Building an HPSG Chinese grammar (Zhong)

Fan, Zhenzhen

2019

Fan, Z. (2019). Building an HPSG Chinese grammar (Zhong). Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/87331>

<https://doi.org/10.32657/10220/48021>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

**BUILDING AN HPSG CHINESE GRAMMAR
(ZHONG)**

**ZHENZHEN FAN
SCHOOL OF HUMANITIES
2018**

Building An HPSG Chinese Grammar (ZHONG)

Zhenzhen Fan

School of Humanities

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2018

Statement of Originality

I certify that all work submitted for this thesis is my original work. I declare that no other person's work has been used without due acknowledgement. Except where it is clearly stated that I have used some of this material elsewhere, this work has not been presented by me for assessment in any other institution or University. I certify that the data collected for this project are authentic and the investigations were conducted in accordance with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

Jan 25, 2019

Date

Fan Zhenzhen

Supervisor Declaration Statement

I have reviewed the content of this thesis and to the best of my knowledge, it does not contain plagiarised materials. The presentation style is also consistent with what is expected of the degree awarded. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accordance with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

Jan 25, 2019

Date

A/Prof Francis Bond

Authorship Attribution Statement

This thesis contains material from 2 papers published in the following peer-reviewed proceedings where I was the first author.

Two sections each from Chapter 4 and Chapter 5 are published as Fan, Zhenzhen, Sanghoun Song and Francis Bond. Building Zhong, a Chinese HPSG Meta-Grammar. In *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar (HPSG 2015)*.

The contributions of the co-authors are as follows:

- A/Prof Bond provided the initial project direction, the guidance in the overall HPSG framework and edited the manuscript drafts.
- Dr. Song gave valuable suggestions and assistance in the implementation of the analysis in codes. He also revised the manuscript drafts.
- I did the literature search, analyzed the phenomena, implemented and tested them in the grammar. I also prepared the manuscript drafts.

Parts of Chapter 3 and Chapter 6 are published as Fan, Zhenzhen, Sanghoun Song, and Francis Bond. An HPSG-based Shared Grammar for the Chinese Languages: Zhong. *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 ACL workshop* (2015). The contributions of the co-authors are as follows:

- A/Prof Bond provided the overall direction and guidance, and edited the manuscript drafts.
- Dr. Song worked closely with me on the enhancement of grammar, especially in helping the implementation of the analyses in codes. He also revised the manuscript drafts.
- I did the literature search, designed and implemented lexical acquisition, analyzed the phenomena, wrote the grammar extensions, tested them and wrote the bulk of the paper. I also presented the paper at the conference and revised it for subsequent publication.

Jan 25, 2019

Date

A handwritten signature in black ink, appearing to be 'Fan Zhenzhen', written on a light gray background.

Fan Zhenzhen

Acknowledgement

This has been a long journey since 2012, juggling my study, a fulltime job, and family along. Now, approaching the end of this journey, I have a lot of people to show my gratitude to. Without them, I would not have survived all the challenges along the way.

First of all, I would like to give my heartiest thank to my supervisor, Francis Bond, who inspired me to work on the topic of my thesis, and who has always been there to give me prompt guidance and help in my area of study. He gives me warm encouragement and kind support throughout this journey, nudging me over numerous hurdles, especially at moments when I was so tempted to surrender to the stresses.

Secondly, I would like to say thank you to my fellow members in the NTU Computational Linguistics Lab. Thanks to Sanghoun Song for easing me into the implementation of HPSG grammar, and numerous help in the work on Zhong; to Michael Goodman for all the support in using DELPH-IN tools; to Luis Morgado da Costa and Wang Wenjie for all the insightful discussions over Chinese during grammar engineering sessions; to David Moeljadi, Lê Tuấn Anh and Hannah Choi for their warm moral support.

I'd also like to show my gratitude to the DELPH-IN community. Specifically, to Emily Bender, who kindly hosted me during my exchange study to University of Washington and gave me the opportunity to work with HPSG veterans and learn from them; to Dan Flickinger, who embraced me as a clumsy apprentice in HPSG grammar engineering and shared with me his rich experience working on ERG. To all who have attended the VLAD sessions, you have given me very useful comments and help in treating the various language phenomena discussed in the sessions. Finally, I would like to express my gratitude to Yi Zhang and Rui Wang, the developers of the MCG and Justin Chunlei Yang, the developer of ManGO: I stand on the

shoulders of giants.

Thanks also have to been given to my boss and long time friend, Dr. Leong Mun Kew, for granting me time away from work when needed, and cheering me through the stresses with inspirational stories and jokes.

Last but not least, I'd like to thank my family, for being the solid foundation on which everything else during the journey becomes possible.

Thank you all, for making it work.

This dissertation is typeset with $\text{\LaTeX}$, using Overleaf (www.overleaf.com). The grammar Zhong and associated tools and treebanks is stored on GitHub (<https://github.com/delph-in/zhong>).

July 2018, Singapore

Contents

Statement of Originality	i
Supervisor Declaration Statement	ii
Authorship Attribution Statement	iii
Acknowledgement	v
Summary	xi
1 Introduction	1
1.1 Statement of the Research Topic	1
1.2 Justification	5
1.3 Structure	7
2 Literature Review	9
2.1 Deep Linguistic Analysis and HPSG	9
2.2 Chinese HPSG	10
2.3 Grammar Sharing	12
2.4 Conclusion	13
3 Lexical Acquisition	15
3.1 The Lexicon before Lexical Acquisition	15
3.2 Related Work	18
3.3 Lexicon Extension Method	19
3.3.1 Resources	19
3.3.2 Approach	21

3.4	Results	22
3.4.1	Test suite creation	22
3.4.2	Lexicon generation from Sinica	23
3.4.3	Lexicon generation from Chinese-English WordNet	28
3.5	Grammar Testing	34
3.6	Conclusion	35
4	Language Phenomena of Concern	37
4.1	Selection of Phenomena	37
4.2	DE Constructions Modifying Nouns	38
4.2.1	Associative DE	39
4.2.2	Relativizing DE	39
4.3	Nominalizing DE Constructions	41
4.4	SUO-DE structure	42
4.5	SHI-DE Constructions	45
4.6	DE as a Sentence Final Particle	49
4.7	Reduplicated Adjectives and Verbs	50
4.8	Interrogative Sentences	52
5	HPSG Modeling of Phenomena	57
5.1	DE Construction Modifying Nouns	60
5.1.1	Associative DE	61
5.1.2	Relativizing DE constructions	70
5.2	Nominalizing DE Constructions	81
5.3	Predicative SHI + Nominalization	85
5.4	SUO-DE structure	88
5.5	Emphatic SHI-DE	93
5.6	DE as a Sentence Final Particle	110
5.7	Reduplicated Adjectives and Verbs	111
5.8	Interrogatives	116

5.8.1	Content questions	116
5.8.2	Particle questions	118
5.9	Other Enhancements of the Grammar	120
6	Grammar Engineering	123
6.1	Preprocessing and Postprocessing	123
6.2	Configuration	124
6.3	Treebanking	124
6.4	Ranking	125
6.5	MAL-Extension	128
7	Conclusion	131
	Bibliography	134

Summary

This thesis describes the development of Zhong, a computational resource grammar for Chinese, in the framework of Head-driven Phrase Structure Grammar (HPSG: Pollard & Sag, 1994) using Minimal Recursion Semantics (Copestake et al., 2005). In order to increase the grammar’s coverage for practical applications, a corpus-driven approach was adopted to systematically expand its lexical and syntactic coverage. The lexicon was expanded through semi-automatic learning lexical entries from an annotated Chinese corpus. Various language phenomena commonly observed in corpora have been analyzed and modeled in the grammar, especially those involving the particle 的 DE. The entire grammar and associated tools are available under an open-source license.

A treebank with 798 sentences has been built with the parse trees from the grammar’s output. With appropriate trees manually selected from the parses, the treebank was used as a gold standard to train a statistical model which can be used to rank the grammar’s output parse trees, both to improve its performance in applications and to be helpful to grammar engineers during development and debugging.

To evaluate the grammar’s suitability to support applications like grammar feedback systems for second language learners, a small extension of the grammar is also built with MAL-rules and MAL-types to enable the parsing of sentences containing grammatical errors and detecting the specific errors. The information provided by the grammar would thus allow the feedback system to identify the errors and give appropriate suggestions to the learner.

Chapter 1

Introduction

Chinese is a group of related but sometimes mutually unintelligible languages that originated in China, including Mandarin Chinese, Cantonese, Min, etc. These languages have many grammatical similarities, though their orthography and vocabulary may differ from language to language. Thus, it is advantageous to implement a Chinese grammar as computational resource covering both the common parts of the grammars and the linguistic diversity across the languages. Building such a grammar reduces the cost for resource construction by sharing the common core definitions across languages and also helps reflect the nature of the Chinese languages reliably.

1.1 Statement of the Research Topic

This thesis describes the main task of my PhD study, which is to develop a computational grammar (Zhong) for modern Mandarin Chinese in the framework of Head-driven Phrase Structure Grammar (HPSG: Pollard & Sag, 1994).

There has been various works in the HPSG community to create analyses for various phenomena in Mandarin Chinese, and a few small-scale attempts to create computational grammars for it. Details of these efforts are described in Section 2.2. However, these early works cover only some essential structures of Chinese with a minimal lexicon. The objective of my PhD study is to give a fuller description of the rich phenomena in Chinese language, and implement them in the computational grammar, Zhong, so that Zhong may have sufficient coverage

to be useful for hypothesis testing of linguistic phenomena as well as supporting applications. I consider particularly a computer-aided language learning system that can give constructive feedback on grammatical errors to language learners.

Inspired by the existing works on grammar sharing, such as the LinGO Grammar Matrix system (Bender et al., 2010), CoreGram (Müller, 2013a, 2015), CLIMB (Fokkens et al., 2012), SLaviCore (Avgustinova & Zhang, 2009) and SlaviCLIMB (Fokkens & Avgustinova, 2013) (details to be found in Section 2.3), Zhong aims to model the common parts and the linguistic diversity across the varieties of Chinese in a single hierarchy. The different Chinese grammars in Zhong share some elements, such as basic word order, and have other elements distinct, such as lexemes and specific grammar rules (e.g., classifier constructions).

Zhong is therefore structured as follows:

(1)

All grammars build upon the common constraints and inherit from shared files containing top-level descriptions for `zhong`. The differences between Mandarin and Cantonese, such as NP structures, are reflected in the `cmn` and `yue` subsets. Since Mandarin Chinese has different orthographies, simplified and traditional characters, in different areas, the Mandarin Chinese grammars are further divided into `zhs` and `zht` representing the simplified and traditional version respectively, modeling further distinction in the corresponding files.

This thesis focuses mainly on Mandarin Chinese.

The development of Zhong builds on the early work of Mandarin Grammar Online (ManGO: Yang, 2011) by the Center for the Study of Language and Information (CSLI) at Stanford University, who collaborated with Shanghai International Studies University to work on the grammar from early 2011.

Based on ManGO's earlier work, Zhong is enhanced following a data-driven approach. The lexicon is constructed by semi-automatically learning the entries from a well-annotated Chinese corpus. And the linguistic phenomena to be analyzed are also selected by observing

real sentences from various corpora, focusing on those that are more frequently encountered in the corpora.

The development cycle

The development of a computational grammar modeling a target language falls into the area of Computational Linguistics, which is an inter-disciplinary field crossing Linguistics and Computer Science. Therefore the process of grammar development can benefit from the software development cycles commonly used in software engineering, for example, the spiral model (Boehm, 1988). In a style similar to the spiral model, each targeted phenomenon or construction of Chinese is handled by iteratively going through the following cycle:

Figure 1.1: Grammar Engineering Spiral

Parse: Example sentences containing the targeted phenomenon are parsed by the grammar to identify what the grammar doesn't yet cover. This often reveals work to be done in the lexicon, grammar rules, feature types, etc. Often simple sentences are constructed to allow focusing on a particular phenomenon.

Treebank: When the grammar is mature enough, it becomes possible to treebank naturally occurring text. HPSG grammars implemented in DELPH-IN style sometimes can yield many outputs in order to accurately model valid kinds of ambiguity. A treebank is constructed by examining the parsed sentences and manually selecting the preferred interpre-

tation from the output. This allows us to validate the grammar’s behavior over real data. The treebanking tool allows us to focus on sentences where the analysis has changed while relieving us of the need to inspect those that have not.

Analyse: Syntactic analysis on the phenomenon is conducted while referring to classical Chinese grammar books as well as other relevant literature. The example sentences in the corpus are also considered during the analysis.

Model: The selected analysis will then be formally modeled in the HPSG framework. I generally follow the classical HPSG theory and well-accepted conventions in the DELPH-IN community. Adaptations to cater to Chinese-specific phenomena are provided when necessary.

Implement: The formal analysis is subsequently implemented computationally in Zhong using open-source tools like ACE ¹ and the Linguistic Knowledge Builder (LKB: Copestake, 2002) ² systems from the DELPH-IN community. Documentation is inserted into the grammar itself.

Parse: Finally the modified grammar is tested by parsing the corresponding example sentences again to make sure the implementation has handled the targeted phenomenon correctly. Regression testing is also conducted over the functional test-suite to detect unexpected interaction within the grammar so that unwanted degradation of the performance can be prevented during the development process.

In the Analyse stage we consult several Chinese reference works such as *Modern Chinese Grammar* (Wang, 1959), *Lecture Notes on Grammar* (Zhu, 1982), and *Mandarin Chinese: A Functional Reference Grammar* (Li & Thompson, 1989), as well as the more recent publication *A Reference Grammar of Chinese* (Huang & Shi, 2016).

The treebank produced during grammar engineering is also used to train a statistical ranking model which can score the trees output from the grammar so the more likely and often better trees are ranked higher. This ranking model can improve the performance of the parser

¹<http://sweaglesw.org/linguistics/ace>

²<http://wiki.delph-in.net/moin/LkbTop>

using this grammar as well as increase the efficiency of grammar engineers in checking and debugging the output of the grammar when further enhancing Zhong.

1.2 Justification

Developing computational linguistic models, a.k.a. grammar engineering, has been considered an important task in computational linguistics (Bender et al., 2008). An implemented grammar allows linguistic hypotheses to be tested against real examples from natural language data. Focused analyses of specific language phenomena are abundant in the linguistic literature. However, when implementing the analyses in a computational grammar, it's impossible to start with just one single phenomenon as the full analysis of this phenomenon depends on the availability of the implementation of other structures to make it work. In addition, the analyses of different phenomena may interact with each other. The semi-automatic nature of testing an implemented grammar against a test suite allows prompt discovery of such interactions so bugs can be caught quickly and analyses can be adjusted, if necessary.

The early efforts in building a computational Chinese HPSG grammar mentioned in the previous section, such as ManGO, provide the essential structures for modeling the language. They therefore provide a good foundation for my work, but they had low coverage and were thus not suitable for parsing sentences from real data. Significant development was necessary to bring the grammar to the level of being useful for potential applications.

I give an example sentence in (2). This sentence could not be parsed by ManGO, as its lexicon didn't include all the entries required to cover this sentence and the specific structure in this sentence (the SUO-DE construction) was not handled yet. Here, a successful analysis of the sentence requires proper modeling of two particles in Chinese, 所 *suo* and 的 *de*, especially the latter in its function of relativization. I'll show that, as a result of this research, Zhong can produce the expected parse for this sentence.

(2) 张三 喜欢 李四 所 写 的 书
zhāngsān xǐhuān lǐsì suǒ xiě de shū
Zhangsan like Lisi SUO write DE-rel book

“Zhangsan likes book(s) that Lisi wrote”

A computational grammar for Chinese with sufficient coverage, both lexically and syntactically, will benefit applications that require deep linguistic analysis to provide rich syntactic and semantic information. For example, in the area of computer-aided second language learning, it will be highly useful for a learner of Chinese to be able to obtain feedback from a system that can detect grammatical errors in his/her sentences and also give constructive suggestions of how to correct the errors. This functionality can be enabled with our grammar by incorporating specifications of common grammatical errors that beginners may make, in the form of “MAL-rule” (Schneider & McCoy, 1998; Bender et al., 2004; Suppes et al., 2014).

I provide here two examples of common grammatical errors that beginning learners of Chinese tend to make. One common error is to use NP coordinator 和 (“and”) to connect clauses, as shown in (3).

(3)
*他 不 学 中文, 和 不 学 法文。
tā bù xué zhōngwén, hé bù xué fǎwén。
He not learn Chinese, and not learn French.

“He learns neither Chinese nor French”

The other common error is the wrong position of the adverb 也 (ye) “also”. (4) shows the adverb being put wrongly before the subject of the clause, where its correct position should be after the subject and before the head verb.

(4)

*我 学 工程, 也 我 学 中文。

wǒ xué gōngchéng, yě wǒ xué zhōngwén。

I study engineering, also I study Chinese.

“I study engineering, and I also study Chinese.”

The MAL-extension of Zhong allows such ungrammatical sentences to be parsed with the errors highlighted in the MAL-types and MAL-rules triggered by the corresponding errors in the sentence.

1.3 Structure

This thesis is organised as follows. Chapter 2 reviews the related work done in the area of Chinese grammar modeling in the formal framework of HPSG. Chapter 3 describes a semi-automatic approach in expanding the lexicon to increase Zhong’s coverage. Chapter 4 highlights the major linguistic phenomena investigated during the enhancement of the grammar. Chapter 5 describes how these phenomena are modeled in the HPSG framework and implemented in Zhong. Chapter 6 describes various grammar engineering efforts, including the construction of a treebank using the grammar over a selected corpus as well as a parse ranking model trained using the treebank. Chapter 7 concludes the work with discussions for future work.

Chapter 2

Literature Review

2.1 Deep Linguistic Analysis and HPSG

As the Natural Language Processing community embraces statistical techniques, the value of deep linguistic analysis aiming at content level understanding is still well-recognized in producing high quality machine translation (Uszkoreit et al., 2000). For example, in Machine Translation, the selection of the correct translation for input words with multiple senses often depends on complex grammatical knowledge that can not be provided by local context covered by statistical phrase tables. Although deep analysis is known to lack efficiency and robustness, the improvement in performance experienced in Verbmobil, a large-scale distributed project of speech dialogue translation, has shown the potential for deep linguistic analysis to reach the efficiency required for practical applications (Kiefer et al., 2000).

To further explore the potentials of deep linguistic analysis of human language, researchers from different research centers have started world-wide collaboration in a consortium called Deep Linguistic Processing with HPSG (DELPH-IN).¹ The objective of this initiative is to combine linguistic and statistical processing methods for efficient analysis of texts and utterances to get their meanings. Originally started by DFKI in Germany and Center for the Study of Language and Information (CSLI) at Stanford University, the consortium now involves researchers from eighteen institutions worldwide.

The foundation of deep linguistic analysis is a well-developed grammatical framework that

¹<http://www.delph-in.net/>

supports formal linguistic description at morphological, syntactic, and semantic level. According to Oepen et al. (2002), Head-Driven Phrase Structure Grammar (HPSG), Lexical Functional Grammar (LFG), and Tree Adjoining Grammar (TAG) are the most widely accepted theories of grammar in Computational Linguistics. HPSG became the selected framework used in Verbobil, and later on in DELPH-IN, as HPSG's monostratal, formal representation of phonological, syntactic and semantic information in typed feature structures makes it very suitable for deep linguistic processing (Uszkoreit et al., 2000). An adequately expressive semantic representation, Minimal Recursion Semantics (Copestake et al., 2005), also exists to support a clean interface between syntax and semantics in HPSG.

According to DELPH-IN's website, broad-coverage HPSGs for English (LinGO English Resource Grammar ,ERG: Flickinger, 2000a), German Grammar (GG: Müller & Kasper, 2000; Crysmann, 2005), and Japanese (Siegel & Bender, 2002), have been developed and used in various applications, whereas grammars for a wide variety of other languages, such as Chinese, French, Korean, Modern Greek, Norwegian, Portuguese, and Spanish, are still under development.

2.2 Chinese HPSG

As the world takes on interest in the Chinese language with the rise of China, the demand for Chinese translation has been growing fast. However, it is very challenging to automatically translate between Chinese and a typologically distant language, such as English. Unlike English, Chinese has no grammatical inflections of words to indicate tense, voice, or number (Li & Thompson, 1989). The generation of such information in a target language like English is therefore difficult. The lack of syntactic markers also makes Chinese highly ambiguous and thus hard to analyze, as the functions of words are rather dependent on word order and sentence structure.

The ambiguity in Chinese has been summarized by Chang & Krulee (1991) into five types: morphological (word segmentation), lexical (Part-of-Speech determination, homonymy, and polysemy), syntactic (serial verbs, conjunction scoping, classifier ambiguity, etc.), semantic

(serial verbs and coverbs), and contextual (number/definiteness of NPs, tense/aspect/number of VPs, topic vs. subject, omitted subject, etc.). The resolution of such widely existing ambiguity requires the contextual understanding with integration of syntactic and semantic information.

Therefore, deep linguistic analysis using a Chinese grammar defined following a formal framework, such as HPSG and MRS, has high potential to improve the analysis result of the language and bring the translation quality to a new level.

The research on deep Chinese processing is not abundant. The work on Chinese HPSG is even less. There was some early work from the community of linguistics, such as an account of Chinese noun phrases by Ng (1997) from HPSG's perspective. Bender (2000) had a detailed analysis of the BA-construction in Mandarin Chinese in the framework of Lexical Functional Grammar, treating BA as a verb. Gao (2000) is a more comprehensive analysis of Chinese argument structure in the framework of HPSG. His analysis covered topics, valence alternations (including BA, ZAI, and other constructions), hierarchical argument structures, locative phrases, phrase structures, and resultative structures. Rules were also proposed to handle such structures.

More recent work on Chinese HPSG tries to extend pure linguistic analysis to the implementation of the grammar as computational resources that can be used by applications such as MT systems. A technical report from Zhang (2004) documented some early work at Saarland University building a Chinese HPSG using the Linguistic Knowledge Builder (LKB) system (Copestake, 2002),² an interactive grammar development environment for typed feature structure grammars. Müller & Lipenkova (2009) provided a detailed account of Serial Verb Constructions from the perspective of HPSG and implemented the analysis in the grammar implementation platform TRALE.³

However, systematic development of large-scale Chinese HPSG grammar, aiming at robustness and wide coverage, has only started very recently in the member institutions of DELPH-IN consortium. The Tsujii Laboratory at University of Tokyo has proposed a skeleton design of Chinese HPSG (Wang et al., 2009) and explored a semi-automatic approach to learn lexicon from an HPSG Treebank converted from Chinese Treebank 6.0 (Yu et al., 2010). At

²<http://wiki.delph-in.net/moin/LkbTop>

³<http://www.sfs.uni-tuebingen.de/hpsg/archive/projects/trale/>

the DELPH-IN 2011 Summit ⁴ (June 25-29, 2011), a special session was dedicated to HPSG for Mandarin Chinese. Two member institutions, the Language Technology Lab at DFKI and CSLI at Stanford University reported their work in this area. DFKI has been building a Mandarin Chinese Grammar (MCG) since summer 2010, whereas CSLI has started working on Mandarin Grammar Online (ManGO) since early 2011 collaborating with Shanghai International Studies University (Yang, 2011). Both are still at a very early stage of the development process, with a very limited lexicon and covering only essential structures.

2.3 Grammar Sharing

The idea of letting different grammars share a common core to capture cross-linguistic generalization has been embraced by a number of projects as a more systematic approach for grammar development. The LinGO Grammar Matrix system (Bender et al., 2010) expedites the development of complex grammars through grammar customization by providing a static core grammar that handles basic phrase types, semantic compositionality and general infrastructure. It also provides libraries for cross-linguistically variable phenomena, so that analyses of these can be dynamically generated as code based on user-configured parameters. The generated grammar is then ready to be extended manually by a linguist/grammar engineer. ManGO, the basic Grammar that Zhong stems from, was first created using the LinGO Grammar Matrix system.

CoreGram Müller (2013b) is motivated by a similar assumption that grammars sharing certain properties can be grouped into classes and thus share common files. Fokkens et al. (2012) proposes CLIMB (Comparative Libraries of Implementations with Matrix Basis), a methodology closely related to the LinGO Grammar Matrix. While still sharing implementation across different languages, the emphasis of CLIMB is facilitating the exploration and comparison of implementations of different analyses for the same phenomenon.

There's also existing work sharing a common core grammar among languages within a language family. Avgustinova & Zhang (2009) build a common Slavic core grammar (SlaviCore) shared by a closed set of languages in the Slavic language family. They further extended their

⁴<http://moin.delph-in.net/SuquamishSchedule>

work into SlaviCLIMB (Fokkens & Avgustinova, 2013), a dynamic grammar engineering component based on the CLIMB methodology, to capture language specific variations and facilitate grammar development for individual Slavic languages.

2.4 Conclusion

The above survey shows that the research work on Chinese HPSG is still at a early stage. A broad-coverage Chinese HPSG needs to be developed to enable experiments for the verification of HPSG's value in improving the analysis of Mandarin Chinese so that deep linguistic analysis can provide accurate and rich information for tasks like computer-aided language learning. Therefore, my PhD study focuses on the precise description of Chinese grammar in HPSG framework and its computational implementation in Zhong. The resulting grammar will eventually be used to build a grammar feedback system to support better learning Chinese as a second language.

Chapter 3

Lexical Acquisition

3.1 The Lexicon before Lexical Acquisition

When I started my dissertation (2012-08-08), the existing grammar already covered many major challenging phenomena in the language, with a lexicon containing representatives for various lexical categories.

The number of entries in the existing lexicon was very small with 579 entries in total. Every entry was labeled using the PINYIN of the word with STEM containing the actual Chinese character(s). The predicate of each word was, however, represented using its English equivalent as the value for feature SYNSEM.LKEYS.KEYREL.PRED. Here LKEYS is a convenience feature defined to point to the semantic information of the word. KEYREL points to its main relation (the **key** relation) and PRED is its predicate. An example entry is shown below:

(5)

```
shu4_n := n__le &  
  [ STEM < "树" >,  
    SYNSEM.LKEYS.KEYREL.PRED "_tree_n_rel" ].
```

An ambiguous word with multiple categories is represented as multiple entries, each indicating one category. For example, 帮 *bāng* is a transitive verb, taking a noun phrase (NP) as its argument, as shown in (6). It can also take a verb phrase (VP) as its second argument, as

illustrated in (7). Such different usages of the word are defined in the lexicon as separate entries with their respective categories.

(6) 他 帮 了 我

tā bāng le wǒ

He help ASP me

“He helped me”

(7) 他 帮 我 洗 车

tā bāng wǒ xǐ chē

He help me wash car

“He helps me to wash the car”

Functional words, like demonstratives, conjunctions, adverbs with grammatical functions, various particles, etc., had more or less been included in the lexicon. Such words are typically closed classes in a language. The majority of such words had already been defined with their respective lexical categories in the lexicon as the related language phenomena had been handled in the existing grammar, although the lexicon did not exhaustively listed all instances of such classes. Generation of the new instances of these categories is performed with the help of POS-tagged corpora. These lexical categories are mapped to the corresponding POS tags used in the corpora. When new instances are encountered, new lexical entries are generated automatically and added into the lexicon. Manual checking and testing then follows afterwards.

The lexical categories and the number of entries within each category are summarized in Table 3.1.

Lexical category	Explanation	No. of entries
adv_._adj_le	Adverb modifying adjs only (ji2_adv)	1
adv_._prepred_le	Adverb appearing before predicates (e.g. tai4_adv)	22
adv_._preverb_le	Adverb appearing before verbs (zhi2_adv, an4shi2_adv, zheng4zai4_adv)	3

adv_--sup_le	Adverb indicating superlatives (zui4_adv)	1
assoc-de_le	Associative marker (de_4)	1
ba_le	BA (ba_cv)	1
bei_le	BEI (bei_cv, gei3_bei)	2
c_--le	Classifier (e.g. ge4_cl)	26
comp-gap-de_le	Complement gapping relative DE (de_3)	1
conj_--le	Conjunction (he2_conj, you4_conj)	2
d_--le	Demonstrative (zhe4_d, na4_d, etc.)	5
de_adj_manner_le	To form manner adverb (de_adv1)	1
de_adj_stative_le	To form adverb with stative adj (de_adv2)	1
excl_--le	Exclamation (a_e, ne_e, ba_e)	3
guo_am_le	Perfective marker (guo4)	1
le_am_le	Perfective marker (le_am)	1
le_sf_le	Sentence final particle (le_sf)	1
loc_np*_le	Location (eg. Shang4mian4_loc, na4bian_loc)	42
loc_np_le	Locative particle (e.g. li3_loc)	8
neg_adv_le	Adverb for negation (bu4_neg, mei2_neg)	2
no-gap-de_le	De for non-gapping relative and complementing DE clauses (de_1)	1
num_--c_le	Numeral (eg. yi1_num)	36
om_--le	Ordinal marker (di4_om)	1
p_np_le	Preposition (eg. xiang4_p)	8
qpart_le	Question particle (ma_q)	1
rslt_--le	The second verb in a compound, indicating in- transitive and resultive (e.g. jin4lai2_rslt)	15
rslt_np_le	The second verb in a compound, indicating transitive and resultive (e.g. lai2_rslt)	18
subj-gap-de_le	subject gapping relative DE (de_2)	1
zai_p_le	Preposition for locative phrases (zai4_p)	1

zhe_am_le	Continuous marker (zhe_am)	1
Subtotal:		208

Table 3.1: Functional lexical categories in the original lexicon

Content words like adjectives, nouns, and verbs are generally open classes in a language. The existing lexicon contained 45 adjectives, 178 nouns, and 148 verbs. 8 sub-categories of verbs could be found in the lexicon. Table 3.2 shows the details. This part of the lexicon definitely could be extended using some automatic or semi-automatic approach.

Lexical category	Explanation	No. of entries
a_-le	Adjective	45
n_-le	Noun	178
v_-le	Intransitive verb (e.g. shui4jiao4_v)	24
v_-loc*_le	Intransitive verbs taking post-verbal locative phrases (e.g. shui4_v)	6
v_ap_le	Transitive verb taking adj (bian4_v)	1
v_np-np	Ditransitive verb (song4_v)	6
v_np-vp_le	Transitive verb with vp as arg2 (bang1_v, dang1_v)	2
v_np_le	Transitive verb	87
v_np_loc*_le	Transitive verbs taking post-verbal locative phrases (e.g. gua4_v)	4
v_vp_le	Transitive verb taking vp (e.g. xiang3_v2)	18
Total:		371

Table 3.2: Contentful lexical categories in the original lexicon

3.2 Related Work

Unknown words (words not included in a grammar’s lexicon) affect the grammar’s performance as they cause parsing failures. Lexical acquisition is therefore a well-studied problem in natural language processing and computational linguistics with many researchers working on the problem using different approaches. Baldwin (2005) has categorized the approaches of deep lexical acquisition (learning precise lexical description for a target grammar) into either

in vivo or *in vitro* methods, based on how they determine the lexical similarity of the unknown word to existing words in the lexicon.

In vivo methods rely on the grammar itself to model the lexical similarity, where as *in vitro* methods utilize language resources which are external to the target grammar. Baldwin suggests that *in vitro* methods are more applicable when the target grammar doesn't yet have enough coverage to parse a large amount of corpus data, as required by *in vivo* methods.

In vitro methods may learn lexical description by mining corpus data annotated by an external POS tagger, chunker, or parser (Korhonen, 2002; Im Walde, 2006; Baldwin & Bond, 2003). Or they may approach the problem by mapping from one lexical resource, such as a machine-readable dictionary or WordNet, onto the one with the desired format (Sanfilippo & Poznański, 1992; Daudé et al., 2000).

Inspired by the above works, the lexicon extension task is done in two ways: to generate the lexical entries either from a well-prepared annotated Chinese corpus, or from a Chinese lexical resource.

3.3 Lexicon Extension Method

We automatically extended the content word categories (adjectives, nouns, and verbs) using publicly available linguistic resources, such as a good quality Chinese corpus tagged with POS information, or a WordNet equivalent for Chinese.

3.3.1 Resources

Chinese Corpus Linguistic Data Consortium (LDC) is well-known for supplying high quality parsed corpora for various languages, including Mandarin Chinese. Its Chinese Treebank (Xue et al., 2010) provides a 100,000-word corpus of Mandarin Chinese text annotated with POS tags and syntactic bracketing. However, the granularity of the POS tag set is too coarse to provide directly useful syntactical information needed for the lexical entries. For example, the tags defined for verbs are VA, VC, VE and VV, referring to predicative adjectives, copula verbs, existential verbs, and all other verbs, respectively (Xia, 2000). Syntactic information of

the verbs, such as whether it's transitive or intransitive, or whether it expects a sentential object, can't be directly derived from the POS tags. Additionally, since a fee is required for accessing the LDC corpora, alternative resources freely distributed are preferred.

The alternative resource identified for Zhong's lexicon expansion is the Sinica Treebank Corpus Sample¹ distributed with Python Natural Language Toolkit (Bird et al., 2009). It's a sample corpus containing 9,999 parsed sentences drawn from the Academia Sinica Balanced Corpus of Modern Chinese (Hsu & Huang, 1995), built by the Language and Knowledge Processing Group in the Institute of Information Science, Academia Sinica, Taiwan.

Sinica Corpus is the first balanced modern Chinese Corpus with topics ranging from philosophy, science, and society to art, life, and literature.² Like LDC's corpora, it's also POS-tagged. The POS-tags used by the corpus are in accordance with the CKIP Tagset of 178 syntactic categories designed by the same group (Chang & Chen, 1995), including detailed sub-categories for verbs and nouns, ideal for Zhong's lexicon.

The sample corpus is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike License,³ and readily accessible with the NLTK package. This, together with the detailed syntactic sub-categories, makes the sample corpus an ideal resource for the lexicon extension task.

The only concern is that Zhong for simplified Chinese targets at mainland Mandarin Chinese whereas Sinica was created in Taiwan where traditional Chinese is used. However, a review of the sentences in the sample corpus shows that they represent typical Modern Chinese with no significant difference in vocabulary and grammar. The only major difference to be handled is its traditional Chinese (Fan Ti) character representation, which we converted to simplified Chinese (Jian Ti) characters.

Chinese-English WordNet Xu et al. (2008) constructed a bilingual Chinese-English WordNet by translating the Princeton English WordNet into Chinese by a combination of manual and automatic means. The resulting Chinese-English WordNet contains information about WordNet synset IDs, their POS, and the Chinese translations of the corresponding lexemes. This can

¹http://nltk.org/_modules/nltk/corpus/reader/sinica_treebank.html

²<http://rocling.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm>

³<http://creativecommons.org/licenses/by-nc-sa/2.5/>

be another resource for the Zhong lexicon expansion task.

A major issue found with this resource is that the Chinese glosses of the synsets may be phrases or even sentences as, when there is a lexical gap in Chinese, they are actually the translation of the English glosses for the synsets. Filtering must be performed in order to generate from glosses that are more appropriate to be used as lexical entries.

The other issue is that the POS tags provided are very coarse with only four categories - *a*, *n*, *r*, and *v*, corresponding to adjectives, nouns, adverbs, and verbs. There is no detailed differentiation among verbs, such as transitive or intransitive verbs. Such differentiation is critical to the parsing of Chinese sentences and therefore must be obtained through other means.

Wordnets contain rich semantic information (hyponymy, meronymy, etc.) of the words. However, these are not included in our lexicon since parsing requires only the syntactic information of the words.

3.3.2 Approach

Test suite creation Around 1,000 sentences were randomly selected from Sinica sample corpus and converted into Simplified Chinese. These sentences were used to create the test suite, which means they would not be used for generating the lexical entries.

Lexicon generation from Sinica The tagged words (word-tag pairs) in the remaining sentences in the corpus were extracted and converted into simplified Chinese characters. The extracted and converted words were compared with the STEMs in Zhong lexicon. Words that couldn't be found in the lexicon became the candidate unknown words to be added.

The definition of Sinica POS-tags was studied and mapped to the equivalent or most similar lexical category in Zhong. For each word in the unknown list, if its POS-tag had an equivalent mapping, a Zhong lexicon entry was generated.

Lexicon generation from Chinese-English WordNet Each Chinese gloss was scanned to filter off phrases or sentences from subsequent processing. The remaining glosses were compared with the STEMs in Zhong lexicon. Those not found in the lexicon became the candidate unknown words to be generated.

Mappings from POS-tags to Zhong lexical categories were created. For Chinese glosses marked with ‘v’, its synset ID was used to obtain the English verb synset from WordNet. The verb frames of the English synset were borrowed as indication of different verb sub-categories, under the assumption that words with similar meaning behave similarly syntactically (Fujita & Bond, 2007). Zhong lexicon entries were then generated.

Grammar testing The test suite created above were used to test Zhong’s coverage using the Linguistic Knowledge Builder (LKB) system (Copestake, 2002), a grammar and lexicon development environment for typed feature structure grammars.

Testing was done under three conditions, measuring Zhong’s coverage on the test suite using i) its original lexicon, ii) the original lexicon and the lexicon generated from the remaining Sinica sentences, iii) the original lexicon and the lexicon generated from Chinese-English Wordnet. The results from these tests were compared to check if the generated lexicon helps to improve the coverage of the grammar, and which method offers better improvement.

3.4 Results

3.4.1 Test suite creation

The length of sentences in the Sinica sample corpus ranged from 1 word to 50 words. A quick scan of the sentences revealed that the single-word sentences were mostly the result of breaking an originally long sentences into shorter segments based on in-sentence punctuations. For example, the sentence with index number 36 (index starts from 0) just contained one word “这时” (“now”). Such single-word sentences are not sentences in the strict sense.

Another discovery was that many long sentences in the corpus contained Taiwan-related vocabulary. To make the test suite more general, shorter sentences were preferred.

As a result, 1,040 sentences, with length between 2 to 10 words inclusively, were randomly selected from the corpus. A test suite was created from these sentences. To quickly create a file conforming to the LKB requirements for test suites, dummy information was used for the rest of the fields, except for “Source” and the sentence input itself.

The remaining 8,959 sentences were used to generate new lexical entries for Zhong’s lexicon.

3.4.2 Lexicon generation from Sinica

Conversion of characters from traditional Chinese to simplified Chinese was performed using a free python package called **JianFan**.⁴ It can be called from python programs while accessing Sinica corpus in NLTK to convert an input Unicode string of traditional Chinese to a Unicode string of simplified Chinese. However, some characters couldn’t be converted and they were handled afterwards by mapping from a manually-crafted list in program.

The remaining 8,959 sentences had 85,507 tagged tokens in total. The number of unique word-tag pairs was 19,268, of which 18,474 were for words unknown in the Zhong lexicon. Zhong Lexicon entries were generated from the latter group.

The Sinica sample corpus has a comprehensive list of POS-tags with detailed subcategories for adjectives, nouns, and verbs. Based on the Technical Report of Academia Sinica Balanced Corpus of Modern Chinese (Huang & Chen, 1998), the definitions of the Sinica POS tags were compared with the definitions of lexical categories in Zhong. They were mapped to their closest Zhong equivalents if available, as summarized in Table 3.3.

Sinica tag	Explanation	Zhong category	Comments
A	Non-predicative adj	a_-le	
Caa, Cab, Cba, Cbb, Cbc	Conjunction	conj_-le	Further differentiation needed
Dbb	Adverb	adv_-prepred_le	
Daa, Dab, Dbaa, Dbab, Dc, Dd, Dg, Dh, Dj	Adverb	adv_-preverb_le	
Dfa	Pre-verbal Adverb of degree	adv_-preverb_le	
Dfb	Post-verbal Adverb of degree	-	No equivalent
Dk	Sentential Adverb	adv_-preverb_le	No equivalent, tentative
DM	Quantitative adverb (e.g. 五十块)	-	No equivalent
I	Interjection	-	No equivalent

⁴<http://code.google.com/p/python-jianfan/>

Naa, Nab, Nac, Nad, Naea, Naeb	Common noun	n_ _le	
Nba, Nbc	Proper Noun	n_ _le	
Nca, Ncb, Ncc, Nce	Place Noun	n_ _le	
Ndaa, Ndab, Ndc, Ndd	Time Noun	n_ _le	
Nhaa, Nhab, Nhac, Nhb, Nhc	Pronoun	n_ _le	
Ncda	Localizer (single character)	loc_np_le	
Ncdb	Localizer	loc_np*_le	
Nep, Neqa, Nes	Determinative	d_ _le	Further differentiation needed
Neu	Numeral	num_ _c_le	
Nfa Nfi	Measure, classifier	c_ _le	
Ng	Postposition	-	No equivalent
P	Preposition	p_np_le	Need further investigation
Ta, Tb, Tc, Td	Sentence end particle	excl_ _le	
VA11, 12, 13, VA3, VA4	Active Intransitive Verb	v_ _le	
VAC VA2	Active Causative Verb	v_ _le	
VB11, 12, VB2	Active Pseudo-transitive Verb	v_ _le	
VC2, VC31, 32, 33	Active Transitive Verb	v_np_le	
VCL VC1	Active Verb with a Locative Object	v_np_le	
VD1, VD2	Ditransitive Verb	v_np-np_le	
VE11, VE12, VE2	Active Verb with a Sentential Object	v_np_le	No equivalent in Zhong, tentative to v_np_le to capture its transitive nature
VF1, VF2	Active Verb with a Verbal Object	v_vp_le	
VG1, VG2	Classifying verb (e.g. 定为)	-	No equivalent, need further investigation
VH11, 12, 13, 14, 15, 17, VH21	Stative Intransitive Verb	a_ _le (for VH15, 17, 21 only)	Arguable for this group. Many are ambiguous words. Some can be v_ _le and some are a_ _le. Needs further study.
VHC VH16, VH22	Stative Causative Verb	a_ _le	May be ambiguous
VII, 2, 3	Stative Pseudo-transitive Verb	v_ _le	
VJ1, 2, 3	Stative Transitive Verb	v_np_le	

VK1, 2	Stative Verb with a Sentential Object	v_np_le	No equivalent in Zhong, tentative to v_np_le to capture its transitive nature
VL1, 2, 3, 4	Stative Verb with a Verbal Object	v_vp_le	

Table 3.3: Mapping from Sinica POS-tags to Zhong lexical categories

Using the POS-tag mapping scheme above, new entries for the Zhong lexicon were generated for words with matching tags. Table 3.4 gives the number of entries created and the remaining word-tag pairs. The generated lexicon is thereafter referred to as the Sinica lexicon.

	Total	%
Entries generated	14,664	79.38
Remaining word-tag pairs	3,810	20.62
Total	18,474	100.00

Table 3.4: Number of entries generated and remaining word-tag pairs

The label of a generated entry was the word in Chinese character followed by its category information. For ease of future debugging, its Sinica POS tag information and the frequency of this word-tag pair in the remaining 8959 sentences were included as comment. An example entry is shown below:

(8)

```
恰好_adv := adv_-_preverb_le & ;;;ASpos: Daa freq: 1
[ STEM < "恰好" >,
  SYNSEM.LKEYS.KEYREL.PRED "_恰好_a_rel" ].
```

The number of entries created for each category is listed in Table 3.5. The largest category is noun (n_-le), which accounts for 61.44% of all generated entries, followed by transitive verbs (v_np_le) which is 18.34% of all entries.

Category	No.	%	Example entry for the most frequent word
----------	-----	---	--

a- _{-le}	437	2.98	<p>一般_a := a-_{-le} & ;;;ASpos: A freq: 15</p> <p>[STEM < "一般" >, SYNSEM.LKEYS.KEYREL.PRED "_一般_{a_rel}"].</p>
adv- _prepred _{le}	78	0.53	<p>就_{adv} := adv-_prepred_{le} & ;;;ASpos: Dbb freq: 65</p> <p>[STEM < "就" >, SYNSEM.LKEYS.KEYREL.PRED "_就_{a_rel}"].</p>
adv- _preverb _{le}	549	3.74	<p>才_{adv} := adv-_preverb_{le} & ;;;ASpos: Daa freq: 38</p> <p>[STEM < "才" >, SYNSEM.LKEYS.KEYREL.PRED "_才_{a_rel}"].</p>
c- _{-le}	129	0.88	<p>艘_c := c-_{-le} & ;;;ASpos: Nfa freq: 1</p> <p>[STEM < "艘" >].</p>
conj- _{-le}	10	0.07	<p>并_{conj} := conj-_le & ;;;ASpos: Cbcb freq: 47</p> <p>[STEM < "并" >, SYNSEM.LKEYS.KEYREL.PRED "_并_{c_rel}"].</p>
d- _{-le}	150	1.02	<p>其_d := d-_le & ;;;ASpos: Nep freq: 99</p> <p>[STEM < "其" >, SYNSEM.LKEYS.KEYREL.PRED "_其_{q_rel}"].</p>
excl- _{-le}	22	0.15	<p>呀_{excl} := excl-_le & ;;;ASpos: Tc freq: 25</p> <p>[STEM < "呀" >].</p>
loc _{np*} _{le}	51	0.35	<p>西方_{loc} := loc_{np*}_{le} & ;;;ASpos: Ncdb freq: 14</p> <p>[STEM < "西方" >, SYNSEM.LKEYS.KEYREL.PRED "_西方_{loc_rel}"].</p>

loc_np_le	21	0.14	<div>东_loc := loc_np_le & ;;;ASpos: Ncda freq: 10</div> <div>[STEM < "东" > ,</div> <div>SYNSEM.LKEYS.KEYREL.PRED "_东_loc_rel"].</div>
n_-le	9,010	61.44	<div>运动_n := n_-le & ;;;ASpos: Nad freq: 24</div> <div>[STEM < "运动" > ,</div> <div>SYNSEM.LKEYS.KEYREL.PRED "_运动_n_rel"].</div>
num_- _c_le	135	0.92	<div>第一_num := num_-c_le & ;;;ASpos: Neu freq: 29</div> <div>[STEM < "第一" > ,</div> <div>SYNSEM.LKEYS.KEYREL.PRED "_第一_num_rel"].</div>
p_np_le	108	0.74	<div>将_p := p_np_le & ;;;ASpos: P07 freq: 88</div> <div>[STEM < "将" > ,</div> <div>SYNSEM.LKEYS.KEYREL.PRED "_将_p_rel"].</div>
v_-le	1050	7.16	<div>飞行_v := v_-le & ;;;ASpos: VA11 freq: 6</div> <div>[STEM < "飞行" > ,</div> <div>SYNSEM.LKEYS.KEYREL.PRED "_飞行_v_rel"].</div>
v_np- np_le	94	0.64	<div>送给_v := v_np-np_le & ;;;ASpos: VD1 freq: 21</div> <div>[STEM < "送给" > ,</div> <div>SYNSEM.LKEYS.KEYREL.PRED "_送给_v_rel"].</div>
v_np_le	2,690	18.34	<div>讲_v := v_np_le & ;;;ASpos: VE2 freq: 15</div> <div>[STEM < "讲" > ,</div> <div>SYNSEM.LKEYS.KEYREL.PRED "_讲_v_rel"].</div>
v_vp_le	130	0.89	<div>申请_v := v_vp_le & ;;;ASpos: VF1 freq: 13</div> <div>[STEM < "申请" > ,</div> <div>SYNSEM.LKEYS.KEYREL.PRED "_申请_v_rel"].</div>

Total	14,664	100.00
-------	--------	--------

Table 3.5: Number of entries generated for 16 Zhong lexical categories

3.4.3 Lexicon generation from Chinese-English WordNet

The Chinese-English WordNet data file (wn-mcn.tab) contains 157,112 entries in total. Each entry gives a synset ID, POS, and Chinese gloss, as illustrated below:

(9) 00001740-a 可以

SynsetID-POS ChineseGloss

Distribution of the four general POS is summarized in Table 3.6.

POS	Count
Adjectives (a)	28,510
Nouns (n)	100,064
Adverbs (r)	5,851
Verbs (v)	22,687
Total	157,112

Table 3.6: Distribution of four POS in the Chinese-English WordNet

Long glosses Since Chinese-English WordNet was constructed by translating from English WordNet, many entries in the data file contain phrases or sentences. For example, entries like “01128266-n 治安维持会的政策” and “01086549-v 在赛马或赛狗中获得第三名或更好的名次” are actually phrases describing the meaning of the original English terms. Such entries can’t be directly used to generate lexical entries as such phrases or sentences are not proper lexical terms. They should be filtered off.

Entries like “01049606-v 唱; 奏” and “00472426-v 打破; 击倒” have more than one translation in the entry, separated by a semicolon. Entries like “00473322-v 删除: 把多余或不想要的东西去掉” and “01287797-v 扎紧: 烧之前捆住或串起的翅膀或腿” seem to be result of translating a dictionary definition. For such entries, we can extract the first translation or the term before definition using punctuations as the delimiter. For the above examples, the extracted terms are “唱”, “打破”, “删除”, and “扎紧” respectively.

Overall, longer Chinese glosses tend to be such phrases. Based on the distribution of glosses over length (in number of characters), shown in Table 3.7, I set a threshold of 5 characters: lemmas longer than this were discarded.

Chinese gloss length (no. of characters)	Count
2	3,255
3	60,771
4	35,251
5	29,550
6	13,074
7	6,384
8	3,550
9	1,841
10	1,180
11-49	2,256
Total	157,112

Table 3.7: Distribution of Chinese glosses over gloss length

Unknown word identification Some of the remaining entries contained the same Chinese gloss but have different synset ID and POS, e.g. “00001740-a 可以” and “00510348-a 可以”. A mapping was created from a gloss to all its associated synset ID and POS pairs. This resulted in 76,262 unique glosses. The unique glosses were then compared with the STEMs in the Zhong lexicon. 384 glosses were found to be already known.

POS-tag mapping and Zhong lexicon generation Chinese terms marked with ‘v’ needed more information to derive their sub categories. The synset IDs were used to obtain the English verb synset from WordNet through NLTK (Bird et al., 2009). ⁵

The verb frames of the English synsets were borrowed to indicate different verb sub-categories for Chinese terms. English WordNet has 35 verb frames. For each verb frame, the most similar Zhong lexical category was identified if possible. The verb frames, their expression string (according to NLTK source code), and the mapped Zhong categories, are summarized in Table 3.8. The mappings for transitive and intransitive verbs are likely to be correct. However, the mappings for other verb frames only serve as a rough indication of what

⁵<http://nltk.googlecode.com/svn/trunk/doc/howto/wordnet.html>

verb categories a Chinese gloss is likely to have, as translation from English to Chinese often requires a different way of expression.

WordNet Verb Frame	Expression String	Zhong Lexical category
1	Something V	v _- _le
2	Somebody V	v _- _le
3	It is Ving	v _- _le
4	Something is Ving PP	v _- _le
5	Something V something Adjective/Noun	–
6	Something V Adjective/Noun	–
7	Somebody V Adjective	–
8	Somebody V something	v _np _le
9	Somebody V somebody	v _np _le
10	Something V somebody	v _np _le
11	Something V something	v _np _le
12	Something V to somebody	–
13	Somebody V on something	v _- _le
14	Somebody V somebody something	v _np-np _le
15	Somebody V something to somebody	v _np _le
16	Somebody V something from somebody	v _np _le
17	Somebody V somebody with something	v _np _le
18	Somebody V somebody of something	v _np _le

19	Somebody V something on somebody	v_np_le
20	Somebody V somebody PP	v_np_le
21	Somebody V something PP	v_np_le
22	Somebody V PP	v_-_le
23	Somebody's (body part) V	v_-_le
24	Somebody V somebody to INFINITIVE	v_np-vp_le
25	Somebody V somebody IN-FINITIVE	v_np-vp_le
26	Somebody V that CLAUSE	–
27	Somebody V to somebody	–
28	Somebody V to INFINITIVE	v_vp_le
29	Somebody V whether IN-FINITIVE	–
30	Somebody V somebody into V-ing something	v_np-vp_le
31	Somebody V something with something	v_np_le
32	Somebody V INFINITIVE	v_vp_le
33	Somebody V VERB-ing	v_vp_le
34	It V that CLAUSE	–
35	Something V INFINITIVE	v_vp_le

Table 3.8: Mapping from WordNet verb frames to Zhong lexical categories.

Adjectives (a) were mapped to “a_-le” whereas nouns (n) were mapped to “n_-le”. Adverbs (r) were by default mapped to “adv_-preverb_le”, since in the generated Sinica lexicon,

this is a much larger category.

Zhong lexicon entries were then generated based on the mappings. The format of a generated entry is similar to the earlier example except the comment. For each unique word, the original synset-pos pairs associated with the word were kept in the comment, followed by entries generated for each POS (a verb may have multiple entries generated based on the verb frames). An example entry is shown below.

(10)

```
;;;wn-mcn: 邀请, wordnet synsets: 01469770-v 01358534-a
03134550-a 00753881-v 02384686-v 00793580-v 07186148-n
02384940-v 01063695-v
邀请_v := v_np_le &
    [ STEM < "邀请" >,
      SYNSEM.LKEYS.KEYREL.PRED "_邀请_v_rel" ].
邀请_a := a_-_le &
    [ STEM < "邀请" >,
      SYNSEM.LKEYS.KEYREL.PRED "_邀请_a_rel" ].
邀请_n := n_-_le &
    [ STEM < "邀请" >,
      SYNSEM.LKEYS.KEYREL.PRED "_邀请_n_rel" ].
邀请_v := v_np-vp_le &
    [ STEM < "邀请" >,
      SYNSEM.LKEYS.KEYREL.PRED "_邀请_v_rel" ].
```

The example clearly illustrates that the POS information in Chinese-English WordNet is heavily influenced by the English original it was translated from. “邀请” (yāoqǐng ”invite”) in Chinese is either a noun or verb and it’s never an adjective. The adjective synsets for this word come from the English adjectives which were translated into “邀请”.

In total 90,360 entries were generated with details for each category summarized in Table 3.9. Nouns (n_-_le) remain the largest category, followed by adjectives (a_-_le).

Category	No.	%	Example entry for the most frequent word
a_-le	16,036	17.75	冷淡_a := a_-le & [STEM < "冷淡" >, SYNSEM.LKEYS.KEYREL.PRED "_冷淡_a_rel"].
adv_- _preverb_le	3,566	3.95	急速_adv := adv_-preverb_le & [STEM < "急速" >, SYNSEM.LKEYS.KEYREL.PRED "_急速_a_rel"].
n_-le	58,060	64.26	阻碍_n := n_-le & [STEM < "阻碍" >, SYNSEM.LKEYS.KEYREL.PRED "_阻碍_n_rel"].
v_-le	4,597	5.09	深思_v := v_-le & [STEM < "深思" >, SYNSEM.LKEYS.KEYREL.PRED "_深思_v_rel"].
v_np- np_le	152	0.17	叫_v := v_np-np_le & [STEM < "叫" >, SYNSEM.LKEYS.KEYREL.PRED "_叫_v_rel"].
v_np- vp_le	154	0.17	叫_v := v_np-vp_le & [STEM < "叫" >, SYNSEM.LKEYS.KEYREL.PRED "_叫_v_rel"].
v_np_le	7,577	8.39	进入_v := v_np_le & [STEM < "进入" >, SYNSEM.LKEYS.KEYREL.PRED "_进入_v_rel"].
v_vp_le	206	0.23	保持_v := v_vp_le & [STEM < "保持" >, SYNSEM.LKEYS.KEYREL.PRED "_保持_v_rel"].

Total	90,348	100.00
-------	--------	--------

Table 3.9: Summary of lexical categories generated from Chinese-English WordNet

The generated entries contained many wrong entries. For example the original English gloss for “打听” (dǎtīng ”inquire about”) may be a ditransitive verb. However, in Chinese, “打听” can’t take two objects directly after the verb. One of them has to be introduced by a preposition and appear before the verb. This further suggests that the POS information derived from Chinese-English WordNet may not be accurate enough for building Zhong lexicon. Since I started my thesis, a new more accurate wordnet has been constructed: the Chinese Open Wordnet (Wang & Bond, 2013). In future work I would like to redo the experiment using this resource.

3.5 Grammar Testing

Testing with the original lexicon Zhong with the original lexicon was run on the test suite, finding a parse for only 6 sentences in the test suite. This leads to the initial coverage of 0.6% (Table 3.10).

	Total Items	Positive Items	Word String	Lexical Items	distinct analyses	Total results	Overall Coverage (%)
i-length in [10 .. 15]	35	35	10.00	0.00	0.00	0	0.0
i-length in [5 .. 10]	737	737	6.65	8.37	13.00	3	0.4
i-length in [0 .. 5]	268	268	3.25	3.69	1.00	3	1.1
Total	1040	1040	5.88	5.25	7.00	6	0.6

Table 3.10: Coverage test result with the original lexicon

Testing with the Sinica lexicon added After the entries from the Sinica lexicon are added into the lexicon, the grammar’s coverage of the test suite was improved to 21.2%, as shown in Table 3.11.

From the 819 sentences that fail to be parsed, 368 of them, or 44.93% contain unknown words. The rest should be further investigated as they fail to be parsed by the grammar although

they contain no unknown words: instead the lexical entries must be lacking some necessary information, or there is some un-analyzed phenomenon.

	Total Items	Positive Items	Word String	Lexical Items	distinct analyses	Total results	Overall Coverage (%)
i-length in [10 .. 15]	35	35	10.00	14.75	23.50	4	11.4
i-length in [5 .. 10]	737	737	6.65	9.52	44.17	149	20.2
i-length in [0 .. 5]	268	268	3.25	4.34	4.53	68	25.4
Total	1040	1040	5.88	8.11	31.60	221	21.2

Table 3.11: Coverage test result using the lexicon extended with Sinica lexicon

Testing with the WordNet lexicon added The generated WordNet lexicon was very large with more than 90,000 entries. However, since the lexicon contains known problems tracing back to the translated English synsets, and there are no new entries for functional words, the improvement of coverage was expected to be much lower than that brought by Sinica lexicon: more entries are not useful if they are incorrect. The testing indeed reveals a coverage at 10.1% (Table 3.12), much lower than that of Sinica lexicon .

743, or 79.47%, of 935 unparsed sentences contain unknown words.

	Total Items	Positive Items	Word String	Lexical Items	distinct analyses	Total results	Overall Coverage (%)
i-length in [10 .. 15]	35	35	10.00	15.75	343.00	1	2.9
i-length in [5 .. 10]	737	737	6.65	10.55	72.09	64	8.7
i-length in [0 .. 5]	268	268	3.25	5.61	4.00	40	14.9
Total	1040	1040	5.88	8.76	48.73	105	10.1

Table 3.12: Coverage test result using the lexicon extended with WordNet lexicon

3.6 Conclusion

The lexicon extension experiment has proved that we can automatically extract lexical entries from linguistic resources like POS-tagged corpora and wordnets for Chinese. The generated entries can be incorporated into Zhong lexicon to improve the grammar's coverage.

The test results prove that a good quality POS-tagged corpus like Sinica sample gives better coverage improvement than a Chinese-English WordNet translated from the English version. The latter, containing useful information for content words, can be used as a resource complementary to a tagged corpus.

The future work is to further extend the Zhong lexicon by working on the Sinica main corpus or LDC Chinese Treebanks, if access can be obtained. The generated entries for functional words should be manually reviewed to verify that the lexical categories are correctly assigned. They should also be tested with sentences containing the related language phenomena to refine the grammar's rules.

The sentences in the test suite also need to be checked as they seem to be clauses derived from a long sentence by segmenting it on punctuations. Quite a number of them are therefore not complete sentences. Due to this reason, this test suite is not used for testing the grammar in terms of the syntactical coverage of the phenomena described in the next chapter. However, in future work, such clauses can be joined back into complete sentences for testing complex sentences with conjunctions.

It's also observed that in sentences that fail to be parsed by the grammar with the extended lexicon, there are a high percentage of sentences with unknown words. This demonstrates that even after lexical extension, when the grammar is used to parse sentences from real world data, it's very likely to encounter unknown words. Therefore some module handling such unknown words with heuristics is definitely necessary to make the grammar more robust, as discussed in Chapter 6.

Chapter 4

Language Phenomena of Concern

4.1 Selection of Phenomena

This thesis doesn't attempt to cover every phenomenon existing in Mandarin Chinese. The study is scoped by identifying from the corpus what are the common structures that have not yet been covered by the grammar.

As mentioned in 1.2, we plan to apply Zhong in computer-aided language learning, to build an application that can check the well-formedness of sentences created by students who are learning Chinese as a second language. A Chinese corpus called CMNEDU has thus been developed as part of the *Syntactic Well-Formedness Diagnosis and Error-Based Coaching in Computer Assisted Language Learning using Machine Translation Technology* project, PI: Francis Bond, MOE TRF. This corpus contains 798 sentences collected from Chinese textbooks used in college classes teaching Chinese as a second language. These sentences represent the kind of grammatical sentences that the students should grasp at their level of learning. This corpus is the main source used to scope the focus of the study in this thesis. Other well-known corpora, such as the Penn Treebank Chinese corpus, are also considered.

One major observation from the corpora is the frequent usage of 的 *de*, a very important structural particle used in Mandarin Chinese, similar to *of* in English. In the Penn Treebank Chinese corpus of 2,448 news articles (Xue et al., 2005), out of 51,447 sentences in total, there are 64,402 occurrences of DE in 3,2452 sentences, i.e., 63.08% of the sentences contain DE

with an average of 1.98 occurrences/sentence. Thus a good analysis of DE is critical for a grammar of Chinese.

This chapter gives an account of some of the common usages of DE, as well as the other two phenomena frequently observed in the CMNEDU corpus, reduplication of adjectives and verbs, and interrogatives.

4.2 DE Constructions Modifying Nouns

DE is usually attached to another predicative phrase to form a so-called “DE-construction”. A DE construction (XP + DE) can be used as a modifier in front of a noun phrase or be nominalized and used directly as an NP.

When DE is attached to the end of an NP, the construction is an associative phrase. When DE is attached to the end of a clause or an adjective phrase, it forms a nominalization structure. According to Li & Thompson (1989), a nominalization is a relative clause if the head noun being modified refers to an unspecified participant in the situation described by the nominalization. We make a further distinction between two scenarios, one where the head noun refers to an unspecified argument (subject or object) in the clause, as in example (11), and one where the head noun is not such an argument in the clause, as shown in example (12).

- (11) 张三 喜欢 写 书 的 人
zhāngsān xǐhuān xiě shù de rén
Zhangsan like write book DE-rel person
“Zhangsan likes people who write book(s)”

- (12) 张三 喜欢 李四 写 书 的 地方
zhāngsān xǐhuān lǐsì xiě shū de dìfāng
Zhangsan like Lisi write book DE-asc place
“Zhangsan like the place where Lisi wrote books”

We call the former **relativizing** DE constructions, and the latter **associative** DE constructions (following the terminology used in Li & Thompson, 1989). (13) illustrates our subcategorization of DE constructions.

4.2.1 Associative DE

Associative DE with NP Associative DE plays the role of connecting two NPs together to form a larger noun phrase, with the NP preceding DE modifying the NP following DE:

- (14) 我 的 朋友 哭 了
 wǒ de péngyou kū le
 I DE-asc friend cry ASP
 “My friend cried”

- (15) 他 知道 学校 的 地址
 tā zhīdào xuéxiào de dìzhǐ
 He know school DE-asc address
 “He knows the address of the school”

4.2.2 Relativizing DE

A relativizing DE construction consists of a clause headed by a verb or an adjective, followed by the relativizing marker DE. The relative clause has an argument, either a subject or an object,

missing. The DE construction modifies a head noun which then fills the semantic role as the missing argument in the clause. This leads to two possible scenarios - **gapped subject**, and **gapped object**.

Gapped Subject The head noun is the omitted subject of the verb in the relative clause, as illustrated in example (11).

Gapped Object The head noun is the omitted object of the transitive verb in the relative clause, for example:

- (16) 张三 喜欢 李四 写 的 书
zhāngsān xǐhuān lǐsì xiě de shù
Zhangsan like Lisi write DE-rel book
“Zhangsan likes book(s) that Lisi wrote”

If the head noun is preceded by a determiner and a classifier, the DE construction may appear either before or after the determiner+classifier, as illustrated in examples (17) and (18). This applies to DE constructions with gapped object as well as those with gapped subject.

- (17) 张三 喜欢 李四 写 的 这 本 书
zhāngsān xǐhuān lǐsì xiě de zhè běn shù
Zhangsan like Lisi write DE-rel this piece book
“Zhangsan likes this book that Lisi wrote”

- (18) 张三 喜欢 这 本 李四 写 的 书
zhāngsān xǐhuān zhè běn lǐsì xiě de shù
Zhangsan like this piece Lisi write DE-rel book
“Zhangsan likes this book that Lisi wrote”

4.3 Nominalizing DE Constructions

A nominalizing DE construction is quite similar to the DE constructions described in the previous section, except that it doesn't modify any head noun, or we can say the noun is omitted.

For example:

(19) 我 喜欢 很 红 的

wǒ xǐhuān hěn hóng de

I like very red DE-nom

“I like the very red (thing)”

(20) 我 喜欢 他 写 的

wǒ xǐhuān tā xiě de

I like he write DE-nom

“I like what he wrote (the thing that he wrote)”

(21) 我 喜欢 吃 的

wǒ xǐhuān chī de

I like eat DE-nom

“I like food (what can be eaten)”

(22) 我 的 哭 了

wǒ de kū le

I DE-nom cry ASP

“Mine cried”

4.4 SUO-DE structure

In Mandarin Chinese, sometimes a particle 所 *sǒ* can be found in a relative clause before its head verb, as shown in (2), which is repeated below. The role of this particle here is to indicate that the noun gapped in the relative clause must be the object of the verb. An earlier analysis was first presented in Fan et al. (2015). If the relative clause has the subject present, as in (2), its meaning is the same as the sentence without SUO (16). But for ambiguous sentences where both the subject and the object are missing in the relative clause, the existence of SUO (23) constrains the interpretation of the sentence to that of the gapped object.

- (2) 张三 喜欢 李四 所 写 的 书
zhāngsān xǐhuān lǐsì suǒ xiě de shù
Zhangsan like Lisi SUO write DE-rel book
“Zhangsan likes book(s) that Lisi wrote”

- (23) 张三 喜欢 所 写 的 书
zhāngsān xǐhuān suǒ xiě de shù
Zhangsan like SUO write DE-rel book
“Zhangsan likes book(s) that were written”

Lǚ (1999) describes 所 *sǒ* as a particle used before a transitive verb to nominalize the structure “SUO+V” into a noun phrase. According to Lu & Ma (1985), in modern Chinese, SUO is used most commonly in the structure “(NP₁)+SUO+V+DE”, either to modify a noun following it (NP₂) or to act as a noun phrase itself. One of such usages, “NP₁+SUO+V+DE+NP₂”, is shown in example (24).

- (24) 我 看 了 他 所 写 的 书
wǒ kàn le tā suǒ xiě de shū
I read ASP he SUO write DE book
“I read the book he wrote”

SUO may appear in structures with small variations. These variations are listed below in (25a-d). The last variation (25e) is used directly as an noun phrase in formal text.

- (25) a. “ $NP_1 + SUO + V + DE + NP_2$ ”
 b. “ $SUO + V + DE + NP_2$ ”
 c. “ $NP_1 + SUO + V + DE$ ” as NP
 d. “ $SUO + V + DE$ ” as NP
 e. “ $SUO + V$ ” as NP

Example(26), taken from Lǚ (1999), shows “ $NP_1 + SUO + V + DE$ ” as NP, serving as the subject of the sentence.

- (26) 他 所 说 的 未 必 确 实
 tā suǒ shuō de wèibì quèshí
 He SUO say DE not-really true
 “What he said might not be true”

“ $SUO + V + DE$ ” plays similar roles in sentences, as exemplified in (27) and (28), also from Lǚ (1999).

- (27) 所 产生 的 结果
 suǒ chǎnshēng de jiēguǒ
 SUO produce DE result
 “the result produced”

- (28) 所 用 的 还 是 老 方 法
 suǒ yòng de hái shì lǎo fāngfǎ
 SUO use DE still be old method
 “What’s adopted is still the old method”

Some earlier views, as represented by Lǚ (1999), regard SUO as a particle to nominalize “SUO+V” into NP. Lu & Ma (1985) recognizes SUO as a functional word forming SUO construction “(NP₁+) SUO+V+DE”, which is overall an NP. Just like other DE constructions, SUO construction can play the role of adjunct, subject, and object in sentences. SUO has also been analysed as a resumptive pronominal clitic, heading a nominal projection, and it undergoes raising and cliticization to its surface position (Ting, 2003).

However, Deng (2009) notes that although SUO has nominalization role for “SUO+V” in ancient Chinese, the equivalent structure in modern Chinese has evolved into “SUO+V+DE”. “SUO+V” in modern Chinese can’t be the subject or object of a sentence; “SUO+V” as modifier of N must be used with a DE (except some special cases where the V is monosyllabic); and “NP₁+SUO+V” can’t be used alone without DE. Based on these evidences, he argues that the nominalization role in structure “(NP₁+) SUO+V+DE” is actually played by the particle DE, not SUO.

If SUO is not nominalizing the above structure, the question naturally follows is what syntactic or semantic function SUO plays there.

In Deng (2009), SUO is redefined to be an additional marker attached to the verb to reduce its eligibility to be a typical predicate. The resulting structure “SUO+V” still preserves some of the properties of verbs, such as taking prepositional phrase and adverbs as adjuncts. However, it can’t take direct objects, aspect markers, or complements.

The semantic function of SUO is explained in Lu & Ma (1985). In observing the differences between sentences with and without SUO, he points out that for “(SUO+) V+DE”, the distinction lies in the scope of its referent. “V+DE” may refer to the agent, patient, or instrument of the action verb, whereas “SUO+V+DE” can only refer to its patient.

We take the view of Deng (2009) that in structures where both SUO and DE appear (25a-d), DE plays the key role of nominalizing the phrase “(NP₁+)SUO+V+DE”, so that it can either be a noun phrase itself, or be a prenominal adjunct (relative clause) to NP₂. The role of SUO in the construction is to indicate that the missing argument of the verb is its patient or direct object.

4.5 SHI-DE Constructions

DE construction or nominalization can be used as object in a sentence headed by the copula verb SHI, leading to a special sentence structure typically called emphatic SHI-DE construction. An example is given in (29).

- (29) 他是昨天来的
tā shì zuótiān lái de
he is yesterday come DE

“He came yesterday/ It was yesterday that he came”

Such constructions should be differentiated from similar structures which also contain a copular SHI and a DE phrase, such as the sentences shown in (30), (31), and (32). The role of DE in these examples is a nominalizer, as described in Section 4.3 above.

- (30) 那是我的
nà shì wǒ de
that is I DE-asc

“That is mine”

- (31) 苹果是吃的
píngguǒ shì chī de
apple is eat DE-nom

“Apples are food”

- (32) 苹果是甜的
píngguǒ shì tián de
apple is sweet DE

“Apples are sweet”

According to Li & Thompson (1989), the emphatic SHI-DE constructions emphasize the circumstances of the action, suggesting either affirming or denying of some supposition (as part of discourse). They identified a few features to set the emphasizing ones apart from other similar structures:

1. The subject of the sentence is the same as the missing subject of the construction.
2. There's an additional adverb or adverbial phrase or auxiliary verb describing the circumstances of the event in the construction.
3. SHI can be omitted. So an alternative form for (29) as in (33) still stands. In this form, DE acts as a sentence final marker (see 4.6).
4. When a direct object exists in the nominalization (34), "shi ... V O de", an alternative form (35), "shi ... V de O", is also acceptable.

(33) 他 昨天 来 的
tā zuótiān lái de
he yesterday come DE

"He came yesterday / It was yesterday that he came"

(34) 他 是 昨天 完成 任务 的
tā shì zuótiān wánchéng rènwù de
he is yesterday finish task DE

"He finished the task yesterday / It was yesterday that he finished the task"

(35) 他 是 昨天 完成 的 任务
tā shì zuótiān wánchéng de rènwù
he is yesterday finish DE task

"He finished the task yesterday / It was yesterday that he finished the task"

The alternative form, "... V *de* O", also introduces a realis mood. This is more obvious in the following examples, in which (36) is present tense, and (37) is past tense.

- (36) 他 在 北京 学 中文
tā zài běijīng xué zhōngwén
he in Beijing learn Chinese
"He learns Chinese in Beijing"

- (37) 他 在 北京 学 的 中文
tā zài běijīng xué de zhōngwén
he in Beijing learn DE Chinese
"He learned Chinese in Beijing"

The emphatic SHI-DE construction like (29) is also referred to as *shi...de pattern proper* by Paul & Whitman (2008), recognizing that it's a kind of cleft construction putting focus on the element immediately following SHI, while Paris (1998) call them SHI-DE clefts where the bound constituent could be NPs, adverbials, PPs, etc.

From information structure perspective, Huang & Shi (2016) describes this construction as a type of focus construction with the bound (emphasized) constituent carrying new and contrastive information.

The constituent being emphasized in SHI-DE constructions can be adverbials, prepositional phrases, subjects, modal verbs, participle clauses, etc., of the verb that appear between SHI and DE. The example in (29) emphasizes a temporal adverbial. The following examples illustrate three alternative ways of emphasizing the subject of the embedded verb. (39) reflects the alternative position of the direct object of the verb embedded in DeP. (40) illustrates the effect of topicalization of the object combined with the emphasis of the verb's subject.

(38) 是 他 完成 任务 的
shì tā wánchéng rènwù de
is he finish task DE
“It was him who finished the task”

(39) 是 他 完成 的 任务
shì tā wánchéng de rènwù
is he finish DE task
“It was him who finished the task”

(40) 任务 是 他 完成 的
rènwù shì tā wánchéng de
task is he finish DE
“It was him who finished the task”

Example for emphasizing scopal adverbs:

(41) 他 是 绝对 哭 了 的
tā shì juéduì kū le de
he is definitely cry ASP DE
“He DEFINITELY cried”

Example for emphasizing modal verb:

(42) 他 是 会 完成 任务 的
tā shì huì wánchéng rènwù de
he is will finish task DE
“He WILL finish the task”

Emphasizing PP:

- (43) 他 是 在 上 海 完 成 任 务 的
 tā shì zài shànghǎi wánchéng rènwù de
 he is PREP Shanghai finish DE task

“It was in Shanghai that he finished the task”

Emphasizing participial clause:

- (44) 他 是 笑 着 完 成 任 务 的
 tā shì xiào zhe wánchéng rènwù de
 he is PREP Shanghai finish DE task

“He finished the task laughing”

Zhong’s analysis of the emphatic SHI-DE is given in Section 5.5.

4.6 DE as a Sentence Final Particle

DE can be used as a sentence end particle, though this usage is not so common in written text. In Penn Treebank’s Chinese corpus, 1,973 out of 64,402 occurrences of DE are found to be sentence end particles, labeled as “SP” in this annotated corpus. Lǚ (1999) indicates that DE can be used at the end of a sentence to express an affirmative tone or a realis mood.

In example (45) and (46), the two sentences have the same meaning, but (46) has an affirmative tone.

- (45) 他 要 走
 tā yào zǒu
 He will go

“He will go”

(46) 他 要 走 的

tā yào zǒu de

He will go DE-sfp

“He will go (with affirmative tone)”

In some sentences, adding DE to the sentence end can introduce a realis mood. For example, in (47), the event has not happened yet (future tense), while in (48), the sentence final DE indicates that the speaker considers the event to be a known state of affairs, i.e., it has already happened (past tense).

(47) 他 什么 时候 走 ？

tā shénme shíhòu zǒu ?

He what time go ?

“When will he go?”

(48) 他 什么 时候 走 的 ？

tā shénme shíhòu zǒu de

He what time go DE-sfp ?

“When did he go?”

4.7 Reduplicated Adjectives and Verbs

According to Li & Thompson (1989), reduplication is the morphological process of repeating a morpheme to form a new word, which mainly applies to verbs and adjectives in Chinese. When a monosyllabic adjective or verb is reduplicated, the character is repeated ($A \rightarrow AA$), as shown in (49) and (50).

(49) 红红
hónghóng
red-red
“very red”

(50) 看看
kànkàn
look-look
“take a look”

When reduplication is applied to disyllabic words, the two characters are repeated differently for adjectives (AB → AABB) and verbs (AB → ABAB), as illustrated in (51) and (52).

(51) 干干净净
gāngānjìngjìng
AABB-clean
“very clean”

(52) 休息休息
xiūxiūxiūxi
rest-rest
“have a rest”

Syntactically, the reduplicated adjectives can not be modified by degree adverbs (e.g. 很 *hen* “very”, 非常 *feichang* “extremely”, 特别 *tebie* “specially”, 极 *ji* “extremely”, 十分 *shifen* “very much”, 更 *geng* “more”, 最 *zui* “most”, 较 *jiao* “more”, 比较 *bijiao* “more”, etc.), as illustrated in (53).

- (53) *很 干干净净
hěn gāngānjìngjìng
very AABB-clean

Reduplicated verbs, on the other hand, do not accept aspect markers like 了 *le*, 着 *zhe*, and 过 *guo*, as shown in (54).

- (54) *看看 着
kànkàn zhe
look-look ASP
“take a look”

The meaning of the reduplicated adjectives (AA or AABB) is more vivid or intensified than its original form (A or AB) (Li & Thompson, 1989). For verbs, reduplication adds a tentative aspect (Chen et al., 1992), or signals a delimitative aspect (doing something “a little bit”) (Li & Thompson, 1989).

4.8 Interrogative Sentences

Interrogatives, or questions, are one of the five major types of sentences classified according to the functions they play in communication. The rest four types are declaratives, exclamatives, directives and vocatives.

In Mandarin Chinese, questions can be formed by uttering a declarative sentence in a rising tone, as stated in both Li & Thompson (1989) and Li & Cheng (1994). Li & Thompson (1989) also summarizes four grammatical devices that can be used to explicitly mark an utterance as questions.

1. **Content questions** (or *wh* questions or **open questions**) use question words which are similar to words like *who*, *where*, *why*, etc in English.

(55) 谁 哭 了 ？

shuí kū le ？

Who cry ASP ？

“Who cried ?”

2. **Disjunctive questions** offer two options as possible answers. Li & Cheng (1994) further divide this type into **yes/no questions** (or **polar questions**) expecting *yes* or *no* as answers (which belongs to the type of **particle questions** listed below), the **affirmative-negative questions**, and the **alternative questions**.

- (a) The **affirmative-negative questions** are formed by putting the affirmative and negative forms of the predicate together, ie. A-not-A, as shown in (56),

(56) 你 去 不 去 ？

nǐ qù bu qù ？

You go not go ？

“Will you go?”

- (b) The **alternative questions** present the two options as two clauses joined together using a morpheme 还是 *haishi*, as shown in the example below.

(57) 你 去 还是 他 来 ？

nǐ qù háishì tā lái ？

You go or he come?

“Will you go, or will he come?”

3. **Particle questions** (another kind of **open question**), formed by having a question particle 吗 *ma* attached to a declarative sentence in sentence-final position.

- (58) 你 去 吗 ？
nǐ qù ma ？
You go Q-ma ？

“Will you go?”

Li & Cheng (1994) points out that the particle may also be 吧 *ba*, which then suggests the speaker is not sure about a certain situation therefore he asks this question, expecting an affirmative answer.

- (59) 你 去 吧 ？
nǐ qù ba ？
You go Q-ba ？

“You will go, right?”

4. **Tag questions** are formed by attaching an A-not-A form to the end of a statement.

- (60) 你 去 ， 对 不 对 ？
nǐ qù , duì bú duì ？
You go , right not right ？

“You will go, right?”

Li & Cheng (1994) also specifies whether different question types may take up modal particles at sentence end position.

Content questions, affirmative-negative questions, and alternative questions can also be ended with modal particles 呢 *ne* and 啊 *a* to make the tone more tactful/moderate, but 吗 *ma* is not allowed at the end of such sentences.

The modal particle 呢 *ne* can be used at the end of a one-word/phrase sentence to express interrogation of where or how the person or thing concerned is. 吗 *ma* can't be added to the end of such questions.

Questions involving A-not-A (disjunctive questions, tag questions) have already been analyzed and implemented by my colleagues in NTU Computational Linguistics Lab (Wang et al., 2015). They will not be discussed further here.

Chapter 5

HPSG Modeling of Phenomena

Zhong is based on ManGO, which in turn was built using the grammar matrix (Bender et al., 2002). There is extensive documentation for it online, including an excellent series of Frequently Asked Questions from classes on grammar engineering: <http://moin.delph-in.net/GrammarEngineeringFaq>, originally developed by Emily Bender at the University of Washington and ported to the DELPH-IN wiki by Ned Letcher. The basic feature geometry is shown in Figures 5.1 and 5.2, taken from the description of feature geometry in the FAQ.¹

To save space, in the illustrations in this chapter, we use abbreviations to represent the feature structures, for example SYN for `SYNSEM.LOCAL.CAT`, SEM for `SYNSEM.LOCAL.CONT`.

Specifically, the semantic information that deep linguistic analysis focuses on is contained in `SYNSEM.LOCAL.CONT`. Here `HOOK` carries the information available for further semantic composition. Inside `HOOK`, one finds the following:

1. `LTOP`: local top handle
2. `INDEX`: index (individual or event) of the local sign
3. `XARG`: distinguished argument of the local sign, available for control by outside predicates

The above features are used in the modeling of the phenomena in the following sections.

¹<http://moin.delph-in.net/GeFaqFeatureGeometry>

phrase															
SYNSEM	LOCAL	CAT	HEAD	MOD list		KEYS	KEY predsor		ALTKEY	predsort					
				SPR < ... >											
				SUBJ < ... >											
				COMPS < ... >											
				SPEC < ... >											
		CONT	HOOK	INDEX individual											
				LTOP handle											
				XARG individual											
		RELS	HCONS	diff-list											
				diff-list											
AGR	COORD	agr													
		coord													
NON-LOCAL	SLASH	diff-list	QUE	diff-list	REL	diff-list									
MODIFIED	xmod														
LIGHT	bool														
(NON)-HEAD-DTR	LOCAL	CAT	HEAD	MOD list		KEYS	KEY predsor				ALTKEY	predsort			
				SPR < [LOCAL ...] >											
				SUBJ < [LOCAL ...] >											
				COMPS < [LOCAL ...] >											
				SPEC < [LOCAL ...] >											
		CONT	HOOK	INDEX individual											
				LTOP handle											
				XARG individual											
		RELS	HCONS	diff-list											
				diff-list											
AGR	COORD	agr													
		coord													
NON-LOCAL	SLASH	diff-list	QUE	diff-list	REL	diff-list									
MODIFIED	xmod														
LIGHT	bool														
C-CONT	HOOK	INDEX individual													
		LTOP handle													
		XARG individual													
		RELS diff-list													
		HCONS diff-list													

Figure 5.1: Basic Matrix Features for Phrases

5.1 DE Construction Modifying Nouns

In section 4.2, the different types of DE constructions identified are summarised in (13), which is repeated below.

HPSG grammars use the lexical type hierarchy to structure the lexicon, defining the shared properties of the words in a general type, which are inherited by its sub-types. The properties that are specific to the sub-types are then defined at the sub-type level. The different lexical types created for the respective types of DE constructions in (13) are illustrated in the type hierarchy in (61).

At the top of the hierarchy, the super type, `de-super-lex`, specifies the common properties of all DEs which head DE constructions: its headtype is a postposition expecting a complement located before it. The position of the complement is specified with the positive value of the binary feature, `POSTCOMP`. The exact type of complement is underspecified at this level as it might be a noun phrase, an adjective phrase, or a verbal phrase. And its `HEAD` feature, `MOD`, defines that the DE phrase intersectively modifies a noun with its complements saturated. Details of this lexical type are shown in (62).

de-super-lex has two sub-types, de-comp-lex and de-assoc-lex. de-assoc-lex defines the features for the associative DE while de-comp-lex defines the relativizing DE. Details of the two sub-types are described in the next sections.

5.1.1 Associative DE

Associative DE plays the role of connecting an NP or a clause to another NP to form a larger noun phrase, with the NP or clause preceding DE modifying the NP following DE. It is further differentiated into the associative DE for NPs or clauses, depending on whether the modifier is an NP or a clause respectively. Therefore, the lexical type `de-assoc-lex`, is created for the general associative DE, as shown in (63). It inherits all the features from `de-super-lex` and has two sub-types under it, `de-assoc-n-lex` and `de-assoc-v-lex`.

$$\begin{array}{l}
(63) \quad \left[\begin{array}{l} \text{de-assoc-lex} \\ \\ \text{SYN} \left[\begin{array}{l} \text{postp-assoc} \\ \text{HEAD} \left[\text{MOD} \left\langle \left[\text{SEM} \left[\begin{array}{l} \text{INDEX } s_1 \\ \text{LTOP } l \end{array} \right] \right\rangle \right] \right] \\ \text{VAL.COMPS} \left\langle \boxed{1} \right\rangle \\ \text{POSTHEAD -} \end{array} \right] \\ \\ \text{SEM} \left[\begin{array}{l} \text{LTOP } l \\ \text{RELS} \left\langle \begin{array}{l} \text{\textit{de-p-assoc-rel}} \\ \text{ARG1} \quad s_1 \\ \text{ARG2} \quad s_2 \end{array} \right\rangle \end{array} \right] \\ \\ \text{ARG-ST} \left\langle \boxed{1} \right\rangle \left[\begin{array}{l} \text{HEAD +nv} \\ \text{SYN} \left[\begin{array}{l} \text{VAL} \left[\begin{array}{l} \text{SPR} \langle \rangle \\ \text{SUBJ} \langle \rangle \\ \text{COMPS} \langle \rangle \end{array} \right] \\ \text{SEM} \left[\begin{array}{l} \text{INDEX } s_2 \\ \text{SF prop} \end{array} \right] \end{array} \right] \end{array} \right] \end{array} \right]
\end{array}
\right]
\end{array}$$

This lexical type defines the associative DE to be an associative post-position that takes a complement headed either by a noun or a verb. The binary feature `POSTHEAD` is set to be negative, indicating the DE construction has to precede the phrase it modifies. It introduces a predicate *de-p-assoc-rel* with its `ARG1` linked to the noun that the DE structure modifies and its `ARG2` linked to the NP/clause modifier. This predicate covers possessive or genitive meaning, as well as other associative meanings.

Associative DE with NP

The subtype *de-assoc-n-lex*, illustrated in (64), further defines DE for the scenario when the modifier in associative DE construction is an NP itself. Here DE, specifically, is an associative post-position for NP. It takes an NP complement to form an associative DE structure.

$$(64) \left[\begin{array}{c} \text{de-assoc-n-lex} \\ \text{SYN} \left[\begin{array}{c} \text{HEAD postp-assoc-n} \\ \text{VAL.COMPS} \left\langle \left[\text{SYN.HEAD noun} \right] \right\rangle \end{array} \right] \end{array} \right]$$

The examples for associative DE with NP are repeated here:

- (14) 我 的 朋 友 哭 了
wǒ de péngyou kū le
I DE-asc friend cry ASP
“My friend cried”

- (15) 他 知 道 学 校 的 地 址
tā zhīdào xuéxiào de dìzhǐ
He know school DE-asc address
“He knows the address of the school”

Our implementation leads to the tree and MRS in (65) and (66) for the example in (14), and (67) and (68) for (15). In the syntactic tree, the DE phrase is labeled with “deP”. The MRS shows the predicate *de_p_assoc_rel* has its ARG2 pointing to the modifier, and its ARG1 pointing to the head noun it modifies.

(65)

(66)

(67)

(68)

The associative DE might be omitted when the noun preceding DE is a personal pronoun, as in (69). This structure is handled by the *pronoun-noun-compound-phrase* rule which combines a pronoun followed by a noun to form a compound noun phrase. This rule makes the semantics similar by introducing a predicate *_de_p_assoc* with the pronoun (the modifier) as its ARG2 and the head noun as its ARG1, as illustrated in (71). The tree is shown in (70).

- (69) 我 朋友 哭 了
wǒ péngyou kū le
I friend cry ASP
“My friend cried”

In other words, the possessive relation indicated by a personal pronoun might be realized with DE (14) or without DE (69). The semantics of the two realizations are both represented using the same predicate *_de_p_assoc_rel*.

This predicate inherits from *related_p_rel*, which is used to represent the more general relatedness of two entities. It has another predicate as its subtype, *compound_p_rel*, which is used for the relatedness introduced in compound noun phrases. (72) shows the type hierarchy of the predicates.

The associative DE and noun-noun compound may appear in the same noun phrase, as in (73). The semantic representation of this structure (75) contains both *de_p_assoc_rel* and *compound_p_rel*, linking the head noun with the modifier in the DE structure and the modifier in the compound structure respectively.

- (73) 他 解决 了 产品 的 质量 问题
 tā jiéjué le chǎnpǐn de zhìliàng wèntí
 He resolve ASP product DE-asc quality problem
 “He resolved the quality problem of the product”

It's also possible to have more than one associative DE structure modifying a head noun, as in (76). This can lead to bracketing ambiguity of the noun phrase, e.g. (产品的质量) 的问题) or (产品的 (质量 的问题)). Our implementation allows both interpretations, leaving the ambiguity to be resolved by a statistical ranking module afterwards. The semantic representations for the two interpretations differ slightly in whether “产品”(“product”) is linked with “质量”(“quality”) or “问题”(“problem”) in the associative predicates.

- (76) 他 解决 了 产品 的 质量 的 问题
 tā jiéjué le chǎnpǐn de zhìliàng de wèntí
 He resolve ASP product DE-asc quality DE-asc problem
 “He resolved the quality problem of the product”

(78)

Associative DE with Clause

As mentioned above, the associative DE can be also used to link a clausal modifier to a head noun. In this scenario, the head noun is not an argument (neither subject nor object) of the event in the clause, as shown in example (12) and repeated here:

- (12) 我 喜欢 他 写 书 的 地方
wǒ xǐhuān tā xiě shū de dìfāng
I like he write book DE-asc place

“I like the place where he wrote books”

The lexical type for such sentences, *de-assoc-v-lex*, is similar to *de-assoc-n-lex*, except that the head of its complement is constrained to be a verb.²The ARG1 of the

²Or more properly verb or adjective.

predicate *de_p_assoc_rel* is linked to the head noun that the DE structure modifies, and its ARG2 is linked to the verb predicate.

$$(79) \left[\begin{array}{l} \text{de-assoc-v-lex} \\ \text{SYN} \left[\begin{array}{l} \text{HEAD postp-assoc-v} \\ \text{VAL.COMPS} \left\langle \left[\text{SYN.HEAD verb} \right] \right\rangle \end{array} \right] \end{array} \right]$$

This leads to the following results, illustrated in (80) and (81). The DE phrase ("deP") is composed of a clause ("S") and DE("de"). In the DMRS, the predicate *de_p_assoc_rel* links the head noun and the head verb of the clausal modifier as its ARG1 and ARG2, respectively.

5.1.2 Relativizing DE constructions

The features for the relativizing DE are defined in the lexical sub-type, `de-comp-lex`. It inherits the specifications from `de-super-lex` and defines more information specific to the relativizing DE, as shown in (82).

This lexical type further defines that its headtype is `postp-comp`, a complementizing post-position. DE's complement must be a clause headed by a non-copular verb or an adjective (`+vj`: Verb or adjective). This clause has a gapped argument which is an NP.

HPSG grammars use non-local features to handle long-distance dependencies that require gapped argument and the filling of such gap in analysis, such as topicalization (Sag et al., 2003, who call the feature `GAP`). Rules like `extracted-subj-phrase` and `extracted-comp-phrase` already exist in the grammar to perform the task of extracting the respective

argument from the verb or adjective and linking it to the NON-LOCAL feature called SLASH. For example, in topicalization, the topicalized argument is extracted from the clause as gap and then *filler-head-phrase* identifies the gap with the NP preceding the clause.

A similar approach is used here to handle the relativizing DE. Its SLASH feature defines the gapped argument, the INDEX and LTOP of which are co-indexed with those of the NP that the DE phrase modifies, as defined in DE's MOD feature. With the above implementation, the sentence in (11) is parsed into the tree shown in (83). The DE structure is labeled with "deP/" to indicate that this is a DE phrase with a slashed(gapped) argument. The relative clause inside is labeled with "S/NP", signaling a missing NP argument.

- (11) 张三 喜欢 写 书 的 人
zhāngsān xǐhuān xiě shù de rén
Zhangsan like write book DE-rel person

"Zhangsan likes people who write book(s)"

(84) shows the MRS representation, in which the subject(ARG1) of “write” ($_写_v_1$) is identified with the ARG0 of “person” ($_人_n_1$). The same semantic information is displayed more intuitively as dependencies in (85), where an arc points from “write” ($_写_v_1$) to “person” ($_人_n_1$), labeled with ”ARG1”. In all subsequent examples, only the dependency graph will be shown for easier reading.

[Note: The relative clause modifies the head noun intersectively, therefore $_写_v_1$ and $_人_n_1$ share the same LBL, which leads to the “EQ” label on the arc.]

In the case of a gapped object, as in example (16), the parse tree is shown in (86), and the dependencies in (87), in which the noun “book” (书 *n_1*) is correctly identified as the object (ARG2) of “write” (写 *v_1* *rel*).

- (16) 张三 喜欢 李四 写 的 书
zhāngsān xǐhuān lǐsì xiě de shù
Zhangsan like Lisi write DE-rel book

“Zhangsan likes book(s) that Lisi wrote”

(86)

(87)

For (17), when the determiner+classifier is immediately before the head noun:

- (17) 张三 喜欢 李四 写 的 这 本 书
zhāngsān xǐhuān lǐsì xiě de zhè běn shù
Zhangsan like Lisi write DE-rel this piece book

“Zhangsan likes this book that Lisi wrote”

[Note: Here *_本_x* and *_书_n_1* share the same LBL. *_写_v_1* and *prox_dem_q* share the same LBL. This is because the LBL of *_写_v_1* is identified with the LBL of what it modifies, the NP. And at NP (DP+N) level, the HOOK of phrase is the same as the HOOK of the non-head-daughter, the determiner. In previous examples, the HOOK of NP is the same as the HOOK of head daughter, N.]

For (18), where the DP appears before the relative clause:

- (18) 张三 喜欢 这 本 李四 写 的 书
zhāngsān xǐhuān zhè běn lǐsì xiě de shù
Zhangsan like this piece Lisi write DE-rel book

“Zhangsan likes this book that Lisi wrote”

[Note: The DMRS of this example differs from that of the previous example in the label of the arc between `_写_v_1` and `_书_n_1`, which has “EQ”, indicating the two has the same LBL.

Ambiguity in Gapping The SLASH feature is defined to take one element at most. When both subject and object are missing in the clause, it is an ambiguous case with two possible parses, one with the gapped subject interpretation and one with the gapped object interpretation.

In the relative clause of example (92), neither the subject nor the object of the verb appears. So theoretically the head noun that the relative clause modifies could refer to the missing subject or the missing object.

- (92) 张三 喜欢 写 的 书
zhāngsān xǐhuān xiě de shù
Zhangsan like write DE-rel book

“Zhangsan likes book(s) that is(are) written”

Both interpretations can be captured in our implementation. Their labeled trees, (93) for subject interpretation and (94) for object interpretation, may look similar in the simplified representation, but they differ in the rules that have been applied leading to the structure, especially at the argument extraction step. In (93), subject extraction rule *extracted-subj-phrase* identifies the missing argument of the verb to be its subject, illustrated in the tree as *S/NP-subj*. In (94), the rule *extracted-comp-phrase* identifies the missing argument of the verb to be its object, shown as *S/NP-obj* in the tree. The semantic distinction is evident in their MRS representations, as shown in (95) for the subject interpretation where the arc is labeled as ARG1 (subject) and in (96) for the object interpretation where the label is ARG2 (object).

(94)

(95)

(96)

The Role of SUO The marker 所 *SUO* used in a relative clause before its head verb indicates that the noun gapped in the relative clause must be the object of the verb. The detailed modeling of *SUO* is presented in Section (5.4) below.

Gapping with Ditransitive Verbs In Chinese sentences with ditransitive verbs (V + Obj1 + Obj2), the first object (ARG2) is the indirect object and the second object (ARG3) is the direct object.

In relative clauses with ditransitive verbs, the gapping commonly happens to the subject (97) and the direct object (98). When the indirect object is actually the head noun, its original position in the relative clause is typically replaced using a pronoun (Li & Thompson, 1989), as illustrated in (99).

- (97) 张三 是 给 李四 书 的 人
 zhāngsān shì gěi lǐsì shù de rén
 Zhangsan is give Lisi book DE-rel person
 “Zhangsan is the person who gave Lisi book(s)”

- (98) 张三 给 李四 的 书 很 贵
 zhāngsān gěi lǐsì de shù hěn guì
 Zhangsan give Lisi DE-rel book very expensive
 “The book(s) which Zhangsan gave to Lisi is(are) expensive”

- (99) 李四 是 张三 给 他 书 的 人
 lǐsì shì zhāngsān gěi tā shù de rén
 Lisi is Zhangsan give him book DE-rel person
 “Lisi is the person to whom Zhangsan gave book(s)”

We get the correct parse tree for (97) as shown in (100) below. The dependencies we get in (101) also correctly link 人_{n-1} (“person”) to the gapped subject (ARG1) of the verb 给_{v-1} (“give”) and to the ARG2 of 是_{v-cop} (“is”).

(100)

(101)

For the example (98), we get the right tree structure (102), and the gapping of the direct object (ARG3) is displayed in (103).

(102)

(103)

For clauses with neither direct object nor indirect object present, like in (104), normally the gapped argument should be the direct object (ARG3)(Li & Thompson, 1989). (105) and (106) show this interpretation correctly.

(104) 张三 给 的 书 很 贵
 zhāngsān gěi de shù hěn guì
 Zhangsan give DE-rel book very expensive

“The book(s) that Zhangsan gave is(are) very expensive”

(105)

(106)

5.2 Nominalizing DE Constructions

Implementation Forming of the nominalizing DE structure is handled similarly to the relativizing DE or associative DE, as described in the above sections. The nominalization of the DE structure is achieved through the rule *bare-nominal-postp-phrase*, which pushes nominalizable postpositional phrases to be noun phrases. The rule introduces a predicate *generic_n_rel* (defined in *zhong.tdl*) representing a generic noun which is linked with the slashed argument in the relativizing DE structure, or ARG1 in *_de_p_assoc_rel* in the associative DE structure.

Currently for sentence in (19), which is repeated below, we get the parse tree in (107), with the *bare-nominal-postp-phrase* shown boxed and the DMRS representation as in (108).

- (19) 我 喜欢 很 红 的
 wǒ xǐhuān hěn hóng de
 I like very red DE-nom
 “I like the very red (thing)”

For example (20), which is repeated here, the parse tree is in (109). Here the slashed argument for V is the omitted object.

- (20) 我 喜欢 他 写 的
wǒ xǐhuān tā xiě de
I like he write DE-nom
“I like what he wrote (the thing that he wrote)”

For example (21), repeated below, theoretically the missing argument of the verb might refer to either its subject or object. Our implementation can get both interpretations. The parse tree (111) looks identical for both interpretations, but the semantic distinction is reflected in the MRS representations, with (112) showing the subject interpretation and (113) for the object interpretation.

(21) 我 喜欢 吃 的

wǒ xǐhuān chī de

I like eat DE-nom

“I like food (what can be eaten)”

(111)

(112)

(113)

For example (22), which is repeated below, as illustrated in (114) and (115), the associative DE forms the DE structure with the pronoun “我”(“I”), taking the pronoun as its ARG2. The ARG1 of DE is linked to the generic noun introduced by *bare-nominal-postp-phrase*.

- (22) 我 的 哭 了
wǒ de kū le
I DE-nom cry ASP
“Mine cried”

5.3 Predicative SHI + Nominalization

These constructions are composed of the copular SHI followed by the nominalization. In Zhong’s implementation, the copular verb SHI introduces a predicate 是_v_cop_rel, which has its ARG1 linked to the subject of SHI and ARG2 to its object. And as described in the previous section, DE nominalization leads to an NP with a predicate representing a generic noun,

generic_n_rel. This NP is handled normally by the copular SHI, which links the ARG2 of 是_v_cop_rel to *generic_n_rel*.

For (30), repeated below, the ARG2 of 是_v_cop_rel is linked to *generic_n_rel*, which is the ARG1 of _de_p_assoc_rel, as shown in (117).

(30) 那 是 我 的
 nà shì wǒ de
 that is I DE-asc
 “That is mine”

For (31), which is repeated here, the ARG2 of 是_v_cop_rel is linked to the *generic_n_rel*, which is the ARG2 of _吃_v_l_rel, as shown in (119).

- (31) 苹果 是 吃 的
 píngguǒ shì chī de
 apple is eat DE-nom

“Apples are food”

For (32), repeated here, the ARG2 of 是_v_cop_rel is linked to the generic_n_rel, which is the ARG2 of 甜_a_1_rel, as shown in (121). In this implementation, the link between the subject ‘apple’ and the adjective ‘sweet’ is achieved indirectly through generic_n_rel. If necessary, a machine-translation style transfer rule can be used later to derive the direct relation.

- (32) 苹果 是 甜 的
 píngguǒ shì tián de
 apple is sweet DE

“Apples are sweet”

5.4 SUO-DE structure

We take the view of Deng (2009) that in structures where both SUO and DE appear (25a-d), DE plays the key role of nominalizing the phrase “(NP₁+)SUO+V+DE”, so that it can either be a noun phrase itself, or be a prenominal adjunct (relative clause) to NP₂. The role of SUO in the construction is to indicate that the missing argument of the verb is its patient or direct object.

Specifically, for "(NP₁+)SUO+V+DE+NP₂", the lexical entry for the relativizing DE is based on *de-comp-lex*, as presented in (122). Here DE expects a complement which is a clause headed by verb or adjective. The clause misses one NP argument (as shown in its SLASH value which is co-indexed with the NP that the DE construction modifies).

Like BA and BEI, we treat SUO as a "verby" marker which takes a verbal complement. The lexical entry for SUO is shown in (123). Through co-indexing of the semantic feature *HOOK.INDEX*, it defines SUO's subject to be the same as its verbal complement's subject, which has to be unrealized yet. It also defines the verb complement's direct object to be the slashed argument of SUO structure (hence constraining the missing argument to be the object).

With SUO, the argument extraction rule no longer needs to be applied. The HOOK of SUO is identical to that of its complement. Other than this, SUO has nothing to add on semantically. It's worth noting that SUO is redundant when NP_1 is present. When NP_1 is not present, SUO helps to restrict the reading of the gap.

(122) and (123) interact to produce the desired analysis for the example sentence in (2), which is repeated below. In this example, the relative clause has its subject present, its meaning is the same as the sentence without SUO (16).

(2) 张三 喜欢 李四 所 写 的 书
 zhāngsān xǐhuān lǐsì suǒ xiě de shù
 Zhangsan like Lisi SUO write DE-rel book

“Zhangsan likes book(s) that Lisi wrote”

(16) 张三 喜欢 李四 写 的 书
 zhāngsān xǐhuān lǐsì xiě de shù
 Zhangsan like Lisi write DE-rel book

“Zhangsan likes book(s) that Lisi wrote”

The parse tree of (2) is shown in (124). Its DMRS representation is identical to that of (16).

But for ambiguous sentences like (92), the existence of SUO, as in (23), constrains the interpretation of the sentence to that of the gapped object.

- (92) 张三 喜欢 写 的 书
zhāngsān xǐhuān xiě de shù
Zhangsan like write DE-rel book
“Zhangsan likes book(s) that is(are) written”

- (23) 张三 喜欢 所 写 的 书
zhāngsān xǐhuān suǒ xiě de shù
Zhangsan like SUO write DE-rel book
“Zhangsan likes book(s) that were written”

For (23), with SUO constraining the extracted argument to be the object of its verbal complement, the parser gets the correct object interpretation.

(127)

5.5 Emphatic SHI-DE

In emphatic SHI-DE constructions, the constituent emphasized usually describes circumstances of the action, affirming or denying some supposition. The element being emphasized can be adverbials, prepositional phrases, subjects, modal verbs, participle clauses, etc., of the verb that appear between SHI and DE. Examples for emphasizing temporal adverbials (29, 34, and 35) have been given in Section 4.5 above and are repeated here.

- (29) 他是昨天来的
tā shì zuótiān lái de
he is yesterday come DE

“He came yesterday/ It was yesterday that he came”

- (34) 他是昨天完成任务的
tā shì zuótiān wánchéng rènwù de
he is yesterday finish task DE

“He finished the task yesterday / It was yesterday that he finished the task”

- (35) 他是昨天完成的任务
tā shì zuótiān wánchéng de rènwù
he is yesterday finish DE task

“He finished the task yesterday / It was yesterday that he finished the task”

For the emphatic SHI-DE constructions, the existing implementation of nominalization can already provide a reasonable analysis — except for the emphatic information. For example for (29), the grammar produces the parse tree in (128). The time adverbial 昨天 (zuótiān ‘yesterday’) is a temporal noun transformed into a PP to combine with the verb 来 (lái ‘come’) to form a VP. The subject of the VP becomes the slashed argument for the DE construction (deP). Through nominalization, deP is transformed into NP, and a dummy predicate *generic_n_rel* is created and co-indexed with the slashed argument (the subject of 来). This dummy predicate is also linked as ARG2 (object) of the predicate for SHI, 是_v_cop. So for (29), its MRS (129) means “He is the one who came yesterday”. However, additional steps need to be taken so that we can capture the emphatic information into the MRS, as in “It was yesterday that he came”.

(129)

There are two options to approach this. The first option is to reuse the current nominalization analysis and find a way to identify the emphasized constituent. In this way, the syntactic tree of the analysis will look similar to what's presented in the Penn Chinese Tree Bank (130), in which the emphasized temporal phrase (PP-TMP) is a constituent in the nominalized DE phrase (NP-PRD).

(130)

```
<S ID=1852>
( ( IP (NP-PN-SBJ (NR 龙永图))
      (VP (VC 是)
          (NP-PRD (CP (WHNP-1 (-NONE- *OP*))
                      (CP (IP (NP-SBJ (-NONE- *T*-1))
                              (VP (PP-TMP (P 于)
                                          (NP (NT 1 8 日)))
                                          (PP-DIR (P 从)
                                          (NP-PN (NR 波士顿)))
                                          (VP (VP (VV 来)
                                                  (NP-PN-OBJ (NR 华盛顿)))
                                                  (VP-PRP (VV 参加)
                                                  (NP-OBJ (NN 演讲))))))
                      (DEC 的))))))
      (PU 。)) )
</S>
```

Though this option seems to require least work in creating the syntactic tree, getting the emphatic information right requires reaching deep into the phrase structure to identify the left most constitute or leaf node in the DeP. The grammar does not have a readily available mechanism to easily enable such tracing now, and it violates standard assumptions about locality.

The second option follows the ERG's (Flickinger, 2000b) way of analysing English cleft sentences with a special BE verb that takes two complements, one for the emphasized constituent, and the other for the THAT clause. Our implementation takes this approach, creating

a special emphatic SHI which expects two complements, the emphasized constituent, and the DeP. As the emphasized constituent will no longer be combined with the embedded verb as its subject, object, ADVP, or PP, special care needs to be taken to ensure proper linking is created between that constituent and the verb, depending on the type of the emphasized constituent. This means, specific subtypes of emphatic SHI need to be defined to handle such linking correctly.

Lexical types The structure of the lexical subtypes for emphatic SHI is illustrated in (131). Once again, we use the type hierarchy to share relevant information.

At the top level, lexical type `shi-emph-lex` is created to represent the general emphatic SHI. It has two subtypes, `shi-other-dep-lex` (for emphasis of other constituents describing the circumstances of the event) and `shi-subj-dep-lex` (for emphasis of the subject). `shi-other-dep-lex` is further defined to have two subtypes, `shi-mod-dep-lex` and `shi-aux-dep-lex`, with the former for the emphasis of verb modifiers, like ADVP or ADP, and the latter for the emphasis of auxiliary verbs. This distinction is necessary as auxiliaries are linked with the main verb differently from the way ADVP or ADP are linked with the main verb. Details are explained in the following paragraphs. Verb modifiers are further differentiated into intersective modifiers and scopal modifiers, leading to two subtypes of `shi-mod-dep-lex`, `shi-mod-int-dep-lex` and `shi-mod-scop-dep-lex`.

shi-emph-lex The lexical type *shi-emph-lex* is defined in (132). It specifies its head to be a copular verb and takes two compulsory complements (indicated by feature *OPT*). The first complement is underspecified as the emphasized constituent can be a few types of phrases. The second complement is *deP*, therefore it must be headed by *postp-comp* and has its own complement already saturated.

shi-other-dep-lex The lexical type *shi-other-dep-lex* inherits everything from *shi-emph-lex* and further indicates that SHI is a verb that takes three arguments (subject and two complements). Its subject, a noun, is co-indexed as the slashed argument in *deP*, which is linked to *deP*'s MOD value.

shi-subj-dep-lex The lexical type *shi-subj-dep-lex* inherits features from *shi-emph-lex* and further states that SHI only takes two arguments (its two complements) as there's no

subject for SHI when it's emphasizing the subject of the embedded clause (see (38)).

Individual Constraints (ICONS) are used to show the information structure in Zhong, following the method described in Song (2017). For all emphasized constituents except scopal modifiers, `focus-icons-int-lex-item` is used to introduce a `focus` element to `CONT.ICONS`, linking its `IARG1` to the index (`LOCAL.CONT.HOOK.INDEX`) of the second complement (deP), and its `IARG2` to the index of the first complement (the emphasized phrase). But for a scopal modifier, since its `INDEX` is already defined to be the same as the `INDEX` of the modified, `LEYS.KEYREL.ARG0` is used instead to refer to the emphasized item so that the `IARG1` and `IARG2` of `focus` do not point to the same index, as defined in `focus-icons-scop-lex-item`.

Emphasizing verb modifiers `shi-mod-dep-lex` inherits from `shi-other-dep-lex` and further defines that the first complement is either `ADVP` or `PP`, and the `INDEX` and `LTOP` of the modified (its `MOD` value) is linked to the `INDEX` and `LTOP` of DeP. Since the `INDEX` and `LTOP` of DeP are already defined to be the same as those of its embedded clause, this allows us to say that what the `ADVP/ADP` modifies is the clause embedded in DeP.

(137)

(138)

For scopal modifiers as in (41), *shi-mod-scop-dep-lex* is defined in a similar manner, inheriting from *shi-mod-dep-lex* and *focus-icons-scop-lex-item*, and stating that its first complement (the emphasized constituent) modifies its head scopally. This gives the analysis shown in (140) and (141), with the predicate for the adverb *绝对_a.3* and that for the verb *哭_v.1* linked through *qeq*.

(139)
$$\left[\begin{array}{l} \text{shi-mod-scop-dep-lex} \\ \text{SYN.VAL.COMPS} \left\langle \left[\text{SYN.HEAD.MOD} \left\langle \left[\text{LOCAL scopal-mod} \right] \right], \text{synsem} \right\rangle \right. \end{array} \right]$$

(41) 他是绝对 哭了的
 tā shì juéduì kū le de
 he is definitely cry ASP DE
 “He DEFINITELY cried”

(140)

(141)

Emphasizing auxiliary verbs For emphasis of auxiliary verbs, the linking between the aux verb and the verb needs to be done through `qeq`, with its `HARG` pointing to `ARG1` of the predicate of the aux verb, and `LARG` pointing to `LBL/LTOP` of the verb's predicate. This is achieved by `shi-aux-dep-lex`, which states that `SHI`'s first complement is headed by an `AUX` verb, and links the `AUX` verb and the main verb by identifying the `INDEX` and `LTOP` values of the `AUX`'s verbal complement as the same as those of `deP` (the second complement for `SHI`). The lexical type of `DE` defines that the `HOOK` value of `DE` is the same as its complement, by inheriting from `raise-sem-lex-item`.

$$(142) \left[\begin{array}{c} \text{shi-aux-dep-lex} \\ \\ \text{SYN.VAL.COMPS} \left\langle \begin{array}{c} \text{SYN} \left[\begin{array}{c} \text{HEAD} \left[\begin{array}{c} \text{verb} \\ \text{AUX } + \end{array} \right] \\ \\ \text{VAL.COMPS} \left\langle \text{SEM.HOOK} \left[\begin{array}{c} \text{INDEX } i \\ \text{LTOP } l \end{array} \right] \right\rangle \end{array} \right] \\ \\ \left[\begin{array}{c} \text{SEM.HOOK} \left[\begin{array}{c} \text{INDEX } i \\ \text{LTOP } l \end{array} \right] \end{array} \right] \end{array} \right\rangle, \end{array} \right]$$

So for the sentence in (42), we get a tree like (143) and DMRS as in (144) :

(42) 他 是 会 完 成 任 务 的
tā shì huì wánchéng rènwù de
he is will finish task DE
“He WILL finish the task”

The modal verb can be modified by the negation adverb 不 (bù “not”), as in (145). In such case, 不 and the modal verb are combined first before they are taken as the first complement of the emphatic SHI, as shown in (146) and (147).

- (145) 他 是 不 会 完 成 任 务 的
 tā shì bú huì wánchéng rènwù de
 he is not will finish task DE

“He WILL NOT finish the task”

On the other hand, when there's other adverb before (NEG+) AUX, the scope of the emphasized constituent is ambiguous. It might be the adverb after SHI, or the adverb + (NEG+) AUX. The grammar can cover the former (as shown in 149 and 150), but not the latter, since scopal adverbs expects the VP to have saturated COMPS in lexical type definition, which means the AUX must combine with its COMPS (the main verb) first.

(148) 他是绝对 不会 完成任务 的

tā shì juéduì bú huì wánchéng rènwù de

he is absolutely not will finish task DE

“He ABSOLUTELY WILL NOT / will not finish the task”

(149)

(150)

Alternative position of object For the alternative position of the object of the embedded verb, as in (35), a special type for DE, *de-4shi-lex*, is created to identify that the NP appearing after DE is actually the omitted object in the embedded clause of *deP*. The HEAD type of this DE is also *postp-comp*, and it expects two compulsory complements. The first complement, which appears before DE, is a VP still expecting SUBJ and COMP, and the second complement, which appears after DE, is an NP. A special binary feature *CH* is defined at the SYNSEM level, to indicate the position of a COMP as relative to the HEAD: [*CH+*] for COMP before HEAD, and [*CH-*] for COMP after HEAD.

SYN.HEAD noun

[2] CH -
OPT -

Consistent with the definition of other complementizing DE, this DE also links its MOD value to the missing argument of the embedded clause, in this case, the subject, as the object

is already linked to the NP after DE. However, contrary to other complementizing DE, the deP formed using this special DE will not modify another NP, therefore feature [OPT+] is used to stop it. This lexical type works with the emphatic SHI entries to give the expected analysis for (35), as shown in (152) and (153).

- (35) 他是昨天 完成 的 任务
 tā shì zuótiān wánchéng de rènwù
 he is yesterday finish DE task

“He finished the task yesterday / It was yesterday that he finished the task”

Emphasizing the subject shi-subj-dep-lex is defined for SHI emphasizing the subject of a clause. It inherits from shi-emph-lex and further states that it takes no subject and expects two arguments which are both its complements. The first complement is an NP and it’s

co-indexed with the slashed argument of the second complement, deP (recall that the slashed argument is also co-indexed with the noun that deP modifies).

This leads to the analysis of (38) as shown in (155) and (156).

- (38) 是 他 完成 任务 的
 shì tā wánchéng rènwù de
 is he finish task DE
 “It was him who finished the task”

This implementation also works well when the emphasis of the subject and the alternative object position co-appear in the same sentence, as in (39). The results are displayed in (157) and (158).

- (39) 是他完成的任务
 shì tā wánchéng de rènwù
 is he finish DE task
 “It was him who finished the task”

When topicalization of the object also occurs in such sentences (40), the emphasis of the subject is weakened as topicalization also plays the role of shifting the focus to the constituent being topicalized. Therefore, we treat the deP in such sentence as the usual nominalization construction, giving the interpretation of “The task is what he finished”. (159) and (160) shows the results.

- (40) 任务 是 他 完成 的
 rènwù shì tā wánchéng de
 task is he finish DE

“It was him who finished the task”

Note: current definition of NON-LOCAL.SLASH as 0-1-dlist only allows SLASH to have one element at most. If we want the emphatic analysis for the above sentence, the embedded clause in DeP needs to keep track of two extracted arguments, both the subject and the object. We leave this for future work.

In summary, to capture the analysis of DE discussed in the above sections, four lexical entries of DE are created in Zhong’s lexicon, for the Associative (N) DE, the Associative (V) DE, the Relativizing DE, and the special DE for the alternative object position in emphatic SHI-DE constructions. In addition, the lexicon also contains an entry for DE as a sentence end particle.

5.6 DE as a Sentence Final Particle

For DE used as a sentence final particle, the existing lexical category shown in 161, *crs-lex-item* for sentence final particles is reused to define the lexical entry. Here DE is defined to be a marker that can take a complement which is headed by a fully saturated verb or an adjective (i.e. a sentence). The complement’s arguments have to be saturated, that is, it is expecting no subject or complement itself. The feature *R-PERIPH* is defined to be + to ensure it’s the right-most constituent in a sentence.

With this entry, the sentence in example (46), which is repeated here, can be processed by the grammar. The desired tree is produced, as shown in (162). Currently, DE does not have

any representation in the MRS graph (163). This is because information such as the affirmative tone or realis mood is not yet introduced into the grammar implementation. This will be part of the future work with proper modeling of the modalities in Mandarin Chinese, which should be done after a systematic study of relevant examples.

(46) 他 要 走 的

tā yào zǒu de

He will go DE-sfp

“He will go (with affirmative tone)”

(162)

(163)

5.7 Reduplicated Adjectives and Verbs

Based on our position that sentences with similar meaning should have similar semantic representations, we model the semantic representation of reduplicated verbs or adjectives as the predicate of the original word (A or AB) and a predicate that acts as an intensifier.³ Depending

³An earlier version of this analysis appears in Fan et al. (2015).

on the semantic function of the intensifier, it can be either an **amplifier** (making the meaning more intensified) or a **downtoner** (scaling it down), following the nomenclature of Quirk et al. (1985, p589 onwards).

Two predicates are therefore defined, *amplifier_x_rel* and *downtoner_x_rel*, both inheriting from a common parent *intensifier_x_rel*. *redup_up_x_rel* (representing amplification using reduplication) and *redup_down_x_rel* (representing scaling-down using reduplication) inherit from *amplifier_x_rel* and *downtoner_x_rel* respectively, as illustrated in (164). The predicate for the most common intensifier, the degree adverb 很 (*hen*, “very”), is also added into this structure, but more detailed differentiation of degree scales is left to the Chinese Open Wordnet (Wang & Bond, 2013).

We use lexical rules to produce the reduplicated forms from the original form. The super-type of the rules, *redup-type*, introduces the predicate *intensifier_x_rel*, as shown in (165).

Two lexical rules, *redup-a-lr* and *redup-v-lr*, inherit from *redup-type*. *redup-a-lr* (166), which is for adjective reduplication (AA and AABB), requires an adjective, and defines that the predicate introduced is the amplifier *redup_up_x_rel*. It also adds the syntactic constraint that the specifier of the word is empty, preventing it from accepting degree adverbs. The rule for the reduplication of verbs (AA and ABAB), *redup-v-lr* (167), requires a verb, defines the predicate *redup_down_x_rel*, and states that the verb doesn't accept aspect markers.

$$(166) \left[\begin{array}{ll} \text{redup-a-lr} \subset \text{redup-type} \\ \text{CAT.HEAD} & +a \text{ (adjective)} \\ \text{VAL} & \left[\text{SPR} \langle \rangle \right] \\ \text{C-CONT} & \left\langle \left[\text{PRED } \text{redup_up_x_rel} \right] \right\rangle \end{array} \right]$$

ORTHOGRAPHY: A → AA (irregular AB → AABB)

$$(167) \left[\begin{array}{ll} \text{redup-v-lr} \subset \text{redup-type} \\ \text{CAT.HEAD} & +v \text{ (verb)} \\ \text{CONT.HOOK} & \left[\text{ASPECT } \text{non-aspect} \right] \\ \text{C-CONT} & \left\langle \left[\text{PRED } \text{redup_down_x_rel} \right] \right\rangle \end{array} \right]$$

ORTHOGRAPHY: A → AA; A → A—A; (irregular AB → ABAB)

With the above definitions, for a sentence like (168), the dependency graph representing its MRS structure is provided in (169), which basically means “Something called “张三” is redup clean”.

- (168) 张三 干干净净
zhāngsān gāngānjìngjìng
Zhangsan AABB-clean
“Zhangsan is very clean”

(169)

If we generate from an MRS representation “Something called “张三” is amplifier clean”, we can get two possible surface forms:

- (170) a. 张三 很 干净
 zhāngsān hěn gānjìng
 Zhangsan very clean
 “Zhangsan is very clean”
- b. 张三 干干净净
 zhāngsān gāngānjìngjìng
 Zhangsan REDUP-clean
 “Zhangsan is very clean”

The above two lexical rules handle the $A \rightarrow AA$ reduplication for both verbs and adjectives. With pre-processing using regular expressions, another variation of the reduplication pattern of monosyllabic verbs, $A \rightarrow A - (yi \text{ “one”})A$, can also be handled by (167). An example of this pattern is given below in (171).

- (171) 看一看
 kànyīkàn
 look-one-look
 “take a look/look a little”

Since AABB reduplication of AB adjectives and ABAB reduplication of AB verbs are not very productive in Chinese (i.e., there are many AB adjectives or verbs that can not be reduplicated this way), we list them as irregular derivation forms (in `irregs.tab`). We have collected 92 entries for the AABB adjectives, and 74 entries for the ABAB verbs so far.

Another AB verb reduplication pattern is AB → AAB in (172), repeating the first character of some AB verbs. There is a similar pattern for some verbs with three characters. These verbs (so far 76) are also defined in `irregs.tab` to be handled in a similar manner.

(172) 说说话

shuōshuōhuà

AAB-talk

“have a talk/talk a little”

Other forms of AB verb reduplication, such as A_了 (*le*, “asp-marker”)A, and AA看 (*kàn* “see”), are left for future work.

ABB, shown in (173) and (174), is another commonly mentioned adjective reduplication pattern. Like other reduplicated words, it can’t be modified by degree adverbs. However, semantically it can’t be reduced to an A or AB predicate and a general reduplication predicate *redup-up-x-rel*. Either the AB form of the word doesn’t exist, or its A form exists but the different reduplication BB adds different meaning to the same A form. These adjectives are directly added into the lexicon (103 entries) with a lexical type defined with the required syntactic constraint for no intensifier.

(173) 绿油油

lǜyóuyóu

green-oil-oil

“bright green”

(174) 绿茸茸

lǜróngróng

green-downy-downy

“mossy green”

The semantic connection between (173) and (174), that they are more specific but slightly different kinds of green “bright green” and “mossy green”), will be captured in the Chinese Open Wordnet (Wang & Bond, 2013).

5.8 Interrogatives

We adopt directly the modeling and implementation of A-not-A questions detailed in Wang et al. (2015). Here I describe the analysis and modeling of content and particle questions.

As shown in Section 4.8, we need to take care of the constraint that the modal particle 吗 *ma* can not be attached to the end of certain questions like content questions. The information carried by question words like 谁 *shuí* “who” needs to be passed up the phrase structures containing a question word, so that 1) it can be recognized as a question; 2) 吗 *ma* can be forbidden to combine with such phrases, but other particles will be allowed; 3) it can be an embedded question in a matrix sentence, which is not a question itself, like “I don’t know what he wants”.

In Zhong, the non-local feature `NON-LOCAL.QUE` is used to signal that there is a wh-word in a clause. Non-local features are passed up by every grammar rule: the `NON-LOCAL.QUE` will be the concatenation of the `NON-LOCAL.QUE`s of all its daughters, similarly to the `GAP` feature in Sag et al. (2003, Chapter 14. *Long-Distance Dependencies*). For 1), a unary rule pushes a clause with non-empty `QUE` to become a question, with the value of `LOCAL.CONT.HOOK.INDEX.SF` set to be *ques*. For 2), 吗 *ma* is defined to expect its complement to have empty `QUE`. For 3) verb types that can take such embedded questions as complement are defined to expect clausal complement with non-empty `QUE`. The `SF` for such matrix clause should then be *prop*, not *ques*.

5.8.1 Content questions

To indicate that it is a wh-word, in the lexical type for question words like 谁 *shuí* “who”, *wh-pronoun*, we define that its `NON-LOCAL.QUE` has a value that’s co-indexed with the `INDEX` of the word’s predicate. This feature is copied up the phrase structures by Zhong’s rules. Finally a

unary rule, `question-phrase`, detects the non-empty value of `QUE` and sets `SF` to be *ques*.

$$(175) \left[\begin{array}{l} \text{wh-pronoun} \\ \text{SEM.HOOK.INDEX } i \\ \text{NON-LOCAL.QUE } \langle i \rangle \end{array} \right]$$

For the question in (55), the grammar produces the tree as in (176). The MRS representation, shown in (177), displays that the `SF` value of the main predicate is *ques*, indicating it's a question. The more intuitive DMRS representation usually doesn't show extra information like `SF` value. For the purpose of illustration, this information is added as a footnote to the main predicate in the output, as in (178). This representation will be used for the subsequent examples.

(55) 谁 哭 了 ?
 shuí kū le ?
 Who cry ASP ?
 “Who cried ?”

(177)

$$\left[\begin{array}{l}
\text{mrs} \\
\text{TOP} \quad \boxed{0} h \\
\text{INDEX} \quad \boxed{2} \left[\begin{array}{l} e \\ \text{SF} \quad \text{ques} \\ \text{E.ASPECT} \quad \text{perfective} \end{array} \right] \\
\text{RELS} \quad \left\langle \begin{array}{l} \left[\begin{array}{l} \text{person_n_rel} \\ \text{LBL} \quad \boxed{4} h \\ \text{ARG0} \quad \boxed{3} x \end{array} \right], \left[\begin{array}{l} \text{which_q_rel} \\ \text{LBL} \quad \boxed{5} h \\ \text{ARG0} \quad \boxed{3} x \\ \text{RSTR} \quad \boxed{6} h \\ \text{BODY} \quad \boxed{7} h \end{array} \right], \left[\begin{array}{l} \text{哭_v_1_rel} \\ \text{LBL} \quad \boxed{1} h \\ \text{ARG0} \quad \boxed{2} e \\ \text{ARG1} \quad \boxed{3} x \end{array} \right] \end{array} \right\rangle \\
\text{HCONS} \quad \left\langle \begin{array}{l} \left[\begin{array}{l} \text{qeq} \\ \text{HARG} \quad \boxed{0} h \\ \text{LARG} \quad \boxed{1} h \end{array} \right], \left[\begin{array}{l} \text{qeq} \\ \text{HARG} \quad \boxed{6} h \\ \text{LARG} \quad \boxed{4} h \end{array} \right] \end{array} \right\rangle \\
\text{ICONS} \quad \left\langle \begin{array}{l} \left[\begin{array}{l} \text{semantic-focus} \\ \text{IARG1} \quad \boxed{2} e \\ \text{IARG2} \quad \boxed{3} x \end{array} \right] \end{array} \right\rangle
\end{array} \right]$$

5.8.2 Particle questions

Particle questions are formed by attaching a question particle like 吗 *ma* to the end of a declarative sentence. The lexical type `qpart-lex-item` (179) defines the features for question particles. It's defined to be a marker that doesn't modify anything else. Constraints are set

carefully to make sure it only combines with the right constituent. It takes no specifier, no subject, and just expects one complement. This complement can be a verb, adjective or PP, with its SPR, SUBJ and COMPS all saturated. To prevent attaching MA to content questions, the NON-LOCAL.QUE of the complement must be empty, indicating it's not already a question.

As attaching the question particle is usually done at the last step of forming a sentential phrase, it doesn't need to pass QUE information further up, but rather discharges it: setting its NON-LOCAL.QUE is empty and the value of its HOOK. INDEX. SF to be *ques*.

With the above implementation, the question in (58) is parsed to give the following results in (180) and (181).

(58) 你 去 吗 ？

nǐ qù ma ？

You go Q-ma ？

“Will you go?”

5.9 Other Enhancements of the Grammar

We have been enhancing the grammar with the objective to achieve coherent and consistent semantics constrained by syntax. Using the sentences from the MRS testsuite, and supplemented by sentences collected from relevant literature and real corpus, we have improved the grammar on its handling of the known structures in the MRS testsuite, such as BA and BEI structures, NP structures, argument structures, classifiers, etc. At the same time, we have also created analyses to cover linguistically interesting phenomena new to the MRS testsuite, including reduplication of adjectives, resultative VV compounds, A-not-A questions, as well as the handling of particles, interjections, and fragments. Our work is summarized in Table 5.1.

After the implementation of the above work and other small changes, we have conducted a coverage test of the grammar using test sentences from another corpus, the Chinese corpus

Table 5.1: Grammar enhancement

	what's improved	what's added	what's planned
grammar	topic-comment clefts BA and BEI NP structures classifiers argument structure adpositions	relative clauses nominalization reduplication VV compounds particles fragments interjections A-not-A questions honorification	serial verbs conjunctions
engineering	generation	unknown word handling bridging rules test modules full-forest treebanking	Wordnet incorporation transfer rules machine translation

from TUFs Open Language Resources (Kawaguchi, 2007) from Tokyo University of Foreign Studies). It contains 1523 sentences with basic vocabulary used to teach Chinese. The coverage test shows that our grammar can parse 651 or 42.74% of these unseen sentences.

Chapter 6

Grammar Engineering

6.1 Preprocessing and Postprocessing

Zhong includes an unknown word handling module based on the chart-mapping technique of Adolphs et al. (2008). We have built a pipeline for converting raw text into a segmented POS-based lattice for input to the parser. The preprocessing stage for handling unknown words runs with the Stanford tools including the Chinese word segmenter (Tseng et al., 2005) and the Chinese Part-Of-Speech tagger (Toutanova et al., 2003). There are multiple different standards for segmenting the input string in Chinese, viz. Chinese Penn Treebank and Peking University. Between them, we are using the former because our fundamental development corpus NTU-MC (Tan & Bond, 2012) was segmented using that standard. We implemented a wrapper to run these tools in the pipeline using NLTK (Bird, 2006). In addition, the pre-processor includes some generic lexical entry rules for handling particular string patterns, such as numbers, dates, currency, emails, urls, etc. These lattice-based mapping rules work with a set of regular expressions. Building upon these two facilities, many lexical items not registered in the dictionary can be automatically identified and efficiently processed.

For postprocessing, we implemented a proof-of-concept monolingual transfer grammar for paraphrasing simplified Mandarin Chinese. It converts MRS outputs in the parse results into more generic or more specific ones. Currently, this postprocessor works for generating intensifying constructions and classifier constructions.

6.2 Configuration

Zhong has been built up following the premise “parse robustly and generate strictly” (Bond et al., 2008). This means that even a rather infelicitous sentence should be parsed, but the infelicitous sentences should be filtered out in generation. This different approach to parsing and generation can be facilitated using different configurations for compiling the grammars. First, Zhong includes a flag feature [STYLE style] for marking the felicity of particular lexical items and constructions, whose subtypes are `strict`, `robust`, `unproductive`, etc. Second, there are different types of roots: namely, `roots.tdl`, `roots-robust.tdl`, and `roots-strict.tdl`. The first one works for ordinary parsing and generation, the second one works with bridging rules to fill out the chasm between constructions, and the third one is particularly used for generation with the [STYLE `strict`] flag. Third, there are different scripts to load and compile the grammars within LKB and ACE, such as `config.tdl`, `config-robust.tdl`, and `config-strict.tdl`. The last one includes the list of items and rules that should be ignored in generation (`generation.ignore`).

For example, 去 着 ‘go DUR’ may not sound good to Chinese native speakers, because the verb 去 tends not to co-occur with the durative aspect. Our grammar provides a parse tree for the sentence with a flag [STYLE `robust`] but does not generate such a sentence. To take another example, the punctuation markers are optional in the ordinary and robust processing but obligatorily appear in the generation output produced by the grammar compiled by `config-strict.tdl`.

6.3 Treebanking

Treebanks refer to corpora of a natural language annotated with rich linguistic information. They are widely used in linguistic research, providing a rich resource for the investigation of language phenonema. They are also considered highly valuable in the field of natural language processing, serving as training data for the construction of various statistical models. Sentences in treebanks can be annotated manually by linguists, automatically using a parser, or in a hybrid manner combining both ways.

The HPSG community has been building treebanks for languages for which HPSG grammars have been built, both as a test ground to see if the grammar developed is wide and robust enough to cover language phenomena naturally existing in corpora, and as training data to build statistical ranking models that can be used to rank the possible parses produced by a parser, pushing the more likely (to be correct) parses to the top of the view.

For Zhong, the treebank is prepared based on a Chinese corpus called CMNEDU (developed as part of the *Syntactic Well-Formedness Diagnosis and Error-Based Coaching in Computer Assisted Language Learning using Machine Translation Technology* project, PI: Francis Bond, MOE TRF). This corpus contains 798 sentences collected from Chinese textbooks used in college classes teaching Chinese as a second language. It is selected so that we can apply Zhong in computer-aided language learning, to build an application that can check the well-formedness of sentences created by students who are learning Chinese as a second language. These sentences represent the kind of grammatical sentences that the students should grasp at their level of learning.

A testsuite is built for the sentences in CMNEDU. Using the full-forest treebanking tool, (FFTB: Packard, 2015) and ACE, the sentences in the testsuite are parsed with Zhong. The grammar may or may not produce possible parses for each sentence. With the current version of Zhong, 743 out of 798, or 93.11% of the sentences get at least one parse from the grammar. These parses are stored in the database and reviewed manually. For 689 out of 743 sentences, their gold (preferred) trees are manually selected from the results, whereas the results of 54 sentences are rejected.

6.4 Ranking

With the CMNEDU treebank prepared (as described in the section above), using FFTrain (the model builder from FFTB), a statistical ranking model, `cmnedu.mem` can be made, using the techniques of Toutanova et al. (2005). This model can be used when parsing sentences with the grammar to score the trees and rank them accordingly.

When calling ACE to parse sentences, option “`--maxent cmnedu.mem`” can be used to

activate the usage of the trained ranking model. The model will change the ordering of the possible trees produced and push the better trees to the top.

For example, for the sentences shown in (182, 183) shows the first three trees in the results without using the ranking model. We can see the first tree is unlikely as it doesn't combine “点儿”(a bit) and “东西”(thing) into a NP first. The second tree is the most plausible interpretation.

- (182) 我 刚才 去 邮局 给 我 妈妈 寄 了 点儿 东西
 wǒ gāngcái qù yóujú gěi wǒ māmā jì le diǎnér dōngxī
 I just-now go post-office to I mother send ASP a-bit thing
 “Just now I went to the post office to send my mother something”

(183)

In the results obtained using the parser with the ranking model, as shown in (184) we can see that in the first three trees, the correct tree is now at the top.

Due to the limited size of the current treebank, and the time constraints, evaluation of the ranking model has not been done on a reserved set of sentences to see if it can always put the correct tree at the top. However, during the grammar development and debugging, the use of the ranking model has been found to improve the efficiency of work as the correct tree is much more likely to be at the top of the results and therefore easier to identify and check.

6.5 MAL-Extension

In computer-aided second language learning, a system can help a beginning learner of Chinese substantially by giving precise, constructive feedbacks on the grammatical errors in his/her sentences. This can be enabled by extending our grammar with specifications of common grammatical errors that beginners tend to make, in the form of "mal-rules" (Schneider & McCoy, 1998; Bender et al., 2004; Suppes et al., 2014).

For example, for the error shown in (3), the wrong usage of NP coordinator 和 (hé "and") to connect clauses, Zhong normally wouldn't parse this sentence as the constraints defined for 和 (hé "and") wouldn't allow so (it only coordinates NPs). In the MAL-extension, a lexical entry ("mal-entry") tolerating this wrong usage is created, as shown in (185).

(3)

*他 不 学 中文, 和 不 学 法文。
tā bù xué zhōngwén, hé bù xué fǎwén。
He not learn Chinese, and not learn French.

"He learns neither Chinese nor French"

(185)

```
和_c_△ := conj_-_e_le &
[ STEM < "和" >,
  SYNSEM.LOCAL.CAT.HEAD.CHAR [ FCHAR "和", LENGTH one ],
  SYNSEM.LKEYS.KEYREL.PRED "_和_c_△_rel",
  TRAITS native_token_list ].
```

In (185), 和 (hé "and") is defined to be a conjunction for clauses, through the usage of an existing type `conj_-_e_le`, which constraints its two constituents to be coordinated as events. The special symbol "△" in the lexical label allows detection of such "mal-entities" in the results.

The resulting tree from the extended Zhong for (3) is shown in Figure 6.1

Figure 6.1: Tree for example (3).

For the other common error, the wrong position of adverb 也 (“also”), as shown in (4), the similar approach is taken.

(4)

*我 学 工程, 也 我 学 中文。

wǒ xué gōngchéng, yě wǒ xué zhōngwén。

I study engineering, also I study Chinese.

“I study engineering, and I also study Chinese.”

For this error, another “mal-entry” (186) is defined. Here this adverb is defined to be a scoping modifier, with an existing type `adv__scop_le`. In Zhong, a scoping adverb can appear before the subject of the clause it’s attached to, thus allowing this sentence to be parsed.

(186)

```
也_r_Δ := adv__scop_le &
[ STEM < "也" >,
  SYNSEM.LOCAL.CAT.HEAD.CHAR [ FCHAR "也", LENGTH one ],
  SYNSEM.LKEYS.KEYREL.PRED "_也_a_Δ_rel",
  TRAITS native_token_list ].
```


The tree obtained for (4) is shown in Figure 6.2. Similarly, the tolerated error is visible in the tree with the special symbol.

Figure 6.2: Tree for example (4).

Chapter 7

Conclusion

In the above chapters I have described the main work done in my PhD study. Zhong has been enhanced in terms of both its lexicon coverage and the modeling of linguistic phenomena and tested on a corpus of learners' text.

The lexicon was built semi-automatically by learning the lexical entries from an annotated corpus and mapping the original lexical tags to the detailed lexical types defined in Zhong. The resulting lexicon was found to improve the grammar's coverage on a reserved test set significantly.

The syntax coverage of Zhong has been enhanced, building on the foundational work of ManGO. Recognizing that 的 DE is one of the most commonly used particle in modern Chinese, I focused on handling various structures involving DE. Analysis for associative DE, relativizing DE, nominalizing DE, SUO-DE, SHI-DE, etc., as well as that for reduplication and interrogatives has been presented in the previous chapters. The results of the implementation produced, including the parse tree and DMRS representations, have been presented to show its effectiveness. Other smaller modifications and extensions of the grammar during development and debugging have been omitted from this thesis, but can be seen in the grammar itself.

A small extension of the grammar to cover malformed sentences was also written, with the aim to begin to cover some common learners' errors when writing Chinese. A few MAL-types and MAL-rules were created so that sentences containing the corresponding errors can be parsed with the right label to trigger subsequent handling in applications like a Grammar Error

Feedback System.

Other grammar engineering extensions were summarized in Chapter 6. In particular, a treebank has been built by parsing a collection of sentences from Chinese textbooks with Zhong and manually selecting the best tree from the output. The grammar was found to be able to parse 84.21% of sentences in the treebank. A statistical ranking model was then trained from the treebank. The model, when used during parsing, can score and rank the output trees from the grammar to present the more likely trees at the top. This turned out to be very helpful for checking and debugging on the grammar's output during development.

For future work, Zhong will continue to be enhanced. The size of the lexicon is still rather small as only a sample of Sinica corpus was freely available in the NLTK package. The same process can be repeated on the main Sinica corpus to obtain more high quality lexical entries. The LDC Chinese Treebanks can also be explored. Though its lexical tags do not contain as much syntactic information as Sinica corpus provides, the structural similarity of the syntactic trees available in the corpus can be explored to derive words that are likely to belong to the same lexical type or sub-types.

The grammar also continues to require further development in terms of phenomena coverage, for example, serial verb constructions, various conjunctions, etc. The existing implementations of the phenomena already investigated also need to be tested further on more sentences from real data, as interactions of phenomena may lead to unexpected results.

As the coverage of Zhong continues to grow, the treebank should be expanded, including more sentences from corpora of various domains. Further, a test set should be prepared and reserved for proper testing of the coverage of the grammar. Then we can systematically track the change of the coverage performance of the grammar as the development continues. With the availability of more sentences with the best parse identified, a better statistical ranking model can be trained. We can also use the test set to track the ranking model's performance using the percentage of correct trees ranked as top parse.

Finally, as Zhong will be used in grammar feedback system, the MAL-extension of the grammar will continue to be developed as the *Syntactic Well-Formedness Diagnosis and Error-Based Coaching in Computer Assisted Language Learning using Machine Translation Tech-*

nology project progresses.

Bibliography

- Adolphs, Peter, Stephan Oepen, Ulrich Callmeier, Berthold Crysmann, Dan Flickinger & Bernd Kiefer. 2008. Some Fine Points of Hybrid Natural Language Parsing. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis & Daniel Tapias (eds.), *Proceedings of the sixth international language resources and evaluation (LREC-2008)*, 1380–1387. Marrakech, Morocco.
- Avgustinova, Tania & Yi Zhang. 2009. Parallel Grammar Engineering for Slavic Languages. In *Workshop on grammar engineering across frameworks at the acl/ijcnlp*, .
- Baldwin, Timothy. 2005. General-purpose lexical acquisition: Procedures, questions and results. In *Proc. of the 6th meeting of the pacific association for computational linguistics (pacling 2005)*, 23–32. Citeseer.
- Baldwin, Timothy & Francis Bond. 2003. Learning the countability of English nouns from corpus data. In *41st annual meeting of the association for computational linguistics: Acl-2003*, 463–470. Sapporo, Japan.
- Bender, Emily. 2000. The syntax of mandarin bă: Reconsidering the verbal analysis. *Journal of East Asian Linguistics* 9(2). 105–145.
- Bender, Emily M., Scott Drellishak, Antske Fokkens, Laurie Poulson & Safiyyah Saleem. 2010. Grammar Customization. *Research on Language & Computation* 8(1). 23–72.
- Bender, Emily M., Dan Flickinger & Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the workshop on grammar engineering and evaluation at the 19th international conference on computational linguistics*, 8–14. Taipei, Taiwan.
- Bender, Emily M, Dan Flickinger & Stephan Oepen. 2008. Grammar engineering for linguistic hypothesis testing. In *Proceedings of the texas linguistics society x conference: Computational linguistics for less-studied languages*, 16–36. Citeseer.

- Bender, Emily M, Dan Flickinger, Stephan Oepen, Annemarie Walsh & Timothy Baldwin. 2004. Arboretum: Using a precision grammar for grammar checking in call. In *Instil/ical symposium 2004*, .
- Bird, Stephen, Ewan Klein & Edward Loper. 2009. *Natural language processing with python*. O'Reilly. (www.nltk.org/book).
- Bird, Steven. 2006. NLTK: the Natural Language Toolkit. In *Proceedings of the coling/acl on interactive presentation sessions*, 69–72.
- Boehm, Barry W. 1988. A spiral model of software development and enhancement. *Computer* 21(5). 61–72.
- Bond, Francis, Eric Nichols, Darren Scott Appling & Michael Paul. 2008. Improving Statistical Machine Translation by Paraphrasing the Training Data. In *Proceedings of the international workshop on spoken language translation*, 150–157. Hawaii.
- Chang, Chao-Huang & Gilbert K Krulee. 1991. Resolution of ambiguity in chinese and its application to machine translation. *Machine translation* 6(4). 279–315.
- Chang, Li-ping & Keh-jian Chen. 1995. The ckip part-of-speech tagging system for modern chinese texts. In *Proceedings of 1995 international conference on computer processing of oriental languages*, 172–175.
- Chen, Feng-yi, Ruo-ping Jean Mo, Chu-Ren Huang & Keh-Jian Chen. 1992. Reduplication in mandarin chinese: Their formation rules, syntactic behavior and icg representation. In *Proceedings of rocling v computational linguistics conference v*, 217–233. The Association for Computational Linguistics and Chinese Language Processing.
- Copestake, Ann. 2002. *Implementing typed feature structure grammars*. CSLI Publications.
- Copestake, Ann, Dan Flickinger, Carl Pollard & Ivan A. Sag. 2005. Minimal Recursion Semantics. An introduction. *Research on Language and Computation* 3(4). 281–332.
- Crysmann, Berthold. 2005. Relative clause extraposition in German: An efficient and portable implementation. *Research on Language and Computation* 3(1). 61–82.

- Daudé, Jordi, Lluís Padro & German Rigau. 2000. Mapping wordnets using structural information. In *Proceedings of the 38th annual meeting on association for computational linguistics*, 504–511. Association for Computational Linguistics.
- Deng, Dun. 2009. xiàndài hànyǔ suǒ jí suǒ zì jiégòu de chóngxīn shěnsì yǔ dìngxìng 《现代汉语“所”及“所”字结构的重新审视与定性》 a new analysis and definition of suo and the suo construction in modern chinese. *Hànyǔ xuéxí* 《汉语学习》 (*Chinese Language Learning*) 2. 106–112.
- Fan, Zhenzhen, Sanghoun Song & Francis Bond. 2015. Building Zhong [—], a Chinese HPSG meta-grammar. In *Proceedings of the 22nd international conference on Head-Driven Phrase Structure Grammar (HPSG 2015)*, .
- Flickinger, Dan. 2000a. On building a more efficient grammar by exploiting types. *Natural Language Engineering* 6(1). 15–28. (Special Issue on Efficient Processing with HPSG).
- Flickinger, Dan. 2000b. On Building a More Efficient Grammar by Exploiting Types. *Natural Language Engineering* 6(1). 15–28.
- Fokkens, Antske & Tania Avgustinova. 2013. SlaviCLIMB: Combining Expertise for Slavic Grammar Development using a Metagrammar. In *Workshop on high-level methodologies for grammar engineering*, 87–92.
- Fokkens, Antske, Tania Avgustinova & Yi Zhang. 2012. CLIMB Grammars: Three Projects using Metagrammar Engineering. In *Proceedings of the eight international conference on language resources and evaluation*, 1672–1679. Istanbul.
- Fujita, Sanae & Francis Bond. 2007. A method of creating new valency entries. *Machine Translation* 21. 1–28. URL <http://dx.doi.org/10.1007/s10590-008-9032-7>.
- Gao, Qian. 2000. *Argument Structure, HPSG, and Chinese Grammar*: Ohio State University dissertation.

- Hsu, Hui-Li & Chu-Ren Huang. 1995. Design criteria for a balanced modern chinese corpus. *Proceedings of ICCPOL, Honolulu, Hawaii* .
- Huang, Chu-Ren & Keh-Jiann Chen. 1998. *Technical Report of Academia Sinica Balanced Corpus of Modern Chinese*. Tech. rep. Academia Sinica. URL <http://db1x.sinica.edu.tw/kiwi/mkiwi/98-04.pdf>.
- Huang, Chu-Ren & Dingxu Shi. 2016. *A reference grammar of chinese*. Cambridge University Press.
- Im Walde, Sabine Schulte. 2006. Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics* 32(2). 159–194.
- Kawaguchi, Yuji. 2007. Foundations of center of usage-based linguistic informatics (ubli). *Corpus-Based Perspectives in Linguistics* 3–28.
- Kiefer, Bernd, Hans-Ulrich Krieger & Mark-Jan Nederhof. 2000. Efficient and robust parsing of word hypotheses graphs. In *Verbmobil: Foundations of speech-to-speech translation*, 280–295. Springer.
- Korhonen, Anna. 2002. Semantically motivated subcategorization acquisition. In *Proceedings of the acl workshop on unsupervised lexical acquisition*, Philadelphia, USA.
- Li, Charles N & Sandra A Thompson. 1989. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press.
- Li, Dejin & Meizhen Cheng. 1994. *A practical chinese grammar for foreigners*. China Books & Periodicals.
- Lu, Jianming & Zhen Ma. 1985. dé zì jiégòu hé suǒ zì jiégòu 《“的”字结构和“所”字结构》 [de construction and suo construction]. In *xiàndài hànyǔ xūcí sānlùn* 《现代汉语虚词散论》 (*a collection of articles on functional words in modern chinese*), 231–248. Beijing: Beijing University Press.
- Lǚ, Shuxiang (ed.). 1999. *xiàndài hànyǔ bābǎi cí* 《现代汉语八百词》 (*eight hundred words of modern chinese*). Beijing: The Commercial Press Ltd.

- Müller, Stefan & Walter Kasper. 2000. HPSG analysis of german. In Wolfgang Wahlster (ed.), *Verbmobil: Foundations of speech-to-speech translation*, 238–253. Berlin, Germany: Springer.
- Müller, Stefan. 2013a. The CoreGram project: A brief overview and motivation. In Denys Duchier & Yannick Parmentier (eds.), *Proceedings of the workshop on high-level methodologies for grammar engineering (hmge 2013), düsseldorf*, 93–104. URL <http://hpsg.fu-berlin.de/~stefan/Pub/coregram-brief.html>.
- Müller, Stefan. 2013b. The CoreGram Project: Theoretical Linguistics, Theory Development and Verification. *Ms. Freie Universität Berlin*.
- Müller, Stefan. 2015. The CoreGram project: Theoretical linguistics, theory development and verification. *Journal of Language Modelling* 3(1). 21–86. URL <http://hpsg.fu-berlin.de/~stefan/Pub/coregram.html>.
- Müller, Stefan & Janna Lipenkova. 2009. Serial verb constructions in chinese: An hpsg account. In *Proceedings of the 16th international conference on head-driven phrase structure grammar*, 234–254.
- Ng, Say Kiat. 1997. *A double-specifier account of Chinese NPs using head-driven phrase structure grammar*. Department of Linguistics, University of Edinburgh MSc Thesis.
- Oepen, Stephan, Dan Flickinger, Jun-ichi Tsujii & Hans Uszkoreit (eds.). 2002. *Collaborative language engineering*. Stanford: CSLI Publications.
- Packard, Woodley. 2015. *Full forest treebanking*. University of Washington MA thesis. URL <https://digital.lib.washington.edu/researchworks/handle/1773/33194>.
- Paris, Marie-Claude. 1998. Focus operators and types of predication in mandarin. *Cahiers de linguistique-Asie orientale* 27(2). 139–159.
- Paul, Waltraud & John Whitman. 2008. Shi... de focus clefts in mandarin chinese. *The Linguistic Review* 25(3-4). 413–451.

- Pollard, Carl & Ivan A. Sag. 1994. *Head driven phrase structure grammar*. Chicago: University of Chicago Press.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the english language*. London: Longman.
- Sag, Ivan A., Tom Wasow & Emily Bender. 2003. *Syntactic theory: A formal introduction*. Stanford: CSLI Publications 2nd edn.
- Sanfilippo, Antonio & Victor Poznański. 1992. The acquisition of lexical knowledge from combined machine-readable dictionary sources. In *Proceedings of the third conference on applied natural language processing*, 80–87. Association for Computational Linguistics.
- Schneider, David & Kathleen F. McCoy. 1998. Recognizing syntactic errors in the writing of second language learners. In *36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics: Coling/acl-98*, 1198–1204. doi: 10.3115/980432.980765. URL <http://dx.doi.org/10.3115/980432.980765>.
- Siegel, Melanie & Emily M. Bender. 2002. Efficient deep processing of Japanese. In *Proceedings of the 3rd workshop on asian language resources and international standardization at the 19th international conference on computational linguistics*, 1–8. Taipei.
- Song, Sanghoun. 2017. *Modeling information structure in a cross-linguistic perspective*. Berlin: Language Science Press.
- Suppes, Pat, T. Liang, E. E. Macken & Dan Flickinger. 2014. Positive technological and negative pre-test-score effects in a four-year assessment of low socioeconomic status K-8 student learning in computer-based math and language arts courses. *Computers & Education* 71. 23–32.
- Tan, Liling & Francis Bond. 2012. Building and Annotating the Linguistically Diverse NTU-MC (NTU-Multilingual Corpus). *International Journal of Asian Language Processing* 22(4). 161–174.

- Ting, Jen. 2003. The nature of the particle suo in mandarin chinese. *Journal of East Asian Linguistics* 12. 121–139.
- Toutanova, Kristina, Dan Klein, Christopher D Manning & Yoram Singer. 2003. Feature-rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology-volume 1*, 173–180.
- Toutanova, Kristina, Christopher D. Manning, Dan Flickinger & Stephan Oepen. 2005. Stochastic HPSG parse disambiguation using the redwoods corpus. *Research on Language and Computation* 3(1). 83–105.
- Tseng, Huihsin, Pichuan Chang, Galen Andrew, Daniel Jurafsky & Christopher Manning. 2005. A Conditional Random Field Word Segmenter. In *Proceedings of the fourth sishan workshop on chinese language processing*, 169–171.
- Uszkoreit, Hans, Dan Flickinger, Walter Kasper & Ivan A Sag. 2000. Deep linguistic analysis with hpsg. In *VerbMobil: Foundations of speech-to-speech translation*, 216–237. Springer.
- Wang, Li. 1959. "zhòngguó xiàndài yǔfá 《中国现代语法》 modern chinese grammar", vol. 1. "zhōnghuá shūjú 中局".
- Wang, Shan & Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th workshop on asian language resources, a workshop at ijcnlp-2013*, 10–18. Nagoya.
- Wang, Wenjie, Sanghoun Song & Francis Bond. 2015. A constraint-based analysis of A-NOT-A questions in Mandarin Chinese. In *Proceedings of the 22nd international conference on Head-Driven Phrase Structure Grammar (HPSG 2015)*, .
- Wang, Xiangli, Shunya Iwasawa, Yusuke Miyao, Takuya Matsuzaki & Jun'ichi Tsujii. 2009. The design of chinese hpsg for data-oriented parsing. In *言语处理学会第15回年次大会*, 20–23.

- Xia, Fei. 2000. The part-of-speech tagging guidelines for the penn chinese treebank (3.0). *IRCS Technical Reports Series* 38.
- Xu, Renjie, Zhiqiang Gao, Yuzhong Qu & Zhisheng Huang. 2008. An integrated approach for automatic construction of bilingual Chinese-English WordNet. In *3rd asian semantic web conference (aswc 2008)*, 302–341.
- Xue, Naiwen, Fei Xia, Fu-Dong Chiou & Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering* 11(02). 207–238.
- Xue, Nianwen, Zixin Jiang, Xiuhong Zhong, Martha Palmer, Fei Xia, Fu-Dong Chiou & Meiyu Chang. 2010. Chinese treebank 7.0. *Linguistic Data Consortium, Philadelphia*.
- Yang, Justin Chunlei. 2011. Mandarin grammar on-line: ManGO. Presentation at the 2011 DELPH-IN Summit. URL <http://www.delph-in.net/2011/mango.pdf>.
- Yu, Kun, Yusuke Miyao, Xiangli Wang, Takuya Matsuzaki & Junichi Tsujii. 2010. Semi-automatically developing chinese hpsg grammar from the penn chinese treebank for deep parsing. In *Proceedings of the 23rd international conference on computational linguistics: Posters*, 1417–1425. Association for Computational Linguistics.
- Zhang, Yi. 2004. Starting to implement chinese resource grammar using lkb and lingo grammar matrix. Tech. rep. Technical report.
- Zhu, Dexi. 1982. "yǔfá jiǎngyì 《语法讲义》 *lecture notes on grammar*", vol. 1. "shāngwù yìnshūguǎn 商务印书馆局".