# Skeleton Parsing in Chinese: Annotation Scheme and Guidelines

**May Lai-Yin Wong**

Department of Chinese, Translation and Linguistics, City University of Hong Kong
83 Tat Chee Avenue, Kowloon Tong, Hong Kong
maywong@cityu.edu.hk

**Abstract**

This paper presents my manual skeleton parsing on a sample text of approximately 100,000 word tokens (or about 2,500 sentences) taken from the PFR Chinese Corpus with a clearly defined parsing scheme of 17 constituent labels. The manually-parsed sample skeleton treebank is one of the very few extant Chinese treebanks. While Chinese part-of-speech tagging and word segmentation have been the subject of concerted research for many years, the syntactic annotation of Chinese corpora is a comparatively new field. The difficulties that I encountered in the production of this treebank demonstrate some of the peculiarities of Chinese syntax. A noteworthy syntactic property is that some serial verb constructions tend to be used as if they were compound verbs. The two transitive verbs in series, unlike common transitive verbs, do not take an object separately within the construction; rather, the serial construction as a whole is able to take the same direct object and the perfective aspect marker *le*. The skeleton-parsed sample treebank is evaluated against Eyes & Leech (1993)'s criteria and proves to be accurate, uniform and linguistically valid.

## 1. Introduction

While Chinese part-of-speech tagging (Zhang & Sheng, 1997) and word segmentation (Sun et al., 2000) have been the subject of concerted research for many years, the syntactic annotation of Chinese corpora is a comparatively new field. Although much treebanking of English has occurred, relatively little of such work has been done on Asian languages, Chinese included (cf. Han et al., 2002).

Treebanks are simply corpora in which syntactic constituent structure is made explicit by a process of corpus annotation (Leech & Garside, 1991, p. 15; Abeillé, 2003, p. 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 & Leech, 1993; Garside, 1993; Black et al., 1996; Leech & Eyes, 1997). Skeleton parsing seeks to produce simplified constituent-structure annotations. I do not intend to go into a deep or logical annotation which would apply functional labels for constituents (e.g. subject, object, etc.) as traditional reference grammar books do (Quirk et al., 1985). Rather, I will focus on categorial labels such as noun phrase, prepositional phrase, adverb phrase, etc. The reason why I use categorial labels rather than functional labels is that the complexity involved in tagging syntactic functions would inevitably require automation or partial automation followed by manual post-editing. However, before the process of automation can be carried out, a clearly-defined parsing scheme should be made available and be applied manually to real-life language data to assess the viability of the annotation scheme.

In this paper, I will describe how to produce a skeleton treebank by using a sample text of approximately 100,000 word tokens (amounting to about 2,500 Chinese sentences). A clearly-defined parsing scheme will be given, which comprises 17 constituent labels and 11 textual markers, followed by a set of parsing guidelines. Some syntactic properties of the Chinese language found in the sample skeleton treebank will also be discussed. Finally, the sample skeleton treebank will be evaluated against some quality-control criteria.

## 2. PFR Sample Skeleton Treebank: Text Selection

The Chinese language data used in this paper is taken from the PFR (*Peita-Fujitsu-Renmin Ribao*) People's Daily POS Tagged Chinese Corpus (abbreviated to PFR Chinese Corpus hereafter) Release 1.0 (http://www.ling.lancs.ac.uk/corplang/pdcorpus/pdcorpus.htm). From the PFR Chinese Corpus, a sample text of some 100,000 word tokens, yielding approximately 2,500 sentences, was chosen for the production of my treebank.

## 3. Parsing Schemes

As Sampson (1995, p. 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 & Garside, 1991, p. 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]

[1] The table was adapted from UCREL (University Centre for Computer Corpus Research on Language)'s website http://www.comp.lancs.ac.uk/computing/research/ucrel/skeletontags.html.

| 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 |

Table 1: The UCREL annotation scheme

### 3.2. PFR Skeleton Parsing Annotation Scheme

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. The non-terminal labels and the symbols to represent them in text are given for the PFR treebank in Table 2.

| Non-terminal Category | Symbol |
|---|---|
| Adverbial Clause | Fa |
| Correlative Clause | Fc |
| Main Clause (to which the adverbial clause is subordinated) | Fm |
| Adverbial Idiom/Set Phrase | Ia |
| Adjective Phrase | J |
| Adverbial Adjective Phrase | Ja |
| Noun Phrase | N |
| Adverbial Noun Phrase | Na |
| Prepositional Phrase | P |
| Adverbial Prepositional Phrase | Pa |
| Adverb Phrase | R |
| Sentence (including direct speech quotation, also with & and + as co-ordinates) | S |
| Verb Phrase | V |
| Adverbial Verb Phrase | Va |
| Verbal Object | Vo |
| Initial Conjunct | & |
| Non-initial Conjunct | + |

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

## 4. Annotation Guidelines

It is advisable, as Kahrel et al. (1997, p. 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. My documented parsing guidelines include practical issues related to map any parses on to sentences in the application of the parsing scheme. The following issues will be discussed:

a) Underspecification – Use of unlabelled bracketings;
b) Bracketing of multi-word constituents;
c) Bracketing of single-word constituents;
d) Punctuation;
e) Ambiguity.

### 4.1. Underspecification – Use of Unlabelled Bracketings

Brackets may be left unlabelled in cases where a particular grouped sequence of words cannot fit in to 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*; 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'.

Those grammatical constructions marked by the particle *de* are highly controversial: some scholars refer to them as relative clauses (e.g. Li & Thompson, 1989, p. 579ff; Aoun & Li, 1993; Chiu, 1993; Wu, 2000; Xue et al., 2000) or appositive clauses (Chu & Chi, 1999, p. 26), while others do not agree with either approach (e.g. Chao, 1968; Zhu, 1982, 2000; Liu, 2003). Moreover, using some catch-all label, such as '*de* constructions', does not help either because the particle *de* in itself is vague in its functions: it can be a genitive marker, a marker of nominalisation and an adjectival marker. It is therefore not easy to agree upon how the *de* constructions are defined and instantiated in texts. Having considered that there is to date no consensus on this issue of how *de* constructions should be analysed, and any invented label that attempts to refer to them will act as a locus of controversy and disagreement, I decided not to set up a new label for them in my parsing scheme and thus these constructions were enclosed by unlabelled brackets in the treebank.

(b) Serial verb constructions which are used as if they were compound verbs (see also section 5.1.1):

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 5.1.2):

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'.

Idioms or set phrases can function as predicate in a sentence. However, even if context is taken into consideration, it is not obvious whether these idiomatic expressions function as a nominal predicate or a verbal predicate, both of which are allowed in the Chinese syntax (Chao, 1968, p. 90). They were thus not enclosed in the <V> element nor <N> element in the treebank and left unlabelled instead.

(e) Coordinated verbs with shared direct object:

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'.

Two or more transitive verbs in coordination share the same direct object. The coordinated verbs (except the last one) are not constituent-like in the sense that they do not constitute a complete verb phrase structure because the following shared object does not come immediately after them. I did not therefore apply the usual practice of marking conjuncts by enclosing them in the <V&> and <V+> elements, which are only used for coordinated verbs or verb phrases with complete verb phrase structure. Since it is also not worthwhile to set up a new parsing label to mark such a non-frequent phenomenon, I decided to put these verbal segments into unlabelled brackets.

### 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 & Leech (1993, p. 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, or where they are in coordination with other multi-word constituents.

### 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 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.

### 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. In fact, similar problems were encountered in the production of the Penn Chinese Treebank and the annotators of the treebank did not annotate ambiguities either (Xue et al., 2000:, p. 73-178). They believed that in each case one of these ambiguous readings was unlikely and thus they annotated assuming the more plausible reading. In this regard, my treebank may appear unsatisfactory in connection with research on different kinds of ambiguity.

## 5. The Process of Skeleton Parsing

The basic idea of skeleton parsing, as Garside & McEnery (1993, p. 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-parsed treebank is given in Figure 1.

<S N= '28' ><V>环顾_v <N>全球_n</N></V> ，_w <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>

Figure 1: An excerpt of the PFR skeleton-parsed sample treebank

### 5.1. 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.

#### 5.1.1. Serial Verb Constructions

Serial verb constructions in Chinese increase the complexity of parsing. There is an immense literature on Chinese serial verb constructions (see, for instance, Li & Thompson, 1989, p. 594ff; Lin & 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.

(1) <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>
I go.to friend home
<V>*chi* <N>*wanfan*</N></V>
eat dinner
'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 (2) and (3), 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 (2), and坚持_v 奉行_v *jianchi fengxing* 'insist and follow' as in (3)), functioning as if they were a single unit, take the same object, i.e. the following noun phrase.

(2) <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
lead perform PERF one CL
*zhongwai*_j *mingqu*_n</N></V>
Chinese.and.Western popular.songs
'led and performed a variety of Chinese and western popular songs'

(3) <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
insist.on follow independent DE
*heping*_n *waijiao*_n *zhengce*_n</N>
peace diplomatic policy
'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 (2). The verbal -*le* has generally been taken as an aspect marker, indicating completion (Norman, 1988, p. 163; Xiao, 2002), and it is attached to verbs and not to the objects of verbs (Chao, 1968, p. 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.

#### 5.1.2. Idioms and 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 (4) and verb-like set phrases in example (5). To my knowledge, the grammatical categories of this kind of idiomatic expressions have not been documented so far.

(4) <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>
tonight DE Changan.Street
<*liuguangyicai*_l ，_w
filled.with.colourful.lights
*huoshuyinhua*_i>
bright.red.trees.with.silver.flowers
'Tonight the Changan Street was filled with colourful lights and really looked magnificent.'

(5) <N>国民经济_n</N> <稳中求进_l>
<N>*guominjingji*_n</N> <*wenzhongqiujin*_l>
national.economy steadily.progress
'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, p. 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.

## 6. Quality Control

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

(a) Consensual categories: The linguistic categories that were employed in my parsing scheme represent grammatical features largely agreed upon by linguists, rather than features which are theory-specific or deeply controversial.

(b) 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.

(c) Productivity: Productivity (i.e. the number of word tokens parsed within a reasonable length of time) was satisfactory with the simplified syntactic analysis provided by skeleton parsing.

(d) 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.

(e) 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. 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.

<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>

Figure 2: A concordance of the verb *yao*

(f) Linguistic validity: One of the possible uses of the skeleton parsing on a sample text of the PFR Chinese Corpus is that it could be used to gain a better understanding of how to precisely locate adverbial clauses in a piece of POS tagged text; in written Chinese, adverbial clauses are typically overtly marked by subordinating conjunctions (or subordinators) of various sorts.

## 7. Concluding Remarks

In my skeleton treebank, functional labels were put aside in order to give a consistent and accurate manual parsing. However, the annotation of syntactic functions may throw up interesting results regarding the range of functions that a phrasal category can take within a sentence; a phrasal category may assume a syntactic function that is not conventionally associated with it.

It is expected that more large-scale treebanks with expanded size and coverage will be built in the near future (cf. Han et al., 2002). It is also hoped that further research could look into the possibility of automating the parsing so that more data could be treebanked and used in various sorts of linguistic analyses and beyond.

## 8. References


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

Aoun, J. & Li, A. (1993). *Syntax of Scope*. Cambridge: MIT Press.

Black, E., Eubank, S., Kashioka, H., Magerman, D., Garside, R. & Leech, G. (1996). Beyond skeleton parsing: producing a comprehensive large-scale general-English treebank with full grammatical analysis. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*. Copenhagen, pp. 107-112.

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

Chiu, B. (1993). The inflectional structure of Mandarin Chinese. Ph.D. thesis. University of Los Angeles.

Chu, C. & Chi, T.J. (1999). *A Cognitive-Functional Grammar of Mandarin Chinese*. Taiwan: Crane Publishing.

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

Garside, R. & McEnery, A. (1993). Treebanking: the compilation of a corpus of skeleton-parsed sentences. In E. Black, R. Garside, & G. Leech (Eds.), *Statistically-driven Computer Grammars of English: The IBM/Lancaster Approach*. Amsterdam: Rodopi, pp.17-35.

Garside, R. (1993). The large-scale production of syntactically analysed corpora. *Literary and Linguistic Computing*, 8(1), pp. 39-46.

Han, C.H., Han, N.R., Ko, E.S. & Palmer, M. (2002). Development and evaluation of a Korean treebank and its application to NLP. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*. Las Palmas, Canary Islands, Spain.

Kahrel, P., Barnett, R. & Leech, G. (1997). Towards cross-linguistic standards or guidelines for the annotation of corpora. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus Annotation: Linguistics Information from Computer Text Corpora*. London and New York: Longman, pp. 231-242.

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

Leech, G. & Garside, R. (1991). Running a grammar factory: the production of syntactically analysed corpora or 'treebanks'. In S. Johansson, & A.-B. Stenström (Eds.), *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton de Gruyter, pp.15-32.

Leech, G., Barnett, R. & Kahrel, P. (1996). *Recommendations for the Syntactic Annotation of Corpora*. EAGLES Document EAG-TCWG-SASG/1.8 Version of 11th March 1996. URL:

http://www.ilc.cnr.it/EAGLES96/segsasg1/segsasg1.html.

Li, C. & Thompson, S. (1989). *Mandarin Chinese: A Functional Reference Grammar*. Berkeley and Los Angeles: University of California Press.

Lin, H.C. & Soo, V.W. (1994). Hypothesis scoring over theta grids information in parsing Chinese sentences with serial verb constructions. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*. Kyoto, Japan, pp. 942-948.

Liu, G. (1996). On serial verbs in Chinese and their representation in HPSG. In *Proceedings of International Chinese Computing Conference (ICCC-96)*. Singapore, pp. 129-135.

Liu, H.Y. (2003). *A Profile of the Mandarin NP: Possessive Phrases and Classifier Phrases in Spoken Discourse*. München, Germany: Lincom.

Norman, J. (1988). *Chinese*. Cambridge and New York: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Sampson, G. (1995). *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Oxford: Clarendon Press.

Sun, M.S., Sun, H.L., Huang, C.N., Zhang, P., Xing, H.B. & Zhou, Q. (2000). Hua Yu: A word segmented and part-of-speech tagged Chinese corpus. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*. Athens, Greece, pp.1081-1085.

Wu, Z. (2000). Grammaticalisation and the development of functional categories in Chinese. Ph.D. thesis. University of Southern California.

Xiao, Z.H. (2002). A corpus-based study of aspect in Mandarin Chinese. Ph.D. thesis. Lancaster University.

Xue, N.W., Xia, F., Huang, S.Z. & Kroch, A. (2000). *The Bracketing Guidelines for the Penn Chinese Treebank (3.0)*. URL: http://www.cis.upenn.edu/~chinese/.

Zhang, M. & Sheng, L. (1997). Tagging Chinese corpus using statistics techniques and rule techniques. In *Proceedings of 1997 International Conference on Computer Processing of Oriental Languages (ICCPOL'97)*. Hong Kong, China, pp.503-506.

Zhu, D.X. (1982). *Yufa Xiangyi* [Lecture Notes on Chinese Grammar]. Beijing: Commercial Press.

Zhu, D.X. (2000). *Xiandai Hanyu Yufa Yanjiu* [A Grammatical Study of Modern Chinese]. Beijing: Commercial Press.