Annotating the Propositions in the Penn Chinese Treebank

Nianwen Xue
Dept. of Computer and Info. Science
University of Pennsylvania
Philadelphia, PA 19104, USA
xueniwen@linc.cis.upenn.edu

Martha Palmer
Dept. of Computer and Info. Science
University of Pennsylvania
Philadelphia, PA 19104, USA
mpalmer@linc.cis.upenn.edu

Abstract

In this paper, we describe an approach to annotate the propositions in the Penn Chinese Treebank. We describe how diathesis alternation patterns can be used to make coarse sense distinctions for Chinese verbs as a necessary step in annotating the predicate-structure of Chinese verbs. We then discuss the representation scheme we use to label the semantic arguments and adjuncts of the predicates. We discuss several complications for this type of annotation and describe our solutions. We then discuss how a lexical database with predicate-argument structure information can be used to ensure consistent annotation. Finally, we discuss possible applications for this resource.

1 Introduction

Linguistically interpreted corpora are instrumental in supervised machine learning paradigms of natural language processing. The information encoded in the corpora to a large extent determines what can be learned by supervised machine learning systems. Therefore, it is crucial to encode the desired level of information for its automatic acquisition. The creation of the Penn English Treebank (Marcus et al., 1993), a syntactically interpreted corpus, played a crucial role in the advances in natural language parsing technology (Collins, 1997; Collins, 2000; Charniak, 2000) for English. The creation of the Penn Chinese Treebank (Xia et al., 2000) is also beginning to help advance technologies in Chinese syntactic analysis (Chiang, 2000; Bikel and Chiang, 2000). Since the treebanks are generally syntactically oriented (cf. Sinica Treebank (Chen et al., to appear)), the information encoded there is ”shallow”. Important information useful for natural language applications is missing. Most notably, significant regularities in the predicate-argument structure of lexical items are not captured. Recent effort in semantic annotation, the creation of the Penn Proposition Bank (Kingsbury and Palmer, 2002) on top of the Penn English Treebank is beginning to address this issue for English. In this new layer of annotation, the regularities of the predicates, mostly verbs, are captured in the predicate-argument structure. For example, in the sentences “The Congress passed the bill” and “The bill passed”, it is intuitively clear that “the bill” plays the same role in the two occurrences of the verb “pass”. Similar regularities also exist in Chinese. For example, in “这/this 条/CL 法案/bill 通过/pass 了/AS” and “议会/Congress 通过/pass 了 /AS 这/this 条/CL 法案/bill”, “法案/bill” also plays the same role for the verb “通过/pass” even though it occurs in different syntactic positions (subject and object respectively).

Capturing such lexical regularities requires a ”deeper” level of annotation than generally provided in a typical syntactically oriented treebank. It also requires making sense distinctions at the appropriate granularity. For example, the regularities demonstrated for “pass” does not exist in other senses of this verb. For example, in “He passed the exam” and “He passed”, the object “the exam” of the transitive use of "pass" does not play the same role as the subject "he" of the intransitive use. In fact, the subject plays the same role in both sentences.

However, how deep the annotation can go is constrained by two important factors: how consistently human annotators can implement this type of annotation (the **consistency** issue) and whether the annotated information is learnable by machine (the **learnability** issue). Making fine-grained sense distinctions, in particular, has been known to be difficult for human annotators as well as machine-learning systems (Palmer et al., submitted). It seems generally true that structural information is more learnable than non-structural information, as evidenced by the higher parsing accuracy and relatively poor fine-grained WSD accuracy. With this in mind, we will propose a level of semantic annotation that still can be captured in structural terms and add this level of annotation to the Penn Chinese Treebank. The rest of the paper is organized as follows. In Section 2, we will discuss the annotation model in detail and describe our representation scheme. We will discuss some complications in Section 3 and some implementation issues in Section 4. Possible applications of this resource are discussed in Section 5. We will conclude in Section 6.

2 Annotation Model

In this section we describe a model that annotates the predicate-argument structure of Chinese predicates. This model captures the lexical regularities by assuming that different instances of a predicate, usually a verb, have the same predicate argument structure if they have the same sense. Defining sense has been one of the most thorny issues in natural language research (Ide and Vronis, 1998), and the term "sense" has been used to mean different things, ranging from part-of-speech and homophones, which are easier to define, to slippery fine-grained semantic distinctions that are hard to make consistently. Determining the "right" level of sense distinction for natural language applications is ultimately an empirical issue, with the best level of sense distinction being the level with the least granularity and yet sufficient for a natural language application in question. Without gearing towards one particular application, our strategy is to use the structural regularities demonstrated in Section 1 to define sense. Finer sense distinctions without clear structural indications are avoided. All instances of a predicate that realize the same set of semantic roles are assumed to have one sense, with the understanding that not all of the semantic roles for this verb sense have to be realized in a given verb instance, and that the same semantic role may be realized in different syntactic positions. All the possible syntactic realizations of the same set of semantic roles for a verb sense are then **alternations** of one another. This state of affairs has been characterized as **diathesis alternation** and used to establish cross-predicate generalizations and classifications (Levin, 1993). It has been hypothesized and demonstrated that verbs sharing the same disthesis alternation patterns also have similar meaning postulates. It is equally plausible to assume then that verb instances having different diathesis alternation patterns also have different semantic properties and thus different senses.

Using diathesis alternation patterns as a diagnostic test, we can identify the different senses for a verb. Alternating **syntactic frames** for a particular verb sense realizing the same set of semantic roles (we call this **roleset**) form a **frameset** and share similar semantic properties. It is easy to see that each frameset, a set of syntactic frames for a verb, corresponds with one roleset and vice versa. From now on, we use the term **frameset** instead of **sense** for clarity. Each frameset consists of one or more syntactic frames and each syntactic frame realizes one or more semantic roles. One frame differs from another in the number and type of arguments its predicate actually takes, and one frameset differs from another in the total number and type of arguments its predicate CAN take. This is illustrated graphically in Figure 1.

Annotating the predicate-argument structure involves mapping the frameset identification information for a predicate to an actual predicate instance in the corpus and assign the semantic roles to its arguments based on the syntactic frame of that predicate instance. It is hoped that since framesets are defined through diathesis alternation of syntactic frames, the distinctions made are still structural in nature and thus are machine-learnable and can be consistently annotated by human annotators.

So far our discussion has focused on semantic ar-

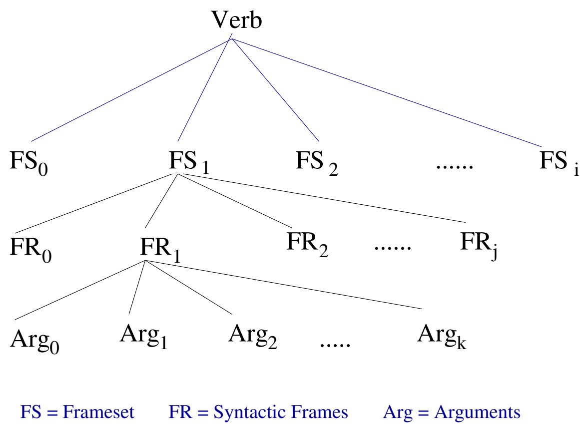

Figure 1: Annotation model

guments, which play a central role in determining the syntactic frames and framesets. There are other elements in a proposition: semantic adjuncts. Compared with semantic arguments, semantic adjuncts do not play a role in defining the syntactic frames or framesets because they occur in a wide variety of predicates and as a result are not as discriminative as semantic arguments. On the other hand, since they can co-occur with a wide variety of predicates, they are more generalizable and classifiable than semantic arguments. In the next section, we will describe a representation scheme that captures this dichotomy.

2.1 Representing arguments and adjuncts

Since the number and type of semantic arguments for a predicate are unique and thus define the semantic roles for a predicate, we label the arguments for a predicate with a contiguous sequence of integers, in the form of *argN*, where N is the integer between 0 and 5. Generally, a predicate has fewer than 6 arguments. Since semantic adjuncts are not subcategorized for by the predicate, we use one label *argM* for all semantic adjuncts. *ArgN* identifies the arguments while *argM* identifies all adjuncts. An *argN* uniquely identifies an argument of a predicate even if it occupies different syntactic positions in different predicate instances. Missing arguments of a predicate instance can be inferred by noting the missing argument labels.

Additionally, we also use secondary tags to generalize and classify the semantic arguments and adjuncts when possible. For example, an adjunct receiving a TMP tag if it is a temporal adjunct. The secondary tags are reserved for semantic adjuncts, predicates that serve as arguments, as well as certain arguments for phrasal verbs. The 18 secondary tags and their descriptions are presented in Table 1.

11 functional tags for semantic adjuncts	
ADV	adverbial, default tag
BNF	beneficiary
CND	condition
DIR	direction
DGR	degree
FRQ	frequency
LOC	locative
MNR	manner
PRP	purpose or reason
TMP	temporal
TPC	topic
1 functional tag for predicate as argument	
PRD	predicate
6 functional tags for arguments to phrasal verbs	
AS	为, 是, 作, 做
AT	在, 于
INTO	成, 入, 进
ONTO	上
TO	到, 至
TOWARDS	向, 往

Table 1: List of functional tags

3 Complications

In this section we discuss several complications in annotating the predicate-argument structure as described in Section 2. Specifically, we discuss the phenomenon of "split arguments" and the annotation of nominalized verbs (or deverbal nouns).

3.1 Split Arguments

What can be characterized as "split arguments" are cases where a constituent that occurs as one argument in one sentence can also be realized as multiple arguments (generally two) for the same predicate in another sentence, without causing changes in the meaning of the sentences. This phenomenon surfaces in several different constructions. One such construction involves "possessor raising", where the possessor (in a broad sense) raises to a higher position. Examples 1a and 1b illustrate this. In 1a, the possessor originates from the subject position and raises to the topic[1] position, while in 1b, the possessor originates from the object position and raises

[1] In Chinese, it is possible to have a topic in addition to the subject. The topic is higher than the subject and plays an important role in the sentence (Li and Thompson, 1976).

to the subject position. The exact syntactic analysis is not important here, and what is important is that one argument in one sentence becomes two in another. The challenge is then to capture this regularity when annotating the predicate-argument structure of the verb.

1. Possessor Raising
a. Subject to Topic
(IP (NP-PN-TPC 中国/China)
(NP-TMP 去年/last year)
(NP-SBJ 进出口/import-export
总值/total volume)
(VP 逾/exceed
(QP-OBJ 三千二百五十亿/325 Billion
(CLP 美元/US. Dollar))))
逾/exceed
arg0-**psr**: 中国/China
arg0-**pse**: 进出口/import-export 总值/total volume
arg1: 三千二百五十亿/325 Billion 美元/US. Dollar

(IP (NP-TMP 去年/last year)
(NP-SBJ (DNP (NP-PN 中国/China)
的/DE)
(NP 进出口/import-export
总值/volume))
(VP 逾/exceed
(QP-OBJ 三千二百五十亿/325 Billion
(CLP 美元/US. Dollar))))
逾/exceed
arg0: 中国/China 的/DE 进出口/import-export
总值/volume
arg1: 三千二百五十亿/325 Billion 美元/US. Dollar

b. Object to Subject
(IP (NP-SBJ (NP-PN 中国/China)
(NP 经济/economy
增长/expansion))
(VP (ADVP 也/also)
(ADVP 将/will)
(VP 放慢/slow down
(NP-OBJ 速度/speed)))
放慢/slow down
arg1-**psr**: 中国/China 经济/economy 增长/expansion
arg1-**pse**: 速度/speed

(IP (NP-SBJ (DNP (NP (NP-PN 中国/China)
(NP 经济/economy
增长/expansion))
的)
(NP 速度/speed))
(VP (ADVP 也/also)
(ADVP 将/will)
(VP 放慢/slow down))
放慢/slow down
arg1: 中国/China 经济/economy 增长/expansion
的/DE 速度/speed

Another case of "split arguments" involves the coordinated noun phrases. In 2a, for example, the coordinated structure as a whole is an argument to the verb "签订/sign". In contrast, in 2b, one piece of the argument, "中国 /China" is realized as a noun phrase introduced by a preposition. There is no apparent difference in meaning for the two sentences.

2. Coordination vs. Prepositional phrase
a. (IP (NP-PN-SBJ 缅甸/Burma
和/and
中国/China)
(VP (ADVP 已/already)
(VP 签订/sign
了/ASP
(NP-OBJ 边境/border
贸易/trade
协定/agreement))))
签订/sign
arg0: 缅甸/Burma 和/and 中国/China
arg1: 边境/border 贸易/trade 协定/agreement

b. (IP (NP-PN-SBJ 缅甸/Burma)
(VP (ADVP 已/already)
(PP 同/with
(NP-PN 中国/China))
(VP 签订/sign
了/ASP
(NP-OBJ 边境/border
贸易/trade
协定/agreement))))
签订/sign
arg0-**crd**: 缅甸/Burma
arg0-**crd**: 中国/China
arg1: 边境/border 贸易/trade 协定/agreement

There are two ways to capture this type of regularity. One way is to treat each piece as a separate argument. The problem is that for coordinated noun phrases, there can be arbitrarily many coordinated constituents. So we adopt the alternative approach of representing the entire constituent as one argument. When the pieces are separate constituents, they will receive the same argument label, with different secondary tags indicating they are parts of a larger constituent. For example, in 1, when possessor raising occurs, the possessor and possessee receive the same argument label with different secondary tags *psr* and *pse*. In 2b, both "中国/China" and "缅甸/Burma" receive the label *arg0*, and the secondary label *crd* indicates each one is a part of the coordinated constituent.

3.2 Nominalizations

Another complication involves nominalizations (or deverbal nouns) and their co-occurrence with light and not-so-light verbs. A nominalized verb, while serving as an argument to another predicate (generally a verb), also has its own predicate-argument structure. For example, in 3, the predicate-argument structure for "怀疑/doubt" should be "怀疑 (读者, 这条 新闻)", where all the arguments of "怀疑/doubt" are embedded in the NP headed by "怀疑/doubt". The complication arises when the nominalized noun is a complement to another verb, as in 4, where the subject "读者/reader" is an argument to both the verb "产生/produce" and the nominalized verb "怀疑/doubt". More interestingly, the other argument "这/this 条/CL 新闻/news" is realized as an adjunct to the verb (introduced by a preposition) even though it bears no apparent thematic relationship to it.

It might be tempting to treat the verb "产生/develop" as a "light verb" that does not have its own predicate-argument structure, but this is questionable because "产生/doubt" can also take a noun that is not a nominalized verb: "我/I 对/towards 她/she 产生/develop 了/LE 感情/feeling". In addition, there is no apparent difference in meaning for "产生/develop" between this sentence and 4, so there is little basis to say these are two different senses of this verb. So we annotate the predicate-argument structure of both the verb "产生 (读者, 怀疑)" and the nominalized verb "怀疑 (读者, 这 条 新闻)".

```
3. (IP (NP-SBJ (NP 读者/reader)
               (DNP (PP 对/towards
                        (NP (DP 这/this
                                (CLP 条/CL))
                            (NP 新闻/news)))
                    的)
               (NP 怀疑/doubt))
       (VP 加深/deepen
           了/LE))
加深/deepen
arg1:读者/reader 对/towards 这/this 条/CL
    新闻/news
```

```
4. (IP (NP-SBJ 读者/reader)
       (VP (PP-DIR 对/towards
                   (NP (DP 这/this
                           (CLP 条/CL))
                       (NP 新闻/news)))
           (ADVP 也/too)
           (VP 会/will
               (VP 产生/develop
                   (NP-OBJ 怀疑/doubt)))))
产生/develop
arg0: 读者/reader
arg1: 怀疑/doubt
```

4 Implementation

To implement the annotation model presented in Section 2, we create a lexical database. Each entry is a predicate listed with its framesets. The set of possible semantic roles for each frameset are also listed with a mnemonic explanation. This explanation is not part of the formal annotation. It is there to help human annotators understand the different semantic roles of this frameset. An annotated example is also provided to help the human annotator.

As illustrated in Example 5, the verb "通过/pass" has three framesets, and each frameset corresponds with a different meaning. The different meanings can be diagnosed with diathesis alternations. For example, when "通过/pass" means "pass through", it allows dropped object. That is, the object does not have to be syntactically realized. When it means "pass by vote", it also has an intransitive use. However, in this case, the verb demonstrates "subject of the intransitive / object of the transitive" alternation. That is, the subject in the intransitive use refers to the same entity as the object in the transitive use. When the verb means "pass an exam, test, inspection", there is also the transitive/intransitive alternation. Only in this case, the object of the transitive counterpart is now part of the subject in the intransitive use. This is the argument-split problem discussed in the last section. The three framesets, representing three senses, are illustrated in 5.

5. Verb: 通过/pass
Frameset.01: 穿越,穿过/pass through
Roles: arg0("passer"), arg1("place")
Example:
(IP (NP-SBJ 火车/train)
    (VP (ADVP 正在/now)
        (VP 通过/pass
            (NP-OBJ 遂道/tunnel))))
通过.01/pass
arg0: 火车/train
arg1: 遂道/tunnel
argM-ADV: 正在/now

(IP (NP-SBJ 火车/train)
    (VP (ADVP 正在/now)
        (VP 通过/pass)))
通过.01/pass
arg0: 火车/train
argM-ADV: 正在/now

Frameset.02: 及格,合格(考试,比赛等)/pass (an exam, etc.)
(IP (NP-SBJ (DNP (NP 他/he)
        的/DE)
    (NP 药检/drug inspection))
    (VP (ADVP 没/not)
        (VP 通过/pass)))
通过.02/pass
arg1: 他/he 的/DE 药检/drug inspection

(IP (NP-SBJ (NP 他/he)
    (VP (ADVP 没/not)
        (VP 通过/pass)))
            (NP-OBJ 药检/drug inspection))
通过.02/pass
arg1-psr: 他/he
arg1-pse: 药检/drug inspection

Frameset.03: 表决通过/pass (a bill, a law, etc.)
(IP (NP-PN-SBJ 美国/the U.S.
        国会/Congress)
    (VP (NP-TMP 最近/recently)
        (VP 通过/pass
            了/ASP
            (NP-OBJ 州际/interstate
                银行法/banking law))))
通过.03/pass
arg0: 美国/the U.S.
arg1: 州际/interstate 银行法/banking law

(IP (NP-SBJ (ADJP 州际/interstate)
        (NP 银行法/banking law))
    (VP (NP-TMP 最近/recently)
        (VP 通过/pass
            了/ASP)))
通过.03/pass
arg1: 州际/interstate 银行法/banking law

The human annotator can use the information specified in this entry to annotate all instances of "通过/pass" in a corpus. When annotating a predicate instance, the annotator first determines the syntactic frame of the predicate instance, and then determine which frameset this frame instantiates. The frameset identification is then attached to this predicate instance. This can be broadly construed as "sense-tagging", except that this type of sense tagging is coarser, and the "senses" are based on structural distinctions rather than just semantic nuances. A distinction is made only when the semantic distinctions also coincide with some structural distinctions. The expectation is that this type of sense tagging is much amenable to automatic machine-learning approaches. The annotation does not stop here. The annotator will go on identifying the arguments and adjuncts for this predicate instance. For the arguments, the annotator will determine which semantic role each argument realizes, based on the set of possible roles for this frameset, and attach the appropriate semantic role label (*argN*) to it. For adjuncts, the annotator will determine the type of adjunct this is and attach a secondary tag to *argM*.

5 Applications

A resource annotated with predicate-argument structure can be used for a variety of natural language applications. For example, this level of abstraction is useful for Information Extraction. The argument role labels can be easily mapped to an Information Extraction template, where each role is mapped to a piece of information that an IE system is interested in. Such mapping will not be as straightforward if it is between surface syntactic entities such as the subject and IE templates.

This level of abstraction can also provide a platform where lexical transfer can take place. It opens up the possibility of linking a frameset of a predicate in one language with that of another, rather than using bilingual (or multilingual) dictionaries where one word is translated into one or more words in a different language. This type of lexical transfer has several advantages. One is that the transfer is made more precise, in the sense that there will be more cases where one-to-one mapping is possible. Even in cases where one-to-one mapping is still not possible, the identification of the framesets of a predicate will narrow down the possible lexical choices. For example, *sign.02* in the English Proposition Bank (Kingsbury and Palmer, 2002) will be linked to "签订.01/enter into an agreement". This type of linking rules out "签署" as a possible translation for *sign.02*, even though it is a translation for other framesets of the word *sign*.

The transfer will also be more precise in another sense, that is, the predicate-argument structure of a word instance will be preserved during the transfer process. Knowing the arguments of a predicate instance can further constrain the lexical choices and rule out translation candidates whose predicate-argument structures are incompatible. For example, if the realized arguments of "sign.01" of the English Proposition Bank in a given sentence are the signer, the document, and the signature, among the translation candidates "签署, 签" ("签订.01/enter into an agreement" is ruled out as a possibility for this frameset), only "签" is possible, because "签署" can only take two arguments, namely, the signer and the document.

6. 他/he 在/at 这/this 个/CL 文件/document 上/LC 签/sign 了/LE 自己/self 的/DE 名字/name
   "He signed his name on this document."

One might argue that the syntactic subcategorization frame obtained from the syntactic parse tree can also constrain the lexical choices. For example, knowing that "sign" has a subject, an object and a prepositional phrase should be enough to rule out "签署" as a possible translation. This argument breaks down when there are lexical divergences. The "document" argument of "签字" can only be realized as a prepositional phrase in Chinese while in English it can only be realized the direct object of "sign". If the syntactic subcategorization frame is used to constrain the lexical choices for "sign", "签字" will be incorrectly ruled out as a possible translation. There will be no such problem if the more abstract predicate-argument structure is used for this purpose. Even when the document is realized as a prepositional phrase, it is still the same argument. Of course, "签署/sign" is also a possible translation. So compared with the surface syntactic frames, the predicate-argument structure constrains the lexical choices without incorrectly ruling out legitimate translation candidates. This is understandable because the predicate-structure abstracts away from the syntactic idiosyncracies of the different languages and thus are more transferable across languages.

7. 他/he 在/at 这/this 个/CL 文件/document 上/LC 签字/sign
   他/he 签署/sign 这/this 个/CL 文件/document
   "He signed this document."

Annotating the predicate-argument structure as described in previous sections will not reduce the lexical choices to one-to-one mappings in call cases. For example, "统一" can be translated into "standardize" or "unite", even though there is only one frameset for both finer senses of this verb. It is conceivable that one might want to posit two framesets, each for one finer sense of this verb. This is essentially a trade-off: either one can conduct deep analysis of the source language, resolve all sense ambiguities on the source side and have a more straightforward mapping, or one takes the one-to-many mappings and select the correct translation on the target language side. Hopefully, the annotation of the predicate-argument provides just the right level of abstraction and the resource described here, with each predicate annotated with its arguments and adjuncts in context, enables the automatic acquisition of the predicate-argument structure.

6 Summary

In this paper, we described an approach to annotate the propositions in the Penn Chinese Treebank. We described how diathesis alternation patterns can be used to make coarse sense distinctions for Chinese verbs as a necessary step in annotating the predicate-structure of predicates. We also described the representation scheme we use to label the semantic arguments and adjuncts of the predicates. We discussed several complications for this type of annotation and described our solutions. We then discussed how a lexical database with predicate-argument structure information can be used to ensure consistent annotation. Finally, we discussed possible applications for this resource.

7 Acknowledgement

This work is supported by MDA904-02-C-0412.

References

Daniel M. Bikel and David Chiang. 2000. Two statistical parsing models applied to the chinese treebank. In *Proceedings of the 2nd Chinese Language Processing Workshop*, Hong Kong, China.

Eugene Charniak. 2000. A Maximum-Entropy-Inspired Parser. In *Proc. of NAACL-2000*.

Keh-Jiann Chen, Chu-Ren Huang, Feng-Yi Chen, Chi-Ching Luo, Ming-Chung Chang, and Chao-Jan Chen. to appear. Sinica Treebank: Design Criteria, representational issues and immplementation. In Anne Abeille, editor, *Building and Using Syntactically Annotated Corpora*. Kluwer.

David Chiang. 2000. Statisitical parsing with an automatically-extracted tree adjoining grammar. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 456-463, Hong Kong.

Mike Collins. 1997. Three Generative, Lexicalised Models for Statistical Parsing. In *Proc. of ACL-1997*.

Mike Collins. 2000. Discriminative Reranking for Natural Language Parsing. In *Proc. of ICML-2000*.

N. Ide and J. Vronis. 1998. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*, Las Palmas, Spain.

Beth Levin. 1993. *English Verbs and Alternations: A Preliminary Investigation*. Chicago: The Unversity of Chicago Press.

Charles Li and Sandra Thompson. 1976. Subject and Topic: A new typology of language. In Charles Li, editor, *Subject and Topic*. Academic Press.

M. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. submitted. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Journal of Natural Language Engineering*.

Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. In *Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece.