Semi-automatically Developing Chinese HPSG Grammar from the Penn Chinese Treebank for Deep Parsing

Kun Yu[1] **Yusuke Miyao**[2] **Xiangli Wang**[1] **Takuya Matsuzaki**[1] **Junichi Tsujii**[1,3]

1. The University of Tokyo
{kunyu, xiangli, matuzaki, tsujii}
@is.s.u-tokyo.ac.jp

2. National Institute of Informatics
yusuke@nii.ac.jp

3. The University of Manchester

Abstract

In this paper, we introduce our recent work on Chinese HPSG grammar development through treebank conversion. By manually defining grammatical constraints and annotation rules, we convert the bracketing trees in the Penn Chinese Treebank (CTB) to be an HPSG treebank. Then, a large-scale lexicon is automatically extracted from the HPSG treebank. Experimental results on the CTB 6.0 show that a HPSG lexicon was successfully extracted with 97.24% accuracy; furthermore, the obtained lexicon achieved 98.51% lexical coverage and 76.51% sentential coverage for unseen text, which are comparable to the state-of-the-art works for English.

1 Introduction

Precise, in-depth syntactic and semantic analysis has become important in many NLP applications. Deep parsing provides a way of simultaneously obtaining both the semantic relation and syntactic structure. Thus, the method has become more popular among researchers recently (Miyao and Tsujii, 2006; Matsuzaki et al., 2007; Clark and Curran, 2004; Kaplan et al., 2004).

This paper introduces our recent work on deep parsing for Chinese, specifically focusing on the development of a large-scale grammar, based on the HPSG theory (Pollard and Sag, 1994). Because it takes a decade to manually develop an HPSG grammar that achieves sufficient coverage for real-world text, we use a semi-automatic approach, which has successfully been pursued for English (Miyao, 2006; Miyao et al., 2005; Xia, 1999; Hockenmaier and Steedman, 2002; Chen and Shanker, 2000; Chiang, 2000) and other languages (Guo et al., 2007; Cramer and Zhang, 2009; Hockenmaier, 2006; Rehbei and Genabith, 2009; Schluter and Genabith, 2009).

The following lists our method of approach: *(1) define a skeleton of the grammar (in this work, the structure of sign, grammatical principles and schemas)*, *(2) convert the CTB (Xue et al., 2002) into an HPSG-style treebank*, *(3) automatically extract a large-scale lexicon from the obtained treebank*.

Experiments were performed to evaluate the quality of the grammar developed from the CTB 6.0. More than 95% of the sentences in the CTB could be successfully converted, and the extracted lexicon was 97.24% accurate. The extracted lexicon achieved 98.51% lexical coverage and 76.51% sentential coverage for unseen text, which are comparable to the state-of-the-art works for English.

Since grammar engineering has many specific problems in each language, although we used the similar method applied in other languages to develop a Chinese HPSG grammar, it is very different from applying, such as statistical parsing models, to a new language. Lots of efforts have been done for the specific characteristics of Chinese. The contribution of our work is to describe these issues. As a result, a skeleton design of Chinese HPSG is proposed, and for the first time, a robust and wide-coverage Chinese HPSG grammar is developed from real-world text.

2 Design of Grammatical Constraints for Chinese HPSG

Because of the lack of a comprehensive HPSG-based syntactic theory for Chinese, we extended the original HPSG (Pollard and Sag, 1994) to analyze the specific linguistic phenomena in Chinese. Due to space limitations, we will provide a brief sampling of our extensions, and discuss several selected constructions.

2.1 Sign, Principles, and Schemas

Sign, which is a data structure to express grammatical constraints of words/phrases, is modified and extended for the analysis of Chinese specific constructions, as shown in Figure 1. *PHON, MOD, SPEC, SUBJ, MARKING,* and *SLASH* are

Coling 2008: Poster Volume, pages 1417–1425,
Beijing, August 2010

features defined in the original HPSG, and they represent the phonological information of a word, the constraints on the modifiee, the specifiee, the subject, the marker, and the long-distance dependency, respectively. *COMPS*, which represents the constraints on complements, is divided into *LCOMPS* and *RCOMPS*, to distinguish between left and right complements. Aspect, question, and negation particles are treated as markers as done in (Gao, 2000), which are distinguished by *ASPECT*, *QUESTION*, and *NEGATION*. *CONT* is also originated from Pollard and Sag (1994), although it is used to represent semantic structures with predicate-argument dependencies. *TOPIC* and *CONJ* are extended features that represent the constraints on the topic and the conjuncts of coordination. *FILLER* is another extended feature that records the grammatical function of the moved argument in a long-distance dependency.

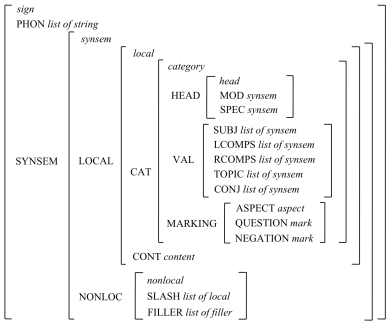


Figure 1. HPSG sign for Chinese.

The principles, including *Phonology Principle*, *Valence Principle*, *Head Feature Principle*, and *Nonlocal Feature Principle*, are implemented in our Chinese HPSG grammar as defined in (Pollard and Sag, 1994). *Semantic Principle* is slightly modified so that it composes predicate-argument structures.

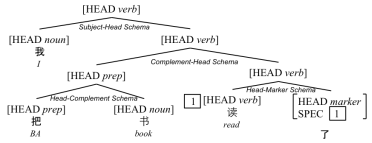
14 schemas are defined in our grammar, among which the *Coord-Empty-Conj Schema*, *Relative-Head Schema*, *Empty-Relativizer Schema*, and *Topic-Head Schema* are designed specifically for Chinese. The other 10 schemas are borrowed from the original HPSG theory.

15 Chinese constructions are considered in our current grammar (refer to Table 1). A detailed description of some particular constructions will be provided in the following subsection.

2.2 An HPSG Analysis for Chinese

2.2.1 BA Construction

The BA construction moves the object of a verb to the pre-verbal position. For example, the sentence in Figure 2 with the original word order is '我/I 读/read 了 书/book'. There were three popular ways to address the BA construction: as a verb (Huang, 1991; Bender, 2000), preposition (Gao, 1992), and case marker (Gao, 2000). Since the aspect markers, such as '了', cannot attach to BA, we exclude the analysis of treating BA as a verb. Because BA, like prepositions, always appears before a noun phrase, we therefore follow the analysis in Gao (1992), and treat BA as a preposition. As shown in Figure 2, BA takes a moved object as a complement, and attaches to the verb as a left-complement.

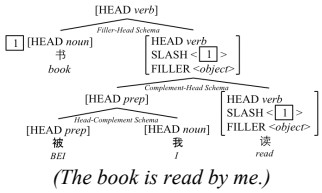


(I read the book.)

Figure 2[1]. Analysis of BA construction.

2.2.2 BEI Construction

The BEI construction is used to make the passive voice of a sentence. Because the aspect marker also cannot attach to BEI, we do not treat BEI as a verb, as done in the CTB. Similar to the analysis of BA construction, we regard BEI as a preposition that attaches to the verb as a left-complement. Additionally, because we can insert a clause '小李/Li 派/send 人/person' between the moved object '他/he' and the verb '打/beat', as is the case for '他/he 被/BEI 小李/Li 派/send 人/person 打/beat 了 (He was beaten by the person that is sent by Li)', we treat the relation between the moved object and the verb as a long-distance dependency. Figure 3 exemplifies our analysis of the BEI construction, in which the *Filler-Head Schema* is used to handle the long-distance dependency, and the *FILLER* feature is used to record that the role of the moved argument.



(The book is read by me.)

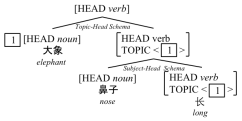
Figure 3. Analysis of BEI construction.

2.2.3 Topic Construction

As indicated in Li and Thompson (1989), a topic refers to the theme of a sentence, which always

[1] In the figures in this paper, we will show only selected features that are relevant to the explanation.

appears before the subject. The difference between the topic and subject is the subject must always have a direct semantic relationship with the verb in a sentence, whereas the topic does not. There are two types of topic constructions. In the first type, the topic does not fill any argument slots of the verb, such as the topic '大象/elephant' in Figure 4. In the second type, the topic has a semantic relationship with the verb. For example, in the sentence '他/he 我/I 喜欢/like (I like him)', the topic '他/he' is also an object of '喜欢/like'. For the first type, we define the *Topic-Head Schema* to describe the topic construction (refer to Figure 4). For the second type, we follow the same analysis as in English, and use the *Filler-Head Schema*.

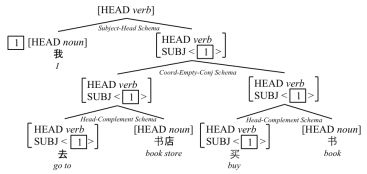


(The nose of an elephant is long.)

Figure 4. Analysis of topic construction.

2.2.4 Serial Verb Construction

In contrast to the definition of serial verb construction in Li and Thompson (1989), we specify a serial verb construction as a special type of verb phrase coordination, which describes several separate events with no conjunctions inside. Similar to ordinary coordination, the verb phrases in a serial verb construction share the same syntactic subject (Muller and Lipenkova, 2009), topic, and left-complement. We define *Coord-Empty-Conj Schema* to deal with it. Figure 5 shows an example analysis.

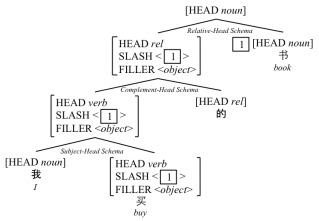


(I go to the book store and buy a book.)

Figure 5. Analysis of serial verb construction.

2.2.5 Relative Clause

In Chinese, a relative clause is marked by a relativizer '的' and exists in the left of the head noun. Because Chinese noun phrases are right-headed in general, we analyze a relative clause as a nominalization that modifies a head noun (Li and Thompson, 1989). Inside of a relative clause, the relativizer is treated as head. When the relativizer is omitted, we define a unary schema, *Empty-Relativizer Schema*, which functions by combining a relative clause with an empty relativizer. Furthermore, we introduce a *Relative-Head Schema* to handle the long-distance dependency for the extracted argument[2] (refer to Figure 6).



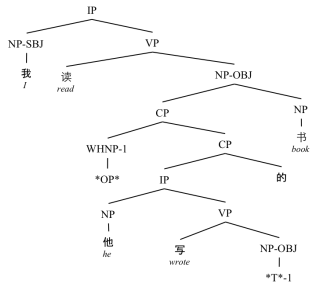
(the book that I buy)

Figure 6. Analysis of relative clause.

3 Converting the CTB into an HPSG Treebank

3.1 Partially-specified Derivation Tree Annotation

In order to convert the CTB into an HPSG treebank, we first annotate the bracketing trees in the CTB to be partially-specified derivation trees[3], which conform to the grammatical constraints designed in Section 2. Three types of rules are defined to fulfill this annotation.



(I read the book that he wrote.)

Figure 7. The CTB annotation for a sentence.

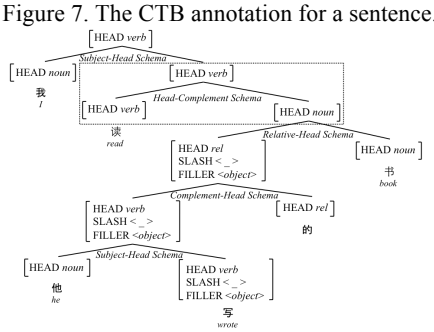


Figure 8. Partially-specified derivation tree for Figure 7.

For example, Figure 7 shows the bracketing tree of a sentence in the CTB, while Figure 8 shows the partially-specified derivation tree after re-annotation.

[2] The extracted adjunct is not treated as a long-distance dependency in our current grammar.

[3] *Partially-specified derivation tree* means a tree structure that is annotated with schema names and some features of the HPSG signs (Miyao, 2006).

3.1.1 Rules for Annotation Conversion

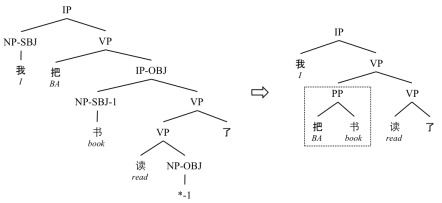
In the CTB, there exist some annotations that do not coincide with our HPSG analysis for Chinese. Therefore, we define pattern rules to convert the annotations in the CTB to fit with our HPSG analysis. 76 annotation rules are defined for 15 Chinese constructions (refer to Table 2). Due to page constraints, we focus on the constructions that we discussed in Section 2.

Construction	Rule #
Relative clause	20
BEI construction	21
Coordination	7
Subject/object control	5
Non-verbal predicate	4
Logical subject	3
Right node raising	3
Parenthesis	3
BA construction	3
Aspect/question/negation particle	2
Subordination	1
Serial Verb construction	1
Modal verb	1
Topic construction	1
Apposition	1

Table 1. Chinese constructions and annotation rules.

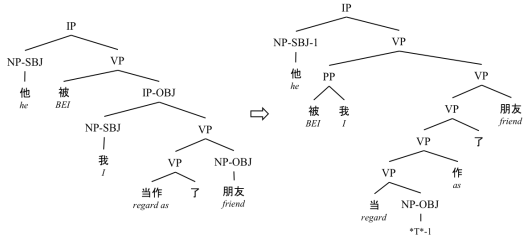
Rules for BA and BEI Construction

As analyzed in Section 2, we treat BA and BEI as prepositions that attach to the verb as left-complements. However, in the CTB, BA and BEI are annotated as verbs that take a sentential complement (Xue and Xia, 2000). By applying the annotation rules, the BA/BEI and the subject of the sentential complement of BA/BEI are re-annotated as a prepositional phrase (as indicated in the dash-boxed part in Figure 9).



(I read the book.)

Figure 9. Conversion of BA construction.



(He is regarded as a friend by me.)

Figure 10. Verb division in BEI construction.

In addition, in the CTB, some BA/BEI constructions are not annotated with trace, which makes it difficult to retrieve the semantic relation between the verb and the moved object. The principal reason for this is that the moved object in these constructions has a semantic relation with only part of the verb. For example, in Figure 10, the moved noun '他/he' is the object of '当/regard', but not for '当作/regard as'. Analysis shows that only a closed set of characters (e.g. '作/as') can be attached to verbs in such a case. Therefore, we manually collect these characters from the CTB, and then define pattern rules to automatically split the verb, which ends with the collected characters, in the BA and BEI construction. Finally, we annotate trace for the split verb. Figure 10 exemplifies the conversion of an example sentence.

Rules for Topic Construction

In the CTB, a functional tag 'TPC' is used to indicate a topic (Xue and Xia, 2000). Therefore, we use this functional tag to detect topic phrases during conversion.

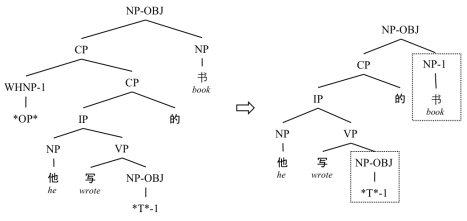
Rules for Serial Verb Construction

We define pattern rules to detect the parallel verb phrases with no conjunction inside (as shown in Figure 11), and treat these verb phrases as a serial verb construction. However, when the verb in the first phrase is a modal verb, such as the case of '我/I 想/want to 唱歌/sing (I want to sing)', the parallel verb phrases should not be treated as a serial verb construction. Therefore, a list of modal verbs is manually collected from the CTB to filter out these exceptional cases during conversion.

(go downstairs and eat meal)

Figure 11. An example of parallel verb phrases.

Rules for Relative Clause



(the book that he wrote)

Figure 12. Conversion of relative clause.

We define annotation rules to slightly modify the annotation of a relative clause in CTB, as shown in Figure 12, to make the tree structure easy to be

analyzed. Furthermore, in CTB, relative clauses are annotated with both extracted arguments and extracted adjuncts. But in our grammar, we only deal with extracted arguments, and the gap in a relative clause (as indicated in the dash-boxed part in Figure 12). When the extracted phrase is an adjunct of the relative clause, we simply view the clause as a modifier of the extracted phrase.

3.1.2 Rules for Correcting Inconsistency

There are some inconsistencies in the annotation of the CTB, which presents difficulties for performing the derivation tree annotation. Therefore, we define 49 rules, as done in (Hockenmaier and Steedman, 2002) for English, to mitigate inconsistencies before annotation (refer to Table 3).

3.1.3 Rules for Assisting Annotation

We also define 48 rules (refer to Table 2), which are similar to the rules used in (Miyao, 2006) for English, to help the derivation tree annotation. For example, 12 pattern rules are defined to assign the schemas to corresponding constituents.

Rule Type	Rule Description	Rule #
Rules for correcting inconsistent annotation	Fix tree annotation	37
	Fix phrase tag annotation	5
	Fix functional tag annotation	5
	Fix POS tag annotation	2
Rules for assisting annotation	Slash recognization	27
	Schema assignment	12
	Head/Argument/Modifier marking	8
	Binarization	1

Table 2. Rules for correcting inconsistency and assisting annotation.

3.2 HPSG Treebank Acquisition

In this phase, the schemas and principles are applied to the annotated partially-specified trees, in order to fill out unspecified constraints and validate the consistency of the annotated constraints. In effect, an HPSG treebank is obtained.

For instance, by applying the *Head-Complement Schema* to the dash-boxed nodes in Figure 8, the constraints of the right daughter are percolated to *RCOMPS* of the left daughter (as indicated as 4 in Figure 13). After applying the schemas and the principles to the whole tree in Figure 8, a HPSG derivation tree is acquired (refer to Figure 13).

3.3 Lexicon Extraction

With the HPSG treebank acquired in Section 3.2, we automatically collect lexical entries as the combination of words and lexical entry templates from the terminal nodes of the derivation trees. For example, from the HPSG derivation tree shown in Figure 13, we obtain a lexical entry for the word '写/write' as shown in Figure 14.

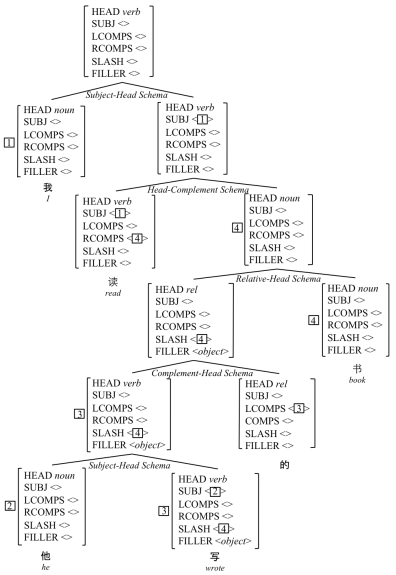


Figure 13. HPSG derivation tree for Figure 8.

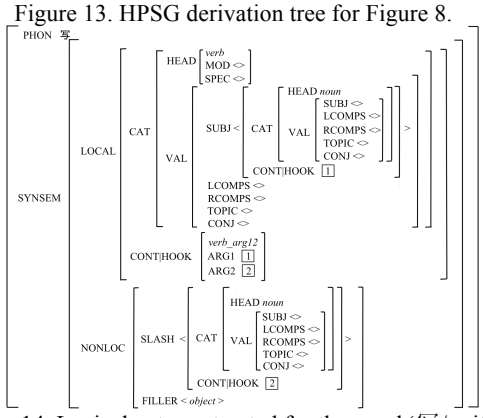
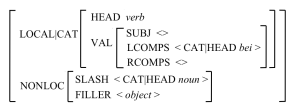
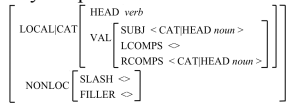


Figure 14. Lexical entry extracted for the word '写/write'.

3.3.1 Lexical Entry Template Expansion



(a) Lexical entry template for the verb in BEI construction



(b) Lexical entry template for the verb in original word order

Figure 15. Application of a lexical rule.

Some Chinese constructions change the word order of sentences, such as the BA/BEI constructions. Therefore, we apply lexical rules (Nakanishi et al., 2004) to the lexical entry templates to convert them into those for the original word order, and expand the lexical entry templates consequently. 18 lexical rules are defined for the verbs in the BA/BEI constructions. For example, by applying a lexical rule to the lexical entry template in Figure 15(a), the moved object indi-

cated by *SLASH* is restored into *RCOMPS*, and the subject introduced by BEI in *LCOMPS* is restored into *SUBJ* (refer to Figure 15(b)).

3.3.2 Mapping of Semantics

In our grammar, we use predicate-argument dependencies for semantic representation. 44 types of predicate-argument relations are defined to represent the semantic structures of 13 classes of words. For example, we define a predicate-argument relation '*verb_arg12*', in which a verb takes two arguments '*ARG1*' and '*ARG2*', to express the semantics of transitive verbs. 72 semantics mapping rules are defined to associate these predicate-argument relations with the lexical entry templates. Figure 16 exemplifies a semantics mapping rule. The input of this rule is the lexical entry template (as shown in the left part), and the output is a predicate-argument relation '*verb_arg12*' (as shown in the right part), which associates the syntactic arguments *SUBJ* and *SLASH* with the semantic arguments *ARG1* and *ARG2* (as indicated by 1 and 2 in Figure 16).

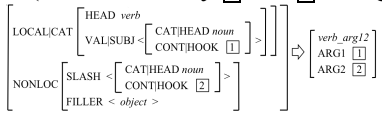


Figure 16. A semantics mapping rule.

4 Evaluation

4.1 Experimental Setting

We used the CTB 6.0 for HPSG grammar development and evaluation. We split the corpus into development, testing, and training data sets, following the recommendation from the corpus author. The development data was used to tune the design of grammar constraints and the annotation rules. However, the testing data set was reserved for further evaluation on parsing. Thus, the training data was further divided into two parts for training and testing in this work. During the evaluation, unknown words were handled in the same way as done in (Hockenmaier and Steedman, 2002).

4.2 Evaluation Metrics

In order to verify the quality of the grammar developed in our work, we evaluated the extracted lexicon by the accuracy for assessing the semi-automatic conversion process, and the coverage for quantifying the upper-bound coverage of the future HPSG parser based on this grammar.

The accuracy of the extracted lexicon was evaluated by *lexical accuracy*, which counts the number of the correct lexical entries among all the obtained lexical entries.

In addition, two evaluation metrics as used in (Hockenmaier and Steedman, 2002; Xia, 1999; Miyao, 2006) were used to evaluate the coverage of the obtained lexicon. The first one is *lexical coverage* (Hockenmaier and Steedman, 2002; Xia, 1999), which means that the percentage that the lexical entries extracted from the testing data are covered by the lexical entries acquired from the training data. The second one is *sentential coverage* (Miyao, 2006): a sentence is considered to be covered only when the lexical entries of all the words in this sentence are covered.

4.3 Results of Accuracy

Since there was no gold standard data for the automatic evaluation of accuracy, we randomly selected 100 sentences from the testing data, and manually checked the lexical entries extracted from these sentences. Results show that 1,558 lexical entries were extracted at 97.24% (1,515/1,558) accuracy.

Error analysis shows all the incorrect lexical entries came from the error in the derivation tree annotation. For example, our current design failed to find the correct boundary of coordinated noun phrases when the word '等/etc' was attached at the end, such as '产权/property right 出让/selling 、 资产/assets 出租/renting 等/etc (property right selling and assets renting etc.)'. We will improve the derivation tree annotation to solve this issue.

4.4 Results of Coverage

Table 3 shows the coverage of the extracted lexical entries, which indicates that a large HPSG lexicon was successfully extracted from the CTB for unseen text, with reasonable coverage. The statistics of the HPSG lexicon extraction in our experiments (refer to Table 4) also indicates that we successfully extracted lexical entries from more than 95% of the sentences in the CTB.

Among all the uncovered lexical entries, 78.55% are for content words, such as verb and noun. In addition, the classification of uncovered lexical entries in Table 4 indicates that about 1/3 of the uncovered lexical entries came from the unknown lexical entry templates ('+*w*/-*t*'). We analyzed the 193 '+*w*/-*t*' failures in the testing data, among which 169 failures resulted from the shortage of training data, which indicated that the correct lexical entry template did not appear in

the training data. The learning curve in Figure 17 shows that we can resolve this issue by enlarging the training data. The other 24 failures came from the error in the derivation tree annotation. For example, our current grammar failed at detecting the coordinated clauses when they were separated by a colon. We will be able to reduce this type of failure by improving the derivation tree annotation.

Sent. Cov.	Lex. Cov.	Uncovered Lexical Entries	
		+w/+t	*+w/-t*
76.51%	98.51%	1.05%	0.43%

Table 3[4]. Coverage of extracted HPSG lexicon.

Data Set	Total Sent #	Succeed Sent #	Word #	Lexical Entry Template #
Training	20,230	19,257(95.19%)	510,815	4,836
Develop	2,067	2,009(97.19%)	55,714	1,582
Testing	2,000	1,941(97.05%)	44,924	1,163

Table 4. Statistics of HPSG lexicon extraction.

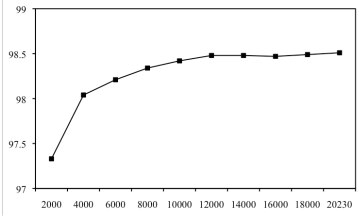


Figure 17. Lexical coverage (Y axis) vs. corpus size (X axis).

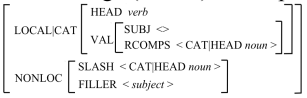


Figure 18. A lexical entry template extracted from testing data.

The other type of failures (‘*+w/+t*’) indicate that a word was incorrectly associated with a lexical entry template, even though both of them existed in the training data. Error analysis shows that 64.39% of failures were related to verbs. For example, for a relative clause ‘投资/invest 台湾/Taiwan 的 商人/businessman (the businessman that invests Taiwan)’ in the testing data, we associated a lexical entry template as shown in Figure 18 with the verb ‘投资/invest’. In the training data, however, the lexical entry template shown in Figure 18 cannot be extracted for ‘投资/invest’, since this word never appears in a relative clause with an extracted subject. Introducing lexical rules to expand the lexical entry template of verbs in a relative clause is a possible way to solve this problem.

4.5 Comparison with Previous Work

Guo’s work (Guo et al., 2007; Guo, 2009) is the only previous work on Chinese lexicalized grammar development from the CTB, which induced wide-coverage LFG resources from the CTB. By using the hand-made gold-standard f-structures of 200 sentences from the CTB 5.1, the LFG f-structures developed in Guo’s work achieved 96.34% precision and 96.46% recall for unseen text (Guo, 2009). In our work, we applied the similar strategy in evaluating the accuracy of the developed Chinese HPSG grammar, which achieved 97.24% lexical accuracy on 100 unseen sentences from the CTB 6.0. When evaluating the coverage of our grammar, we used a much larger data set (including 2,000 unseen sentences), and achieved 98.51% lexical coverage. Although these results cannot be compared to Guo’s work directly because of the different size and content of data set, it indicates that the Chinese HPSG grammar developed in our work is comparable in quality with Guo’s work.

In addition, there were previous works about developing lexicalized grammar for English. Considering the small size of the CTB, in comparison to the Penn Treebank used in the previous works, the results listed in Table 5 verify that, the quality of the Chinese HPSG grammar developed in our work is comparable to these previous works.

Previous Work	Sent. Cov.	Lex. Cov.
Miyao (2006)	82.50%	98.97%
Hockenmaier and Steedman (2002)	-	98.50%
Xia (1999)	-	96.20%

Table 5. Evaluation results of previous work.

4.6 Discussion

There are still some sentences in the CTB from which we failed to extract lexical entries. We analyzed the 59 failed sentences in the testing data and listed the reasons in Table 6.

Reason	Sent #
Error in the derivation tree annotation	31
Short of semantics mapping rule	23
Inconsistent annotation in the CTB	5

Table 6. Reasons for lexicon extraction failures.

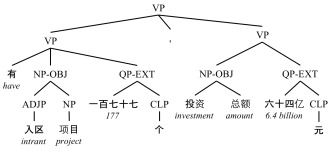
The principal reason for 31 sentence failures, is the error in the derivation tree annotation. For instance, our current annotation rules could convert the regular relative clause shown in Figure 12. Nonetheless, when the relative clause is inside of a parenthesis, such as ‘“ 原始/primitive 的 ”方法/method (the method that is primitive)’, the annotation rules failed at finding the extracted head noun to create a derivation tree. This type of failure can be reduced by improving the annotation rules.

[4] ‘*+w/+t*’ means both the word and lexical entry template have been seen in the lexicon. ‘*+w/-t*’ means only the word has been seen in the lexicon (Hockenmaier and Steedman, 2002).

The second reason, for which 23 sentences failed, is the shortage of the semantics mapping rules. For example, we did not define semantics mapping rule for a classifier that acts as a predicate with two topics. This type of failure can be reduced by adding semantic mapping rules.

The last reason for sentence failures is inconsistencies in the CTB annotation. In our future work, these inconsistencies will be collected to enrich our inconsistency correction rules.

In addition to the reasons above, some sentences with special constructions in the development and training data also could not be analyzed by our current grammar, since the special construction is difficult for the current HPSG to analyze. The special constructions include the argument-cluster coordination shown in Figure 19. Introducing the similar rules used in CCG (Hockenmaier and Steedman, 2002) could be a possible solution to this problem.



(have 177 intrant projects and 6.4 billion investments)

Figure 19. An argument-cluster coordination in CTB.

5 Related Work

To the extent of our knowledge, the only previous work about developing Chinese lexicalized grammar from treebanks is Guo's work (Guo et al., 2007; Guo, 2009). An LFG-based parsing using wide-coverage LFG approximations induced from the CTB was done in this work. However, they did not train a deep parser based on the LFG resources obtained in their work, but relied on an external PCFG parser to create c-structure trees, and then mapped the c-structure trees into f-structures using their annotation rules (Guo, 2009). In contrast to Guo's work, we paid particular attention to a different grammar framework, i.e. HPSG, with the analysis of more Chinese constructions, such as the serial verb construction. In addition, in our on-going deep parsing work, we use the developed Chinese HPSG grammar, i.e. the lexical entries, to train a full-fledged HPSG parser directly.

Additionally, there are some works that induce lexicalized grammar from corpora for other languages. For example, by using the Penn Treebank, Miyao et al. (2005) automatically extracted a large HPSG lexicon, Xia (1999), Chen and Shanker (2000), Hockenmaier and Steedman (2002), and Chiang (2000) invented LTAG/CCG specific procedures for lexical entry extraction. From the German Tiger corpus, Cramer and Zhang (2009) constructed a German HPSG grammar; Hockenmaier (2006) created a German CCGbank; and Rehbei and Genabith (2009) acquired LFG resources. In addition, Schluter and Genabith (2009) automatically obtained wide-coverage LFG resources from a French Treebank. Our work implements a similar idea to these works, but we apply different grammar design and annotation rules, which are specific to Chinese. Furthermore, we obtained a comparative result to state-of-the-art works for English.

There are some researchers who worked on Chinese HPSG grammar development manually. Zhang (2004) implemented a Chinese HPSG grammar using the LinGO Grammar matrix (Bender et al., 2002). Only a few basic constructions were considered, and a small lexicon was constructed in this work. Li (1997) and Wang et al. (2009) designed frameworks for Chinese HPSG grammar; however, only small grammars were implemented in these works.

Furthermore, some linguistic works focused mainly on the discussion of specific Chinese constructions in the HPSG or LFG framework, without implementing a grammar for real-world text (Bender, 2000; Gao, 2000; Li and McFetridge, 1995; Li, 1995; Xue and McFetridge, 1995; Wang and Liu, 2007; Ng, 1997; Muller and Lipenkova, 2009; Liu, 1996; Kit, 1998).

6 Conclusion and Future Work

In this paper, we described the semi-automatic development of a Chinese HPSG grammar from the CTB. Grammatical constraints are first designed by hand. Then, we convert the bracketing trees in the CTB into an HPSG treebank, by using pre-defined annotation rules. Lastly, we automatically extract lexical entries from the HPSG treebank. We evaluated our work on the CTB 6.0. Results indicated that a large HPSG lexicon was successfully extracted with a 97.24% accuracy. Furthermore, our grammar achieved 98.51% lexical coverage and 76.51% sentential coverage for unseen text.

This is an ongoing work, and there are some future works under consideration, including enriching the design of annotation rules, introducing more semantics mapping rules, and adding lexical rules. In addition, the work on Chinese HPSG parsing is on-going, within which the Chinese HPSG grammar developed in this work will be available soon.

References

Emily Bender. 2000. The Syntax of Madarin Ba: Reconsidering the Verbal Analysis. *Journal of East Asian Linguistics*. 9(2): 105-145.

Emily Bender, Dan Flickinger, and Stephan Oepen. 2002. The Grammar Matrix: An Open-source Starter-lit for the Rapid Development of Cross-linguistically Consistent Broad-coverage Precision Grammars. *Procedings of the Workshop on Grammar Engineering and Evaluation*.

John Chen and Vijay K. Shanker. 2004. Automated Extraction of TAGs from the Penn Treebank. *Proceedings of the 6th IWPT*.

David Chiang. 2000. Statistical Parsing with an Automatically-extracted Tree Adjoining Grammar. *Proceedings of the 38th ACL*. 456-463.

Stephen Clark and James R. Curran. 2004. Parsing the WSJ Using CCG and Log-linear Models. *Proceedings of the 42nd ACL*.

Bart Cramer and Yi Zhang. 2009. Construction of a German HPSG Grammar from a Detailed Treebank. *Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks*.

Qian Gao. 1992. *Chinese Ba Construction: its Syntax and Semantics*. Technical report.

Qian Gao. 2000. *Argument Structure, HPSG and Chinese Grammar*. Ph.D. Thesis. Ohio State University.

Yuqing Guo. 2009. *Treebank-based acquisition of Chinese LFG Resources for Parsing and Generation*. Ph.D. Thesis. Dublin City University.

Yuqing Guo, Josef van Genabith and Haifeng Wang. 2007. Acquisition of Wide-Coverage, Robust, Probabilistic Lexical-Functional Grammar Resources for Chinese. *Proceedings of the 12th International Lexical Functional Grammar Conference (LFG 2007)*. 214-232.

Julia Hockenmaier. 2006. Creating a CCGbank and a wide-coverage CCG lexicon for German *Proceedings of COLING/ACL 2006*.

Julia Hockenmaier and Mark Steedman. 2002. Acquiring Compact Lexicalized Grammars from a Cleaner Treebank. *Proceedings of the 3rd LREC*.

C-R Huang. 1991. Madarin Chinese and the Lexical Mapping Theory: A Study of the Interaction of Morphology and Argument Changing. *Bulletin of the Institute of History and Philosophy 62*.

Ronald M. Kaplan et al. 2004. Speed and Accuracy in Shallow and Deep Stochastic Parsing. *Proceedings of HLT/NAACL 2004*.

Chunyu Kit. 1998. Ba and Bei as Multi-valence Prepositions in Chinese. *Studia Linguistica Sinica*: 497-522.

Wei Li. 1995. Esperanto Inflection and its Interface in HPSG. *Proceedings of the 11th North West Linguistics Conference*.

Wei Li. 1997. Outline of an HPSG-style Chinese Reversible Grammar. *Proceedings of the 13th North West Linguistics Conference*.

Wei Li and Paul McFetridge. 1995. Handling Chinese NP Predicate in HPSG. *Proceedings of PACLING-II*.

Charles N. Li and Sandra A. Thompson. 1989. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press, London, England.

Takuya Matsuzaki, Yusuke Miyao, and Junichi Tsujii. 2007. Efficient HPSG Parsing with Supertagging and CFG-filtering. *Proceedings of the 20th IJCAI*.

Yusuke Miyao. 2006. *From Linguistic Theory to Syntactic Analysis: Corpus-oriented Grammar Development and Feature Forest Model*. Ph.D. Thesis. The University of Tokyo.

Yusuke Miyao, Takashi Ninomiya and Junichi Tsujii. 2005. Corpus-oriented Grammar Development for Acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank. *Natural Language Processing - IJCNLP 2005*: 684-693.

Yusuke Miyao and Junichi Tsujii. 2008. Feature Forest Models for Probabilistic HPSG Parsing. *Computational Linguistics*. 34(1): 35-80.

Stefan Muller and Janna Lipenkova. 2009. Serial Verb Constructions in Chinese: A HPSG Account. *Proceedings of the 16th International Conference on Head-Driven Phrase Structure Grammar*. 234-254.

Hiroko Nakanishi, Yusuke Miyao and Junichi Tsujii. 2004. An Empirical Investigation of the Effect of Lexical Rules on Parsing with a Treebank Grammar. *Proceedings of the 3rd TLT*. 103-114.

Say K. Ng. 1997. *A Double-specifier Account of Chinese NPs Using Head-driven Phrase Structure Grammar*. Master Thesis. Department of Linguistics, University of Edinburgh.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press and CSLI Publications, Chicago, IL and Stanford, CA.

Ines Rehbein and Josef van Genabith. 2009. Automatic Acquisition of LFG Resources for German – As Good as it Gets. *Proceedings of the 14th International Lexical Functional Grammar Conference (LFG 2009)*.

Natalie Schluter and Josef van Genabith. 2008. Treebank-based Acquisition of LFG Parsing Resources for French. *Proceedings of the 6th LREC*.

Mark Steedman. 2000. *The Syntactic Process*. The MIT Press.

Xiangli Wang et al. 2009. Design of Chinese HPSG Framework for Data-driven Parsing. *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*.

Lulu Wang and Haitao Liu. 2007. A Description of Chinese NPs Using Head-driven Phrase Structure Grammar. *Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar*. 287-305.

Fei Xia. 1999. Extracting Tree Adjoining Grammars from Bracketed Corpora. *Proceedings of the 5th NLPRS*.

Nianwen Xue, Fudong Chiou, and Martha Palmer. 2002. Building a Large-scale Annotated Chinese Corpus. *Proceedings of COLING 2002*.

Ping Xue and Paul McFetridge. 1995. DP Structure, HPSG, and the Chinese *NP. Proceedings of the 14th Annual Conference of Canadian Linguistics Association*.

Nianwen Xue and Fei Xia. 2000. *The Bracketing Guidelines for the Penn Chinese Treebank*.

Yi Zhang. 2004. Starting to Implement Chinese Resource Grammar using LKB and LinGO Grammar Matrix. Technical report.