Test Sets for Chinese Nonlocal Dependency Parsing

Manjuan Duan, William Schuler
Department of Linguistics
The Ohio State University
1712 Neil Ave. Columbus OH, 43210
{duan.59, schuler.77}@osu.edu

Abstract

Chinese is a language rich in nonlocal dependencies. Correctly resolving these dependencies is crucial in understanding the predicate-argument structure of a sentence. Making full use of the trace annotations in the Penn Chinese Treebank (Xue et al., 2005), this research contributes several test sets of Chinese nonlocal dependencies which occur in different grammatical constructions. These datasets can be used by an automatic dependency parser to evaluate its performance on nonlocal dependency resolution in various syntactic constructions in Chinese.

Keywords: nonlocal dependency, parsing, Mandarin Chinese

1. Introduction

Recovering unbounded dependencies is challenging in English, as reported by Rimell et al. (2009). However, it serves as an important test of an automatic parser which cannot be easily accomplished by shallow language techniques. In spite of the low frequency of some nonlocal dependency constructions, correctly resolving these dependencies is crucial in understanding the underlying predicate-argument structure of a sentence.

Although trace categories are annotated in the Penn Treebank (Xue et al., 2005), few state-of-the-art constituent parsers make use of these annotations to make predictions of the nonlocal dependencies. Categorial grammar annotations of the Treebank are advocated partly for their well-defined representations of filler-gap phenomena in natural languages. Parsers trained on categorial grammar annotations are found to produce superior performance in recovering nonlocal dependencies in English (Rimell et al., 2009; Nguyen et al., 2012).

In this research we focus on various nonlocal dependency types in Mandarin Chinese, a language that makes heavy use of nonlocal dependencies (Kummerfeld et al., 2013). We make full use of the trace categories annotated in the Penn Chinese Treebank (Xue et al., 2005) to generate test sets for eight nonlocal dependency constructions. We evaluate the nonlocal dependency recovery performance of parsers trained on generalized categorial grammar annotations (Bach, 1981; Nguyen et al., 2012; Duan and Schuler, 2015) with the annotated test sets. We hope these test sets can make it easier to automatically evaluate nonlocal dependency recovery in Mandarin Chinese.

2. Nonlocal Dependency Data Sets

2.1. The Constructions

In a nonlocal dependency, a constituent seems to be moved from its canonical position in the predicate-argument pattern, while it still needs to be interpreted in the position from which it is moved. For this research we examined the trace categories annotated in Penn Chinese Treebank and included those constructions where either both the trace and the filler were clearly annotated or they can reliably be recovered from the syntactic trees. For example, the Chinese Penn Treebank has clear annotations about the locations of traces in relative clauses but the fillers are 'WHNP' empty categories. However, the head noun of a relative clause can always be reliably and accurately located given the tree structure. We therefore include relative clauses as a nonlocal construction in this study.

The only exception is a type of relative clause where the head noun is an adverbial modifier of the relative clause, as shown in (1).

(1) Zhe jiu shi ta jingchang dubo de yuanyin
this exactly is he often gamble de reason
'That's why he often gambles.'

In (1), the head noun *yuanyin* 'reason' is annotated as a moved adverbial modifier of the verb *dubo* 'gamble'. We did not include this type of relative clause as a nonlocal construction because unlike other nonlocal constructions we annotated, the filler in this type of construction cannot be put back into the trace location to form a grammatical sentence in Chinese, which makes the nonlocal dependency controversial. Also this type relative clause has identical syntactic structure to an appositive clause in Chinese. Sometimes, they can be semantically indistinguishable.

We annotated eight types of nonlocal dependencies in which both the trace and filler are reliably available in the Treebank annotations. Here is a brief introduction to each of them.

Subject relative clause

Subject relative clauses are constructions in which a subject is extracted from a relative clause and a trace category is annotated in the subject position in the relative clause. In Mandarin Chinese, a relative clause is followed by a particle *de* and the head noun occurs after the *de* particle.

(2) meiyou yingxiang de xinwen yao diudiao
do-not-have video de news needs dump
'dump the news which do not have videos'
1(*meiyou* 'do-not-have', *xinwen* 'news')

For example, in (2), the head noun *xinwen* 'news' occurs after *de* and serves as the subject of the verb *meiyou* 'do-not-have' in the relative clause in (2). We label the relation '1' between the extracted subject, *xinwen* 'news', and the verb, *meiyou* 'do-not-have' , to indicate the extracted subject is the first argument of the verb. We exclude those cases from our test sets where the head noun of the relative clause is not specified linguistically, so that a concrete dependency can be established between a pair of words.

Object relative clause
Object relative clauses are constructions where the object is extracted from the relative clause as shown in (3).

(3) zhexie dou buneng dailai ta xuyao de xingfu
these all cannot bring he needs de happiness
'All these cannot bring the happiness that he needs'
2(*xuyao* 'need', *xingfu* 'happiness')

In (3), a label '2' is given to the dependency between the extracted object, *xingfu* 'happiness', and the verb, *xuyao* 'need'. It means the extracted object is the direct object of the verb. We have not found any extracted indirect objects in relative clauses in the Treebank data.

Topic relative clause
Topic relative clauses are constructions where the topic is extracted from the relative clause. In all the topic relative clauses we examined, we find there is no predicate-argument relation between the extracted topic and the verb in the relative clause. Instead, the extracted topic has a semantic relation with the subject of the relative clause which is similar to the relation in English expressed by 'the subject *of* the topic'.

(4) huzi changguo toufa de welia
beard longer-than hair de Welia
'Welia, whose beard is longer than his hair'
of-asso(*huzi* 'beard', *welia* 'Welia')

For example, in (4), the extracted topic *Welia* is not an argument of the verb *changguo* 'longer-than'. It is related to the subject *huzi* 'beard' instead. Therefore, we labeled an 'of-asso' dependency between the extracted topic and the subject of the relative clause.

Focus construction
Focus constructions, or sometimes called *even*-constructions in Mandarin Chinese, have the constituent in focus occur pre-verbally.

(5) ta shuijiao de difang dou meiyou
he sleep de place even do-not-have
'He does not even have a place to sleep.'
2(*meiyou* 'do-not-have', *difang* 'place')

In (5), *shuijiao de difang* 'place to sleep' is the constituent in focus. It occurs before the verb *meiyou* 'do-not-have', rather than after the verb, the canonical position of a direct object. Since *shuijiao de difang* 'place to sleep' is the direct object of *meiyou* 'do-not-have', there is a labeled '2' dependency between them.

Passivization of direct objects
Passivizations of direct objects are passivized sentences where the direct object is fronted to become the subject of the sentence.

(6) gonggong changhe chouyan yinggai bei jinzhi
public place smoking should bei forbidden
'Smoking in public should be forbidden'
2(*jinzhi* 'forbid', *chouyan* 'smoking')

Most passivized sentences in the Treebank data belong to this category. We give the label '2' to the dependency between the verb and the fronted subject since the subject is the second argument of the transitive verb.

Passivization of indirect objects
Passivizations of indirect objects are passivized sentences where the indirect object is fronted to be the subject of the sentence.

(7) liangren bei geiyu nanmin shenfen
two-people bei given refugee status
'Two people were given the status of refugee'
3(*geiyu* 'give', *liangren* 'two-people')

We give the label '3' to the dependency between the ditransitive verb and the fronted object since it is the third argument of the verb.

Topicalization
Topicalizations are constructions where a word or phrase is moved to the sentence initial position to serve as a topic of the sentence. The dependency label could be '1' or '2', depending on whether the fronted word or phrase is the first argument or the second argument of the verb.

(8) zhezhong shiqing ni bushi diyici
this-sort-of things you not first-time
pengdao
came-across
'It is not the first time that you came across this sort of things.'
2(*pengdao* 'came-across', *shiqing* 'things')

In (8), the direct object *shiqing* 'things' is moved to the front of the sentence. Therefore a label '2' is given to the dependency between *pengdao* 'came-across' and *shiqing* 'things'.

Extraction from an embedded clause
Extractions from an embedded clauses are constructions where a subject or an object is moved across at least two clause boundaries.

(9) shouji keyi shuo shi zhege shidai de yi bufen
cell-phone can say is this age de a part
'Cell phone, we can say, is a part of this age'
1(*shi* 'is', *shouji* 'cell-phone')

In (9), the noun phrase *shouji* 'cell-phone' is extracted to the sentence initial position from its subject position within the complement clause of *shuo* 'say'. A '1' dependency is annotated between *shi* 'is' and *shouji* 'cell phone' to indicate that 'cell-phone' is the first argument of 'is'.

Construction	# sents	# deps	Freq %
Sbj rel	111	147	23.0
Obj rel	112	127	9.0
Tpc rel	97	104	1.4
Foc	87	96	0.6
Pass direct	110	178	2.0
Pass indirect	61	65	0.2
Topicalization	116	141	0.6
Embedded	42	48	0.1

Table 1: Test data size and frequency of the constructions.

2.2. The Data

We divided the Chinese Penn Treebank 6 into training, development and test sets as indicated by Tse and Curran (2010). From the test set and all the other sentences in the Penn Chinese Treebank 8 that are not included in the training set, for each nonlocal construction, we randomly choose 120 sentences to annotate with dependencies. If we could not find 120 cases for some rare constructions, we include all the occurrences from the test set in order to have a test set of reasonably large size. We annotate all the dependencies according to their trace annotations in the Penn Chinese Treebank. We removed the sentences where either the head or the dependent in the nonlocal dependency is not present linguistically. Sometimes, there is more than one occurrence of a particular construction in one sentence. In these cases, more than one nonlocal dependency is annotated. The size of the test set for each construction and the distribution of the construction in the Penn Chinese Treebank are shown in Table 1.

We can see subject relative clauses and object relative clauses are very common in the corpus. Around a third of sentences in the Penn Chinese Treebank contain at least a subject relative clause or an object relative clause. Focus constructions, topicalizations, passivization with fronted indirect objects and extractions from embedded clauses occur relatively rarely in the corpus. It is possible that some constructions, such as the focus construction, can be more frequent in more colloquial text.

3. Experiments

We first examined the proportion of each test set that can be recovered from Stanford dependencies converted from gold Treebank trees to explore the possibility to evaluate nonlocal dependency recovery on an ideal parser producing Stanford dependencies. Then we evaluated the performance of several automatic parsers on the task of recovering nonlocal dependencies to have a preliminary understanding of how difficult it is to recover each of these nonlocal dependencies.

3.1. Stanford dependencies

Stanford dependencies (Marneffe et al., 2006; Chang et al., 2009) are widely used in many dependency parsing evaluations and can be easily obtained from Treebank annotations. Constituent parsers, such as the Stanford parser (Klein and Manning, 2003), the Berkeley parser (Petrov and Klein, 2007) and the Brown parser (Charniak, 2000), can all be trained on Treebank annotations to yield Stanford dependencies. Dependency parsers, such as MaltParser (Nivre et al., 2006), Mate (Bohnet, 2010) and MSTParser (MacDonal and Pereira, 2006), can be trained directly on Stanford dependencies to predict Stanford dependencies. Therefore we implemented a heuristic extraction of nonlocal dependencies from gold Stanford dependencies and evaluated the results against our annotations. Heuristically, we mapped 'nsubj' in Stanford dependencies into '1' dependencies in our annotations, 'dobj' into '2' dependencies and 'iobj' into a '3' dependencies.

For relative clauses, Chinese Stanford dependencies have a dependency labeled 'rcmod' between the head noun and the main verb of a relative clause, regardless of the type of the relative clause. For example, the Stanford dependencies of (2) will contain 'rcmod(news, do-not-have)' and the Stanford dependencies of (3) will contain 'rcmod(happiness, need)'. However, this dependency label does not provide any information about whether the head noun is the first or second argument of the verb. After examining the statistics of relative clauses in Stanford dependencies of the training set, we found that when the subject is present in the relative clause while the object is missing, the relative clause is most likely to be an object relative clause; when the subject is missing but the object is present, it is most likely to be a subject relative clause; and if both subject and object are absent, it is most likely to be a subject relative clause. By these principles, we mapped the rcmod dependencies into subject or object relative clause dependencies in our annotations. However, it is more difficult to recover nonlocal dependencies for topic relative clauses. For example, Stanford dependencies have 'rcmod(Welia, longer-than)' for (4), while 'Welia' is not a participant of 'longer-than'. Although we can map all the relative clauses where both the subject and object are present into topic relative clauses, our statistics show that relative clauses with both subject and object present are most likely to be a relative clause relativizing an adjective or adverbial modifier as shown in (1). Mapping the relative clause with both subject and object present into topic relative clauses is not supported by our data. Therefore, we cannot have labeled dependency results of topic relative clauses from gold Stanford dependencies (sd-orcl) in Table (2).

For passive voice, Stanford dependencies have a 'nsubjpass' dependency between the subject and the verb in a passivized sentence. In most cases, the subject is the second argument of the verb. Therefore, 'nsubjpass' dependencies are converted into '2' dependencies in our annotations.

Considering the possibility of mismatching the dependency labels between Stanford dependencies and our annotations, we also set up the evaluation of unlabeled dependencies to check whether pairs of words which have nonlocal dependencies in our annotations are related in any type of dependency in Stanford dependencies. The results of recovering nonlocal dependencies, either labeled (L) or unlabeled (U), from gold Stanford dependencies (sd-orcl) are shown in Table 2.

Table 2 shows that some nonlocal dependencies, even when unlabeled, are absent in the Stanford dependencies converted from gold Treebank trees. This could indicate

Construction	L/U	sd-orcl	gcg-s	gcg-l
Sbj rel	L	.81	.51	.50
	U	.81	.57	.61
Obj rel	L	.63	.63	.77
	U	.87	.70	.81
Tpc rel	L	-	.49	.48
	U	-	.53	.50
Focus	L	.28	.06	.33
	U	.97	.82	.80
Pass_do	L	.42	.83	.89
	U	.71	.89	.92
Pass_io	L	0	.60	.62
	U	.69	.74	.88
Topicalization	L	0	.08	.11
	U	.86	.59	.62
Embedded	L	.02	.38	.31
	U	.04	.42	.33

Table 2: Nonlocal dependency recovery results from gold Chinese Stanford dependencies (sd-orcl) and from two automatic parsers trained on generalized categorial grammar annotations of Mandarin Chinese (gcg-s and gcg-l).

that syntactic dependencies extracted from the Penn Treebank are not a close approximation of semantic predicate-argument structures extracted from categorial grammar annotations where filler-gap constructions are well defined in the syntactic derivations.

3.2. A GCG parser

To examine the performance of recovering nonlocal dependencies from an automatic parser, we experimented on parsers trained by the Berkeley latent variable PCFG trainer (Petrov and Klein, 2007) on generalized categorial grammar annotations (Bach, 1981; Nguyen et al., 2012; Duan and Schuler, 2015) of the Penn Chinese Treebank. We experimented with two different training sets, a small training set (gcg-s) and a large training set (gcg-l). The small training set is the same training set used in Tse and Curran (2010) which contains 15,957 sentences. The large training set includes all the sentences from the Penn Chinese Treebank 8 except those sentences used in nonlocal dependency test sets. The large training set contains 50,635 sentences. We trained two parsers with these two training sets and parsed the sentences in the nonlocal dependency test sets with these two parsers. Then we extracted the dependencies out of the parse outputs and evaluated the results against the dependencies annotated in the eight test sets.

4. Discussion

In general, a larger training set is beneficial for recovering nonlocal dependencies, as shown in Table 2. The improvements for object relative clauses and focus constructions are large. Small drops in accuracy are observed for subject relative clauses, topic relative clauses and extractions from an embedded clause. These usually only involve one or two more wrongly predicted dependencies on the side of the large training set.

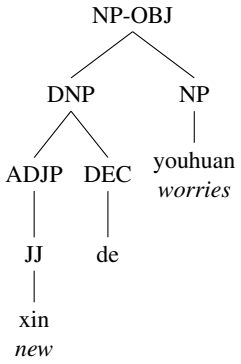

Figure 1: Treebank annotations for *xin de youhuan* 'new worries'

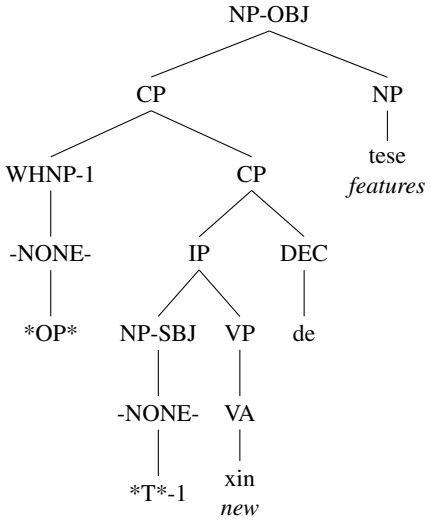

Figure 2: Treebank annotations for *xin de tese* 'new features'

In order to better inspect the parsing errors, we annotated a small development set for seven nonlocal constructions. We did not have a development set for extractions from embedded clauses because they rarely occur in the data. The development set for each construction consists of seven to ten sentences. The error analysis below is conducted based on the parsing outputs of the development sets.

In spite of their frequent occurrence in the corpus, subject relative clauses are not easy to parse correctly, as suggested by the results in Table 2. Examining the failed cases in the development set suggests two potential difficulties for parsing subject relative clauses correctly. The first difficulty is that there are inconsistent Treebank annotations for the same noun phrase construction, as shown in Figure 1 and 2. There is no apparent semantic or syntactic reason to analyze the noun phrases in Figure 1 and 2 differently. This confusion caused the parser to be unable to reliably predict the internal structure of this type of noun phrases.

Another mistake observed in parsing subject relative clauses is caused by noun-verb confusion in Chinese.

(10) shamao de xiansheng
silly de man
'a silly man'

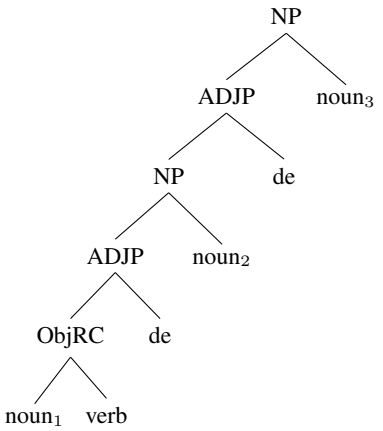

Figure 3: A possible parse for *noun + verb + de + noun + de + noun*

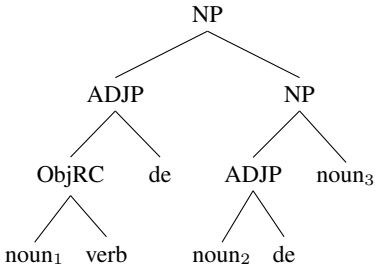

Figure 4: Another possible parse for *noun + verb + de + noun + de + noun*

Shamao 'silly' in Chinese can either be a verb *be silly* or a noun *a silly*. *Shamao* is annotated as a verb in (10) in Treebank annotations and *shamao de* forms a relative clause modifying *xiangsheng* 'man'. However, the gcg-l parser parses the word as a noun and the structure of the noun phrase is 'noun + de + noun', which is also a very common structure for noun phrases in Mandarin Chinese.

The errors in parsing object relative clauses are often caused by the difficulty of predicting correct internal structure for noun phrases. For example, a word sequence of 'noun$_1$ + verb + de + noun$_2$ + de + noun$_3$' is structurally ambiguous between two possible parses as shown in Figures 3 and 4. The parse in Figure 3 yields the dependency 2(verb, noun$_2$) from the object relative clause. The parse in Figure 4 yields the dependency 2(verb, noun$_3$).

The difficulty of parsing topic relative clauses correctly is that there are multiple ways to parse the word sequence of a topic relative clause. A topic relative clause usually has a word sequence of 'noun$_1$ + verb + (noun$_2$) + de + noun$_3$' where noun$_3$ is the topic of the relative clause and we expect to have the dependency of-asso(noun$_1$, noun$_3$). The parser can parse noun$_1$ as adverbial modifier and the rest as a subject relative clause and yield the dependency 1(verb, noun$_3$). If the noun$_2$ is not present, the parser is also likely to parse this word sequence as an object relative clause and yield the dependency 2(verb, noun$_3$). Since subject relative clauses and object relative clauses are both very frequent in the data, these two parsing possibilities are very competitive.

Table2 shows that the parsing of a passivization with a fronted direct object is comparatively easy. This is because the passivization particle *bei* gives an unequivocal indication of the predicate-argument structure of the sentence. Since all ten sentences in the development set are parsed correctly, we speculate that the errors in the test set come from mistaken parsing of some noun phrase involved in the construction.

The parsing of passivizations with fronted indirect objects tends to be more challenging. Error analysis of the development set seems to suggest the parsing confusion between the second and third argument of a ditransitive verb accounts for most of the errors. A sentence 'John gave me a book' can be expressed in Chinese with a passive voice either as 'I + bei + give + a book' or 'A book + bei + give + me'. When presented with the word order 'noun$_1$ + bei + verb + noun$_2$', the parser can have a hard time deciding whether noun$_1$ is the second argument (direct object) or the third argument (indirect object) of the verb.

Focus constructions usually come with the word order 'noun$_1$ + noun$_2$ + verb' where noun$_1$ is the subject and noun$_2$ is the fronted object. However noun$_1$ is often not present linguistically because Chinese is a pro-drop language. In that case, noun$_2$ is often parsed mistakenly as the subject. Even if both noun$_1$ and noun$_2$ are both present, there is also chance to parse either noun$_1$ or noun$_2$ as an adverbial modifier because of the relatively low frequency of focus constructions in the corpus.

Table 2 suggests that it is hard to correctly resolve the nonlocal dependencies in topicalizations and extractions from embedded clauses. The fact that these two constructions occur very rarely in the data definitely contributes to the difficulty. Also, unlike relative clauses and passivized sentences, there is no syntactic marker to indicate the possible predicate-argument structure. The distance between the head and the dependent in these two constructions can be very long. An extra noun phrase at the beginning of a sentence can be parsed as a modifier or an adverbial rather than an extracted argument. Correctly resolving the nonlocal dependencies in these two constructions is very challenging.

5. Conclusions

In this study we contribute several test sets for nonlocal dependencies occurring in various constructions in Mandarin Chinese. The poor match between gold Chinese Stanford dependencies and the nonlocal dependencies we annotated indicates that Stanford dependencies cannot provide sufficiently specific information for nonlocal dependencies. The preliminary parsing results suggest that resolving nonlocal dependencies can be a challenging task for some nonlocal constructions. In the future, we are going to experiment with more dependency parsers against the test sets. We hope the availability of these test sets can help to promote better performance of nonlocal dependency recovery in Mandarin Chinese.

6. Bibliographical References

Bach, E. (1981). Discontinuous constituents in generalized categorial grammars. *Proceedings of the Annual Meeting of the Northeast Linguistic Society (NELS)*, 11:1–12.

Bohnet, B. (2010). Top accurancy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 89–97.

Chang, P.-C., Tseng, H., Jurafsky, D., and Manning, C. D. (2009). Discriminative Reordering with Chinese Grammatical Relations Features. In *Proceedings of the third Workshop on Syntax and Structure in Statistical Translation*, pages 51–59.

Charniak, E. (2000). A maximum-entropy inspired parser. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL'00)*, pages 132–139, Seattle, Washington.

Duan, M. and Schuler, W. (2015). Parsing Chinese with a Generative Categorial Grammar. In *Proceedings of Grammar Engineering Across Frameworks*, pages 25–32.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan.

Kummerfeld, J. K., Tse, D., Curran, J. R., and Klein, D. (2013). An empirical examination of challenges in chinese parsing. In *Proceedings of ACL'13*, pages 98–103, Sofia, Bulgaria.

MacDonal, R. and Pereira, F. (2006). Online learning of approximate dependency parsering algorithms. In *Proceedings of the 11th Conference of the European Chapter of the ACL (EACL 2006)*, pages 81–88.

Marneffe, M.-C. d., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*.

Nguyen, L., van Schijndel, M., and Schuler, W. (2012). Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of COLING '12*, pages 2125–2140, Mumbai, India.

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *In Proc. of LREC*, pages 2216–2219.

Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *Proceedings of NAACL HLT 2007*, pages 404–411, Rochester, New York, April. Association for Computational Linguistics.

Rimell, L., Clark, S., and Steedman, M. (2009). Unbounded dependency recovery for parser evaluation. In *Proceedings of EMNLP 2009*, volume 2, pages 813–821.

Tse, D. and Curran, J. R. (2010). Chinese CCGbank: extracting CCG derivations from the penn chinese treebank. In *Proceedings of COLING '10*, pages 1083–1091.

Xue, N., Xian, F., Chiou, F.-D., and Palmer, M. (2005). The Penn Chinese Treebank: Phrase Structure annotation of a large corpus. *Natural Language Engineering*, 11:207–238.