Dependency-based Syntactic Annotation of a Chinese Corpus

Tom B.Y. LAI
City University of Hong Kong
Tsinghua University, Beijing
cttomlai@cityu.edu.hk

HUANG Changning
Microsoft Research, China
cnhuang@microsoft.com

Abstract

We discuss a syntactic annotation scheme for Chinese text corpora following a dependency-based framework that admits no intermediate phrasal nodes and allows no crossing of syntactic dependency links. While one particular approach to syntactic analysis is being followed, dependency annotation facilitates the use of annotated corpora by followers of other approaches.

1 Introduction

Major linguistic theories like GB/MP (Chomsky, 1986; Chomsky, 1995), HPSG (Pollard and Sag, 1994) and LFG (Bresnan, 1982) agree to represent syntactic structures in terms of phrase structures, but disagree about what kinds of phrase structures should be assigned to the same linguistic expressions. In a project on syntactic annotation of Chinese corpora, we represent syntactic structures in terms of dependency. [1]

We follow an approach (Lai and Huang, 1998a; Lai and Huang, 1999a) to Dependency Grammar (Tesnière, 1959; Gaifman, 1965; Hays, 1964; Robinson, 1970), that requires syntactic dependency to be single-headed and projective. Unlike Dependency Grammar schools that allow multiple-headed and non-projective dependency structures (Hudson, 1984; Mel'čuk, 1988; Starosta, 1988; Hajičova, 1991), single-headedness and projectivity are maintained in a syntactic skeleton, with reference to which constraints to capture non-projective phenomena in language are anchored. In experimental implementations (Lai and Huang, 1998b; Lai and Huang, 1999b) of this approach, projective syntactic dependency structures are generated subject to the constraints of subcategorization properties of the words concerned as well as other grammatical considerations. This approach is different from many works in dependency-based parsing (Hellwig, 1986; Covington, 1990; Courtin and Genthial, 1998; Bourdon et al., 1998) in that the relationship between the governor and all its dependents are immediate and no intermediate phrasal nodes are necessary. Similar "flat" syntactic structures have recently been suggested in phrase-structure grammars (Bouma et al., 1998; Przepiórkowski, 1999).

In preparation for large-scale annotation, we are carrying out manual syntactic annotation of a small Chinese legal text corpus. The text is first processed using a "segmentation" and "tagging" tool (Lai et al., 1992; Lai et al., 1998). The tokens are then subjected to morphological analysis to confirm and adjust word boundaries. Words, as the units that are operated on in syntactic analysis, form the basis of an SGML-based annotation scheme. The annotation scheme also recognizes larger parsing units like phrases and sentences and smaller units like *characters* and dictionary entries, which may or may not coincide with the words.

Following accepted practices of text corpora annotation, the original character sequences of the raw corpus are preserved as the terminally tagged elements. This enables recovery from possible errors in morphological and syntactic analysis. The representation of syntactic relationships in terms of dependency also facilitates the use of the annotated corpus by followers of other approaches.

2 Projective dependency syntax without intermediate phrasal nodes

2.1 Projective syntactic dependency skeleton

In Dependency Grammar (Tesnière, 1959), words are linked to one another by asymmetrical governor-dependent relationships. Syntactic dependency structures are constrained by Robinson (1970) as follows:

(1) a. One and only one element is independent.

b. All others depend directly on some element.

c. No element depends directly on more than one other.

[1] The Academica Sinica (Taipei) Chinese treebank and the LDC Chinese Treebank Project should be noted.

d. If A depends directly on B and some element C intervenes between them (in linear order of string), then C depends directly on A or on B or some other intervening element.

Robinson requires that a word should not depend on more than one word. She also requires that syntactic dependency structures be *projective* in the sense that dependency links should not cross one another.

For example, the projectivity criterion will be violated if the word *ta* ('he') in the Chinese sentence (2) is considered to depend on the matrix verb *xiang* ('wanted') and the embedded verb *xiao* ('laugh') at the same time.

(2) Ta xiang xiao.

he want laugh

'He wants/wanted to laugh.'

In Figure 1, the dependency link between *ta* and *xiao* crosses the branch linking the matrix verb *xiang* to the root node. This situation is dealt with by suggesting a projective skeleton structure as in Figure 2. The link between *ta* and *xiao* is severed. The fact that *ta* is the subject of *xiao* is accounted for by a specification in the lexical entry of the word *xiang*: the subject of *xiang* is the subject of its predicate complement *xiao*. Other non-projective linguistic phenomena are dealt with similarly by grammatical constraints defined in terms of the nodes and arcs of the projective syntactic dependency skeleton.

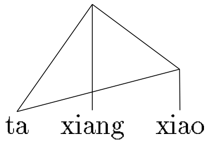

Figure 1: Non-projective syntactic structure

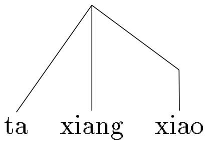

Figure 2: Projective syntactic structure

2.2 Dependency rules

Projective dependency structures can be generated using Hays' (1964) dependency rules.

(3) a. X(A, B, C, . . . , *, Y, . . . , Z)

b. X(*)

c. *(X)

In (3), dependents of the governor X are listed between a pair of brackets, with the asterisk * indicating the position of the governor itself.

Hays' rules have the disadvantage of having word order (of dependents with the same governor) built into the rule mechanism. This disadvantage is removed by making dependency rules binary-branching.

Repeated application of binary-branching dependency rules will over-generate, but subcategorization properties of the governing word and global grammatical constraints of the language will co-operate to function as a filter and account for the correct ensemble of dependent elements in the "domain" of the governor.

In dependency rules, the governor and its placeholder * are not only of the same *type*, as in phrase-structure rewrite rules, but also *token*-identical. The result is that a "phrase" is indistinguishable from its head word, and the ensemble of a head word and its dependent is a "flat" structure without intermediate phrasal nodes.

3 Dependency-based annotation

3.1 A small text corpus

We begin with manually annotating a small corpus, with a plan to scale up with the help of the experience gained. We use a small corpus of Chinese text segmented and tagged using a bigram-based segmentation-tagging tool (Lai et al., 1998). The corpus is two "chapters" of a statute in an East Asian Chinese community (Hong Kong). It contains 4797 tokens produced by the segmentation-tagging tool. Because of its small size, this corpus is stylistically not balanced, which is to be borne in mind.

3.2 SGML-style annotation scheme

The annotation scheme is based on SGML (SGML, 1986). Its design is explained with the help of the following example:

```
<pu pi=1>
<mu mi=i wu=1><du tg=hm><cu>"di4"
<wu wi=1 gv=2 fn=nm ct=mx sm="seventh">
  <du tg=mx ><cu>"qi1"
<wu wi=2 gv=0 ct=ncl sm="chapter">
  <du tg=cnb><cu>"zhang1"
</pu>
<pu pi=2>
<wu wi=1 gv=0 fn=sub ct=nc mh="de"
    sm="contract">
  <du tg=ncd><cu>"he2"<cu>"yue4"
<mu mi=1 wu=1><du tg=ed><cu>"de"
<wu wi=2 gv=0 fn=sub ct=na sm="form">
  <du tg=nad><cu>"xing2"<cu>"shi4"
```

```
</pu>
<pu pi=3>
<wu wi=1 gv=0 ct=mx sm="7.1">
  <du tg=mx><cu>7<du tg="."><cu>.
  <du tg=mx><cu>1
</pu>
<pu pi=4>
<wu wi=1 gv=0 ct=na sm="form">
  <du tg=nad><cu>"xing2"<cu>"shi4"
</pu>
<pu pi=5>
<wu wi=1 gv=0 ct=mx sm="7.1.1">
 <du tg=mx><cu>7<du tg="."><cu>.
 <du tg=mx><cu>1<du tg="."><cu>.
 <du tg=mx><cu>1
</pu>
<pu pi=6>
<wu wi=1 gv=7 fn=sub ct=nc sm="contract">
  <du tg=ncd><cu>"he2"<cu>"yue4"
<wu wi=2 gv=5 fn=mks ct=cnj sm="because">
  <du tg=jom><cu>"yin1"
<wu wi=3 gv=5 fn=sub ct=na sm="form">
  <du tg=nad><cu>"xing2"<cu>"shi4"
<wu wi=4 gv=5 fn=neg ct=adv sm="neg">
  <du tg=bu><cu>"bu4"
<wu wi=5 gv=7 fn=ajt ct=vt sm="conform">
  <du tg=vnm><cu>"he2"
<wu wi=6 gv=5 fn=obj ct=na
    sm="stipulation">
  <du tg=nad><cu>"gui1"<cu>"ge2"
<wu wi=7 gv=0 ct=aux sm="can">
  <du tg=ud><cu>"ke3"<cu>"neng2"
<wu wi=8 gv=7 fn=axo ct=aj sm="invalid">
  <du tg=aod><cu>"wu2"<cu>"xiao4"
<wu wi=9 gv=11 fn=mkc ct=cnj sm="or">
  <du tg=jom><cu>"huo4"
<wu wi=10 gv=11 fn=neg ct=av sm="neg">
  <du tg=bu><cu>"bu4"
<wu wi=11 gv=8 fn=cjt ct=aux sm="can">
  <du tg=um><cu>"neng2"
<wu wi=12 gv=11 fn=axo ct=vt
    sm="carry out">
  <du tg=vnd><cu>"li3"<cu>"xing2"
<wu wi=13 gv=7 fn=pt ct="..">
  <du tg=".."><cu>..
</pu>
```

The terminal text elements, marked by the $< cu >$ tags, are Chinese (*Han*) characters in a two-byte encoding scheme. They are written in the phonetic *pingyin* script in this paper for the benefit of the reader.

The largest text unit for syntactic analysis shown here is not the *sentence*, but the *parsing unit* $< pu >$. There are six such units in the example: an ordinal numerical phrase, a chapter title, two numeral construction in Arabic numberals, a section heading, and a one-sentence subsection text.

The words, as the basic units of subsequent syntactic analysis, play a key role in the annotation scheme. The semantic glosses of the words are given in the *sm* attribute. In general, sub-word morphological units are contained within the scope of a $< wu >$ tag. The $< du >$ tag marks dictionary entries as a sub-word units, which, especially in Chinese, often do not coincide with the words they constitute.

The tags $< wu >$ and $< cu >$ correspond to the $< w >$ and $< c >$ tags for *linguistic segmentation elements* in CES (Ide et al., 1996). The usage of $< mu >$, however, is different from that of $< m >$ in CES. We do not have tags corresponding to $< cl >$ and $< phr >$ in CES. As noted earlier, $< pu >$ can be a word, a phrase or a sentence.

The $< wu >$ elements have an index attribute *wi* to mark their positions within the $< pu >$ unit. They also have a *gv* attribute recording the *wi* indices of their governors. A value of 0 shows that the word is the head element of the $< pu >$. The syntactic category of the word is given by the *ct* attribute, and its relation to its immediate governor is the fn attribute.

When values are assigned to the *gv* attribute of $< wu >$, care is taken to have Robinson's "axioms" of well-formedness (1) observed. Projectivity is ensured by checking that the *gv* value of a word is neither smaller than that of any words preceding it in the same $< pu >$ nor greater than that of any other words following it in the $< pu >$.

In our annotation scheme, morphemes (*mu*) are marked only when they are not adequately covered by the words and the dictionary entries. When they morphemes are marked, as in $< pu\ pi = 1 >$ and $< pu\ pi = 2 >$, they are not marked as constituents of a *wu*. This will be explained later in this paper.

4 Basic features of the annotation scheme

4.1 Preservation of raw text elements

Chinese texts are stored as sequences of "characters" without explicit word boundary marks. With very few exceptions, Chinese characters are meaning-bearing syllables. They may function either as one-morpheme words or as morphemes that combine to form words. Unfortunately, Chinese linguists do not always agree about how a given sequence of characters should be "segmented" in words. It is thus important that the raw character sequence of the original text should be preserved for the benefit of people who do not agree with us.

The basic encoding units of European texts are the letters of the alphabet. Letters combine to form words, which are marked off from one another by white spaces (though sometimes "words" will have to be combined to form compound words). It is thus not uncommon for the terminally tagged units of annotated European texts to be (lemmatized) words. This will be fine if the morphological analysis is always correct, but will make recovery from errors like mistaking "bake[PAST]" for "bake[PART]" difficult.

In our annotation scheme, punctuation marks are also preserved and marked as such in our annotation scheme. Besides providing hints for syntactic and pragmatic analysis, they also mark off small chunks of character sequences for the segmentation-tagging program to operate on.

4.2 Flexibility of parsing units

In Chinese test, as well as in texts in other languages, the chunk of text that one has to feed into a "sentence" analyzer are often not a sentence. In the example in the previous section, $< pu\ pi = 1 >$, $< pu\ pi = 2 >$, $< pu\ pi = 3 >$, $< pu\ pi = 4 >$ and $< pu\ pi = 5 >$ are all not sentences. In an English translation of the text, they may be rendered as *Chapter Seven*, *The Form of a Contract*, *7.1*, *Form*, and *7.1.1* respectively, which should also be treated as non-sentence parsing units.

Thus, we allow parse units to be anything suggested by the text itself. shown in the example, they may be words, phrases and, of course, sentences. Parse units form separate domains for the position indices of their constituent words. Head words of parse units are not assigned governors of of their own. Relationships between parse units are considered to belong to the realm of pragmatics.

4.3 Dependency marking

We do not mark phrase structures. There no bracketing as in the PENN Treebank. There are no intermediate phrase nodes as in GB/MP, HPSG and LFG, and the relationship between governor and dependent is always direct. This does not not solve the problem of different approaches producing different syntactic structures for the same linguistic expression, but rather accentuates it in a somewhat positive sense. This will be discussed in greater detail in a later section.

5 Morphological complications

5.1 Deriving words from dictionary units

In Chinese, as in all other languages, words may combine to form larger compounds words. It is often justified to have compound words listed in our dictionary. Sometimes, however, listing a word formed from simpler words in a dictionary can be unrealistic or unreasonable.

The third $< pu >$ in Section 3.2, repeated below, serves to illustrate this.

```
<pu pi=3>
<wu wi=1 gv=0 ct=mx sm="7.1">
  <du tg=mx><cu>7<du tg="."><cu>.
  <du tg=mx><cu>1
</pu>
```

The segmentation-tagging program outputs three "tokens" as candidates for separate words. We adjust the word boundaries and group the three characters together to form only one word ($< wu >$), which is a kind or numerical label.

We do not consider this an error of the segmentation-tagging program. Numeral characters like *7*, "." and *1* are dictionary entries that are, with the helps of marks like the point ".", capable of combining transparently to form an infinite number of words like *7.1*, which is not, and cannot all be, listed in a dictionary. In the example, we mark the three characters as a word, but also retain the information that this word is composed from the three one-character dictionary entries. The grammatical information originally attached to the three constituent "tokens" are retained. They are tagged as dictionary entries ($< du >$).

Obviously, word units like *7.1* are also possible in other languages. Besides, in Chinese and in other languages, it is sometimes difficult to decide whether a number of space-separated "words" should combine to form a compound word. In view of this, dictionary unit tags are a good way to ensure the usefulness of the annotated corpus.

5.2 Derivational and inflectional affixes

Derivational morphology is encountered in first $< pu >$ in the example in Section 3.2:

```
<pu pi=1>
<mu mi=i wu=1><du tg=hm><cu>"di4"
<wu wi=1 gv=2 fn=nm ct=mx sm="seventh">
  <du tg=mx><cu>"qi1"
<wu wi=2 gv=0 ct=ncl sm="chapter">
  <du tg=cnb><cu>"zhang1"
</pu>
```

In *di4qi1*, the prefix *di4* is attached to the cardinal number *qi1* ('seven') to turn it into an ordinal number. Like *-th* and *-ieme* in English and French, *di4* is a bound morpheme in modern Chinese. However, this prefix is listed as a separate entry in all Chinese dictionaries, and the ordinary native speaker has difficulty in seeing it as different from "real" words in the language. Anyway, no graphical hints are available to distinguish between free and bound morphemes in Chinese text.

In respect of the rather general practice of the Chinese computational linguistics community to mark off bound grammatical morphemes like *di*4 as separate "words". affixes are marked as $< mu >$ and placed immediately under $< pu >$ like $< wu >$. However, as we choose to consider affixes as part of the words to which they are attached, they are not given an independent position index, and their *mi* indices are meaningful only within the scope of the words to which they are attached. The syntactic categories and the meanings of the word units are those of the derived words.

The second < *pu* > is an example of inflectional morphology.

```
<pu pi=2>
<wu wi=1 gv=0 fn=sub ct=nc mh="de"
    m="contract">
  <du tg=ncd><cu>"he2"<cu>"yue4"
<mu mi=1 wu=1><du tg=ed><cu>"de"
<wu wi=2 gv=0 fn=sub ct=na sm="form">
  <du tg=nad><cu>"xing2"<cu>"shi4"
</pu>
```

Inflectional affixes are dealt with like derivational affixes. The suffix *de* is a genitive marker in Chinese. It is marked as an < *mu* > and is assigned an *mi* and assigned an *mi* index that is only meaningful to the word *he2yue4* ('contract'). One significant difference from the treatment of derivational affixes should be noted. Inflectional affixes do not change the meanings and syntactic categories of the words to which they are attached to. To show their effects as grammatical morphemes, an attribute *mh* is added to the stem word units.

5.3 Discontinuous morphological phenomena

It should be noted that as far as affixes discussed above are concerned, we could also have them included under the < *wu* > tags of the stems they attach to. The more complicated treatment described above is in fact motivated by the discontinuous morphological phenomena as shown in the following examples, which are not attested in our corpus (not so much because of its small size, but because of its stylistic bias)

```
<pu>
<wu wi=1 gv=0 ct=vn mh="perf" sm="have meal">
  <du mu="2" tg=vnm sm="eat"><cu>"chi1"
<mu mi=1 wu=1><du tg=el><cu>"le"
<mu mi=2 wu=1><du tg=ncm sm="meal"><cu>"fan4"
</pu>
<pu>
<wu wi=1 gv=0 ct=adj mh="dup, de" sm="happy">
  <du mu="2" tg=ad><cu>"gao1"
<mu mi=1 wu=1><cu>"gao1"
<mu mi=2 wu=1><cu>"xing4"
<mu mi=3 wu=1><cu>"xing4"
<mu mi=4 wu=1><du tg=ed><cu>"de"
</pu>
```

In the first example, an infix *le* is inserted between the two characters of the word *chi1fan4*. As is common in computational linguistics research on Chinese, the segmentation program "segments" the text rather "lemmatize". It outputs three one-charter tokens, which, in Chinese, are all valid dictionary entries. The treatment of *le* ('PERF') is like *de*, which is to be expected. The constituent "word" *fan4* is separated from its "major" partner *chi1* in the compound word *chi1fan4* ('have meal'). It has to be marked as an < *mu* > attached to its major partner (as representative of the whole compound word) in order not to get into the way of subsequent syntactic analysis. The meaning of the < *wu* > is that of the compound word.

We "lemmatize", but we take care to make sure that the original output of the segmentation-tagging program is recoverable, just in case users of our annotated corpus are not happy with our lemmatization results.

The second example is even more interesting. The word is "inflected" form of the two-character dictionary entry *gao1xing4* ('happy'). The morphological process of reduplication has been applied, and each of the two characters is repeated to give a four-character surface form. As the segmentation-tagging program does not lemmatize and meddle with the order in which characters occur, its output is (somewhat erroneously) the four-token sequence of *gao1 gao1 xing4 xing4*. These characters are all valid dictionary entries in Chinese themselves, but they do not "combine" to form the word *gao1gao1xing4xing4*, which is obtained from *gao1xing4* by a kind of reduplication. In our annotation scheme, the four characters are one < *wu* > (with a number of < *mu* >'s attached to it. The original output of the segmentation program is preserved.

6 Problems with sharing

6.1 Incompatible syntactic structures

Syntactic structures produced according to different theoretical approaches are incompatible. Efforts like the PENN Treebank has tried to minimize the differences by adopting a basic bracketing scheme. But consider the following example from our corpus:

```
<pu>
<wu wi=1 gv=4 fn=mks ct=cnj sm="because">
  <du tg=jom><cu>"yin1"
<wu wi=2 gv=4 fn=sub ct=na sm="form">
  <du tg=nad><cu>"xing2"<cu>"shi4"
<wu wi=3 gv=4 fn=neg ct=adv sm="neg">
  <du tg=bu><cu>"bu4"
<wu wi-4 gv=8 fn=ajt ct=vt sm="conform">
  <du tg=vnm><cu>"he2"
<wu wi=5 gv=5 fn=obj ct=na
    sm="stipulation">
  <du tg=nad><cu>"gui1"<cu>"ge2"
<wu wi=6 gv=8 fn=mkm ct=cnj sm="then">
  <du tg=jom><cu>"er2"
```

```
<wu wi=7 gv=8 fn=neg cat=adv sm="neg">
  <du tg=bu><cu>"bu4"
<wu wi=8 gv=11 fn=ajt ct=aux sm="can">
  <du tg=um><cu>"neng2"
<wu wi=9 gv=8 fn=axo ct=vt sm="carry out">
  <du tg=vnd><cu>"li3"<cu>"xing2"
<wu wi=10 gv=8 fn=mka ct=de sm="rel">
  <du tg=ez><cu>"zhi1"
<wu wi=11 gv=0 fn=sub ct=nc sm="contract">
  <du tg=ncd><cu>"he2"<cu>"yue4"
</pu>
```

Linguists generally agree that the chunk from $< wu\ wi = 1 >$ to $< wu\ wi = 5 >$ is a "subordinate clause". but they may disagree about the internal structure of the clause. In GB/MP, the subordinating conjunction *yin1* ('because') will be the head of the tree hierarchy as shown below (unnecessary details skipped):

```
(yin1 (xing2shi4 bu4 he2 gui1ge2))
```

In HPSG, *yin1* may be analyzed as a "marker" or a "preposition" with a sentential complement. When analyzed as a marker, it will be a dependent of the verb *he2* ('conform')

```
((yin1) (xing2shi4) (bu4) he2 (gui1ge2))
```

When *yin1* is analyzed as a preposition, it may be considered the head of the phrase structure. However, it has been argued within HPSG that the preposition is a dependent of the verb *he2*.

The position of the subordinating conjunction in the syntactic structure of a subordinate clause is thus different depending on the syntactic theory followed. In our annotation, *yin1* is marked as a subordinating conjunction ($ct = cnj$). It is marked as being governed by the head word *he2*, for which it functions somewhat like a *marker* in HPSG (Pollard and Sag, 1994).

As we can obviously not claim impartiality for our analysis, our use of dependency annotation is of course not a solution to the problem. However, clearly indicating the dependency relationships in the parse structure will *accentuate* the disagreement. If the other researcher who wants to use our annotated corpus happens to favour an analysis that gives the same dependency relationships as we have marked in the annotation, then it will be up to him to flesh up the dependency structure with intermediate phrasal nodes according to his grammatical formalism. More often, the other researcher will find that the annotated syntactic structure does not agree completely with his own opinions, it should then be *easier* for him to make the necessary transformations if annotation consists only of skeletal dependencies without the complications arising intermediate nodes.

6.2 Shareable annotated corpora

It will be in vain for one to attempt to find parse structures that are universally accepted. What we can do, and have done, is to label arcs of our parse structures with the dependency relation names, thus leaving the hope alive that our parse structures may be convertible for use by researchers following other approaches.

While it will be out of our control whether other researchers will find our annotated corpora useful, we will be eager to be able to convert linguistic corpora annotated by other researchers for our own use.

Conclusion

We are giving SGML-based syntactic annotation to a small corpus of Chinese text as a piloting effort leading to large-scale syntactic annotation. We are also investigating the potential of importing and adapting annotated corpora prepared by researchers following other approaches. With the experience gained in this pilot effort, we will try to scale up to annotate a stylistically more balanced corpus. We also explore the possibility of making use of resources prepared by other researchers.

Acknowledgement

This research is supported by Grant No. 7000888, City University of Hong Kong.

References

Gosse Bouma, Rob Malouf, and Ivan A. Sag. 1998. A unified theory of complement, adjunct, and subject extraction. In Gosse Bouma, Geert-Jan M. Kruijff, and Richard T. Oehrle, editors, *Proceedings of the FHCG98. Symposium on Unbounded Dependencies*, pages 83–97, Saarbrücken. Universität des Saarlandes und DFKI.

Marie Bourdon, Lyne Da Sylva, Michel Gagnon, Alma Kharrat, Sonja Knoll, and Anna Maclachlan. 1998. A Case Study in Implementing Dependency-Based Grammars. In Alan Polguère and Sylvain Kahane, editors, *Proceedings of COLING-ACL'98 Workshop on Processing of Dependency-Based Grammars*, pages 88–94, Université de Montréal, August.

Joan W. Bresnan, editor. 1982. *The Mental Representation of Grammatical Relations*. The MIT Press, Cambridge, Massachusetts.

Noam Chomsky. 1986. *Barriers*. The MIT Press, Cambridge, Massachusetts.

Noam Chomsky. 1995. *The Minimalist Program*. The MIT Press, Cambridge, Massachusetts.

Jacques Courtin and Damien Genthial. 1998. Parsing with Dependency Relations and Robust Parsing. In Alan Polguère and Sylvain Kahane, editors, *Proceedings of COLING-ACL'98 Workshop* *on Processing of Dependency-Based Grammars*, pages 95–101, Université de Montréal, August.

Michael A. Covington. 1990. Parsing Discontinuous Constituents in Dependency Grammar. *Computational Linguistics*, 16(4):234–236.

Haim Gaifman. 1965. Dependency Systems and Phrase-Structure Systems. *Information and Control*, 8:304–337.

Eva Hajičova. 1991. Free Word Order Described Without Unnecessary Complexity. *Theoretical Linguistics*, 17:99–106.

David G. Hays. 1964. Dependency Theory: A Formalism and Some Observations. *Language*, 40:511–525.

Peter Hellwig. 1986. Dependency Unification Grammar. In *Proceedings of 11th International Conference on Computational Linguistics (COLING'86)*, pages 195–199.

Richard Hudson. 1984. *Word Grammar.* Blackwell, Oxford.

Nancy Ide, Greg Priest-Dorman, and Jean Véronis. 1996. Corpus Encoding Standard. Expert Advisory Group on Language Engineering Standards. Last modified 20 March 2000.

Tom Bong-Yeung Lai and Changning Huang. 1998a. An Approach to Dependency Grammar for Chinese. In Yang Gu, editor, *Studies in Chinese Linguistics*, pages 143–163. Linguistic Society of Hong Kong, Hong Kong.

Tom Bong Yeung Lai and Changning Huang. 1998b. Complements and Adjuncts in Dependency Grammar Parsing Emulated by a Constrained Context-Free Grammar. In Alan Polguère and Sylvain Kahane, editors, *Proceedings of COLING-ACL'98 Workshop on Processing of Dependency-Based Grammars*, pages 102–108, Université de Montréal, August.

Tom Bong Yeung Lai and Changning Huang. 1999a. Unification-based Parsing Using Annotated Dependency Rules. In Jost Gippert and Peter Olivier, editors, *Multilinguale Corpora: Codierung, Strukturierung, Analyse - 11. Jahrestagung der Gesellschaft für Linguistische Daten Verarbeitung (GLDV'99)*, pages 235–244. Enigma Corporation, Praha, December.

Tom Bong Yeung Lai and Changning Huang. 1999b. Unification-based Parsing Using Annotated Dependency Rules. In *Proceedings of 5th Natural Language Processing Pacific Rim Symposium 1999 NLPRS'99*, pages 102–107, Beijing, November.

T.B.Y. Lai, S.C. Lun, C.F. Sun, and M.S. Sun. 1992. A Tagging-Based First-Order Markov Model Approach to Automatic Word Identification for Chinese Sentences. In Julius T. Tou and Joseph J. Liang, editors, *Intelligent Systems for Processing Oriental Languages (Proceedings of the 1992 International Conference on Computer Processing of Chinese and Oriental Languages)*, pages 17–23, Tampa, Florida, December. Chinese Language Computer Society.

Tom B.Y. Lai, M.S. Sun, S.C. Lun, and B.K. Tsou. 1998. Using Syntactically Motivated Tags in Markov Model Word Segmentation. In *Proceedings of 1998 International Conference on Chinese Information Processing (ICCIP'98)*, pages 215–222, Beijing.

Igor A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice.* State University of New York Press, New York.

Carl Pollard and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar.* University of Chicago Press, Chicago.

Adam Przepiórkowski. 1999. On complements and adjuncts in Polish. In Robert D. Borsley and Adam Przepiórkowsk, editors, *Slavic in HPSG*, pages 183–210. CSLI Publications, Stanford.

Jane J. Robinson. 1970. Dependency Structures and Transformation Rules. *Language*, 46:259–285.

SGML. 1986. Information Processing - Text and Office Systems - Standard Generalized Markup Language (SGML). Genf: International Organization for Standardization. ISO8897.

Stanley Starosta. 1988. *The Case for Lexicase.* Pinter Publishers, London.

Lucien Tesnière. 1959. *Elements de syntaxe structurale.* Klincksieck, Paris.