</td>
<td>N</td>
<td>“a hundred years’ history”</td>
</tr>
<tr>
<td>Adverbial Noun Phrase</td>
<td>Na</td>
<td>“at present”</td>
</tr>
<tr>
<td>Prepositional Phrase</td>
<td>P</td>
<td>“to Hong Kong”</td>
</tr>
<tr>
<td>Adverbial Prepositional Phrase</td>
<td>Pa</td>
<td>“in this year”</td>
</tr>
<tr>
<td>Adverb Phrase</td>
<td>R</td>
<td>“not…though”</td>
</tr>
<tr>
<td>Sentence (including direct speech quotation, also with &amp; and + as co-ordinates)</td>
<td>S</td>
<td>“In this year, the Chinese government has gained great success in its diplomacy.”</td>
</tr>
</tbody>
</table>
### Nonterminal Category | Symbol | Example
--- | --- | ---
Verb Phrase | V | “have hope in the new era”
Adverbial Verb Phrase | Va | “at the moment we are looking forward to the prosperous year of 1998”
Verbal Object | Vo | “wish to rely on you”
Initial Conjunct | &c | “China’s revolution”
Non-initial Conjunct | + | “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><联合国_nt 和_c 其他_r 国际_n 组织_n 的_u> 协调_v</N> 金<br>融_n 金融_v “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_v 金融_v “the recent financial crises happened in some countries and districts”;<br>

(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> <zhibei_v yanpou_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. <大气磅礴_i ,_w 波澜壮阔_i> <daiqibangbo_i ,_w bolanzhuangkuo_i> “powerful wind, fierce waves”;
- e.g. <流光溢彩_i ,_w 火树银花_i> <liuangangyicai_i ,_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”;<br>
- 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 guili_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>_往_ đ_v_ ᶵ_d</R> <R>_hai_ đ_bu_ đ_d/R> “not…though”; \(<R>_永远_ đ_不再_ đ_d</R> <R>_yongyuang_ đ_buzai_ đ_d/R> “never forever”; \(<R>_yizhi_ đ_dou_ đ_d</R> “constantly”), and multi-word attributive adjectival phrases containing an adjective premodified by at least one adverb (e.g. \(<J>_朆ⷠ_慵_天_ ᶱ_u</J> <J>_feichang_ đ_zhongyao_ a_ de_u</J> “very important”; \(<J>_很_ đ_不_ đ_平凡_ a_的_u</J> <J>_hen_ đ_bu_ đ_pingfan_ a_ de_u</J> “very extraordinary; \(<J>_十分_ a_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>_人民_n 生活_vn</N> <R>_进一步_d</R> <V>_改善_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&>_全党_n</N&> _和_c <N+>_全国_n 各族_r 人民_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>\(<P>_为_p <N>_“_w_ 两手抓_ l_ 、_w_ 两手_m 都_d 要_v 硬_a ”_w</N></P>
   \(<P>_提供_v_了_u <N>_新_a 的_u 理论_n 根据_n</N></P>\(<P>_提供_v_了_u <N>_新_a 的_u 理论_n 根据_n</N></P>
   “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_n</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_1 ”_w , _w 高度_d 自治_v 的_u> 方针_n</N></P> <V>保持_v <N>香港_n 的_u 繁荣_an 稳定_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 themselves’ 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.
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, the 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 the ba noun phrase (i.e. the NP directly following ba) followed by a verb.

\[(5) \text{subject ba NP verb}\]

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.

\[(6) \text{subject ba NP verb}\]

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.

\[(6) \text{subject ba NP verb}\]
follows the *ba* construction lacks a direct object, which is equivalent to the *ba* noun phrase, i.e. *ba* NP, verb Ø, where Ø denotes the empty position of the preposed verb object that shares the same reference as the prepositional complement of *ba*. 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>㈲_p<N>_ナ_大_听众_耳所能详的_歌曲</N></P> <V>改编_Øi</V> <V>为_wei</V> <N>器乐曲</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_人_N_人们</N></P> <V>带入_v</V> <N>迷人_的艺术_境地</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_的_长安街</N> <W_火树银花_i>今_流光溢采</W>_<I>Changanjie</I> <N>Changanjie</N></N> <W_火树银花_i>“Tonight the Changan Street was filled with colourful lights and really looked magnificent.”

建_好_a, 建-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.
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 de at the end, as in example 11. The particle de is traditionally treated as a marker of modification (Chao 1968: 285).

(11) <N><中国_n 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 bezuo_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</V> <N>朋友家</N> <V>吃</V> <N>晚饭</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> “led and performed a variety of Chinese and western popular songs”

(15) <V><坚持_v 奉行_v> <N>独立自主_l 的_u 和平_n 外交_n 政策_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 \(-le\), 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

---

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


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.


