Engineering a Deep HPSG for Mandarin Chinese

Yi Zhang **Rui Wang** **Yu Chen**
LT-Lab, German Reserach Center for Artificial Intelligence, Saarbrücken, Germany
{yzhang,rwang}@coli.uni-sb.de, yuchen@dfki.de

Abstract

In this paper, we present our on-going grammar development effort towards a linguistically precise and broad coverage grammar for Mandarin Chinese in the framework of HPSG. The use of LinGO Grammar Matrix facilitates the quick start of the development. We propose a series of linguistic treatments for a list of interesting phenomena. The analyses are largely compatible with the HPSG framework. In addition, the grammar also composes semantic representations in Minimum Recursion Semantics. Preliminary tests of the grammar on a phenomenon-oriented test suite show encouraging precision and coverage.

1 Introduction

Broad coverage in-depth and accurate linguistic processing is desirable for both linguistic studies and practical NLP applications. In recent years, several competing linguistic frameworks emerge with proper expressive power and good computational properties. Typically offered by such frameworks are not only the description of the syntactic structures, but also the ways in which meanings are composed. Among the most popular frameworks are CCG, TAG, LFG and HPSG.

With the increasing availability of deep linguistic processing platforms, large-scale grammar resource development becomes possible. The past experience on large-scale grammar engineering shows that it is a long-term undertaking, which amounts to years or decades of both labor- and intelligence-intensive work. More recently, it has been shown that such process could be largely accelerated by the accumulative experience from various grammar development projects. Also, the data-driven techniques reduce the tedious repetitive work and allow grammar writers to focus on the challenging phenomena.

Encouraged by these breakthroughs, we have seen the emergence of various grammar development projects in the last decade, not only for languages with large speaker population, but also for endangered or extinct languages (Bender, 2008). Despite the huge population of Mandarin Chinese native speakers, strikingly few attempts have been made so far to formally describe the language within the above-mentioned modern linguistic frameworks. This is partially due to the fact that Mandarin Chinese is relatively less grammaticalized in the sense that the wellformedness of a sentence cannot be clearly judged from the syntactic structure alone. But given that some modern frameworks (such as HPSG) integrates the syntactic and semantic representations, a joint analysis is feasible.

Another trendy approach in deep grammar engineering is the corpus-driven approach. For instance, Miyao et al. (2004) showed that by enriching the PTB annotation with HPSG feature structures and applying top-down unifications, one can automatically acquire detailed lexical templates. The similar procedure was practiced by Hockenmaier and Steedman (2005) (though in a different framework) in the creation of the CCGbank. Recently, some of these success stories have been transferred to the development of Mandarin Chinese grammars on the Penn Chinese Treebank (CTB; (Xue et al., 2005)). Nevertheless, we believe that the corpus-driven approach does not replace the need for a carefully engineered core grammar, with which the basic linguistic generalizations could be captured and consistently applied to various instantiations in the corpus. Thus, we believe that a hand-written grammar will constitute the very foundation of the deep linguistic processing.

In this paper, we report on the on-going development of a Mandarin Chinese grammar (MCG) in the framework of HPSG. With the modern grammar engineering setup, we were able to cover a

Proceedings of the 9th Workshop on Asian Language Resources, pages 18–22,
Chiang Mai, Thailand, November 12 and 13, 2011.

large number of interesting phenomena with satisfactory accuracy from both syntactic and semantic points of view. The evaluation of the grammar resource is an important aspect of the development. At the current stage, we value the correct choice of linguistic solutions to be more important than the less meaningful parsing coverage on arbitrary "gold standard" annotation. For this reason, we choose to test the core-grammar on a phenomenon-oriented test suite instead of a corpus of naturally occurring texts.

2 An HPSG Analysis of Mandarin

2.1 Design of sign & schemata

The design of the HPSG sign in MCG is compatible with the design in the LinGO Grammar Matrix. Four valcence features were employed: SUBJ for subjects, COMPS for complements, SPR for specifiers, and SPEC for back-reference from the specifier to its head. Unlike Yu et al. (2010) who separate complement list into LCOMPS and RCOMPS, we keep all complements on the same complement list (COMPS), and use an additional boolean feature $\left[\text{RC} \quad \pm\right]$ to indicate whether the complement is to the right or to the left of the head.

The grammar currently contains about 20 rule schemata. It should be noted that most of these rule schemata are very general. They are be used to handle multiple types of constructions, some of which will be illustrated below.

2.2 HEAD types

The HEAD types in HPSG identify the major categories of part-of-speech for the language. The structure of MCG's HEAD type hierarchy is show in Figure 1. Worth noticing is that we have adjectives being a sub-type of predicative, so it can serve as the predicate of a sentence (similar to verb) without *"type-raising"*. A special category *coverb* is designed to cover words which share certain properties of verbs, but usually do not serve as the main predicate of a sentence, such as prepositions (在,用), BA (把), BEI (被), and resultative coverbs (e.g. 来, 开).

2.3 Topic Construction

According to Li and Thompson (1989), a topic of a sentence refers to the theme of the sentence and appears before the subject. For a better account of the semantics, we further distinguish the following types of topics and treat them separately with different schemata.

- When the sentential topic equals the subject, the composition is done with SUBJ-HEAD, with no special treatment involved
- Temporal or location topics are treated as modifiers with ADJ-HEAD
- A special rule SUBJ2-HEAD is used to fill topics headed by noun or verb into the SPR valence of the main sentence. This is also referred to as the "double subject" constructions

Yu et al. (2010) introduce an extra valence feature (TOPIC) for the topic construction. Tse and Curran (2010) distinguish two types of topics,*gap* or *non-gap*. Both solutions are rather similar to ours nonetheless.

2.4 Numeral-classifiers & demonstratives

Numeral-classifier structures are analyzed as a phrase with rule SPEC-HEAD, and they together serve as a specifier to the head noun. A feature "CL" in the HEAD type of *noun* identifies the suitable groups of classifiers. Demonstratives are also treated as specifiers to nouns (similar to the double specifier account in (Ng, 1997)), though specific word order constraints are further enforced for the correct NP structure. Both specifiers of nouns are optional. The numeral before the classifier can be optional too, unless the NP is in a subject position and no demonstrative is available (e.g. * 头 大象 爱 吃 苹果).

2.5 DE-Constructions

DE (的) is involved in two major types of phrases:

- *Associative* DE-*phrase* where a semantic relation is created to associate the NPs before and after DE. The relation is similar to (and more general than) the possessive relation
- *Nominalizing* DE-*phrase* where DE combines with the predicative phrase before it to make a nominal phrase

While the *associative* DE-*phrase* is straightforward to model, the semantics of the *nominalizing* DE-*phrase* is more intriguing. We further categorize the nominalizing DE-phrase into the following three types:

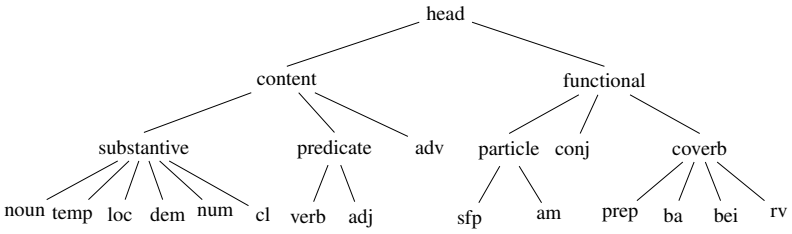

Figure 1: HEAD type hierarchy

- `subject gapping relative` DE
  where the NP after DE will serve as the subject to the predicative before DE
- `complement gapping relative` DE
  where the NP after DE will serve as the complement to the predicative before DE
- `non-gapping` DE
  where neither of the above two cases applies

Yu et al. (2010) mentioned the treatment of relative clauses using DE as a relativizer. However it is not clear whether different sub-types of the relative clauses (with different argument composition) are captured with specialized rules. Guo et al. (2007) differentiated three types of DE-constructions, ADJ-REL (relative clause), ADJUNCT (adjective), and POSS (possessive DE). We have a more fine-grained inventory for the relative clauses and treat the adjective case in the subject gapping relative DE-phrases (since we allow adjectives to be predicates, as shown in Figure 1). For example, 大 的 苹果 (big apple) will be analyzed as 大 (big) is the (adjectival) predicate of 苹果 (apple).

2.6 Locatives & temporals

Locative phrases serve as both pre-verbal and post-verbal modifiers, and generally take the form of *zai* + *NP* + *Loc*, e.g. 在 桌子 上 (on the table), 在 房子 东面 (to the east of the house), etc.

Locative phrases can always serve as pre-verbal modifiers. But only certain verbs can take post-verbal locatives with the HEAD-ADJ rule. The treatment of locative phrases as normal prepositional phrases as in (Wang et al., 2009) may lead to massive over-generation.

The analysis of temporal phrases is similar to the locative phrases.

2.7 BA-Construction

BA-construction moves the direct object of a verb to the pre-verbal position. In our analyses, we use a specialized unary rule BA-FRONTED to change the last element of the verb's complement list from $\begin{bmatrix} \text{HEAD} & \textit{noun} \\ \text{RC} & + \\ \text{INDEX} & \boxed{1} \end{bmatrix}$ to $\begin{bmatrix} \text{HEAD} & \textit{ba} \\ \text{RC} & - \\ \text{INDEX} & \boxed{1} \end{bmatrix}$.

There are various discussions on BA in the literature. Bender (2000) considered it as a verb, Gao (2000) and Wang et al. (2009) treated it as a case-marker, and Yu et al. (2010) as a preposition. We categorize BA as a special coverb. This makes it similar to prepositions. But it will be subcategorized by (instead of modifying) the verb phrase.

2.8 BEI-Construction

BEI-construction is used to compose passive voice sentences in Chinese. Similar to the analysis of BA, we use a specialized unary rule to promote the complement of the verb into the subject list, and change the original subject $\begin{bmatrix} \text{HEAD} & \textit{noun} \\ \text{INDEX} & \boxed{1} \end{bmatrix}$ into a *"bei"* headed left complement $\begin{bmatrix} \text{HEAD} & \textit{bei} \\ \text{RC} & - \\ \text{INDEX} & \boxed{1} \end{bmatrix}$.

Consistent with their analysis of BA, (Yu et al., 2010) treat BEI as a preposition. They view the complement of BEI as an extracted subject and use filler-head rule to combine the subject and the predicate. Guo et al. (2007), on the other hand, assume that the NP and VP following BEI is in one constituent, and will be case-marked by BEI jointly.

2.9 Various Markers

Several types of constructions were covered by the HEAD-MARKER/MARKER-HEAD rule, among them are the aspect markers (着, 了, 过), sentence-final particles (了, 吗), ordinal numeral prefix (第), etc. Various specific semantic information is supplemented by the marking construction.

2.10 Resultative verb compound

The resultative verb compounds refer to the compounding of a verb together with a resultative coverb (e.g., 来, 去, 开, 到, etc.), taking HEAD type *rv*, to signal the *"result"* of the action or process conveyed by the first verb. This is different to the normal modification in that the valency of the compound is mainly determined by the resultative coverb. We capture the compounding with a special RVC rule which will pass upward the head type from the first verb, and the complements from the resultative coverb.

2.11 Serial verb constructions

Serial verb construction refers to a group of complex phenomena in Mandarin Chinese where multiple verb phrases or clauses occurs in a sentence without any marker indicating the relationship between them. According to Li and Thompson (1989), it can be divided into four groups: i) two or more separate events; ii) one verb phrase or clause serving as the subject or direct object of another verb; iii) pivotal constructions; iv) descriptive clauses. We have adopted different analyses for each of them.

Yu et al. (2010) dealt mainly with the first case of the serial verb constructions. Two or more verbs were treated as coordinations, which can share subjects, topics or left-complements. Tse and Curran (2010) treated both serial verb constructions and resultative verb compound (see Section 2.10) as *verbal compounding*. Müller and Lipenkova (2009) offered more detailed theoretical analyses of certain Chinese serial verb constructions, capturing subtle semantic differences in the descriptive clauses category with additional constructional semantic relations. We intend to investigate their solutions in the future.

3 Development & Evaluation

The MCG is currently developed on the LKB platform (Copestake, 2002), which implements the typed feature structure formalism in TDL (Krieger and Schäfer, 1994). The first stage of grammar development was done with the help of the LinGO Grammar Matrix customization system, which took care of the general design of the feature geometry in HPSG, as well as the definition of the universal types for basic rule schemata and corresponding semantic compositions. Significant amount of development time were spent on the careful revision of the design and the constant debate on the treatment of various Chinese specific phenomena, while trying to keep in line with the classical HPSG theory and the conventions from other DELPH-IN grammars. As it currently stands, in addition to the types provided by the grammar Matrix, the MCG contains over 200 new type descriptions, and over 3000 lines of code in TDL. A small hand-crafted lexicon containing over 500 entries is currently used for development and testing.

Also developed is a phenomenon-oriented test suite of over 700 sentences (with both positive and negative test items). We randomly sampled 129 previously unseen sentences from the test suite and parsed them with MCG, among them are 110 wellformed sentences and 19 illformed.

		Gold standard	
		Positive	Negative
System	Positive	82	2
	Negative	28	17

Table 1: Test suite parsing performance of MCG

While the test set contains only short sentences, the phenomena are non-trivial from the linguistic view point. A sentence is considered to be successfully parse when there is a reading that is both syntactically and semantically correct. We achieve a high precision (82/84=97.6%) with an acceptable recall (82/110=74.5%). Among all negative sentences, the grammar only generated reading for two of them. One was due to the incorrect classifier constraints from a noun lexical entry. The other was due to the over-relaxed head selection in adjective+head modification. Both errors are fixed after observing the error. The parser outputs on average 5.04 readings per sentence, which attributes to the constraints we encoded in the grammar to avoid over-generation. Full coverage over phenomena such as coordinations is still lacking in MCG.

4 Summary

An overview of the MCG grammar design is presented, though the detailed presentation of our linguistic solutions does not fit in the short-paper format the workshop organizer chose for us. Nevertheless, the grammar is in line with the open-source spirit of DELPH-IN, and freely available for research purposes (http://mcg.opendfki.de/).

References

Emily M. Bender and Dan Flickinger. 2005. Rapid prototyping of scalable grammars: Towards modularity in extensions to a language-independent core. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing IJCNLP-05 (Posters/Demos)*, Jeju Island, Korea.

Emily M. Bender. 2000. The syntax of mandarin ba: Reconsidering the verbal analysis. *Journal of East Asian Linguistics*, 9(2):105–145, April.

Emily M. Bender. 2008. Evaluating a crosslinguistic grammar resource: A case study of Wambaya. In *Proceedings of ACL-08: HLT*, pages 977–985, Columbus, Ohio, June. Association for Computational Linguistics.

Ann Copestake, Dan Flickinger, Carl J. Pollard, and Ivan A. Sag. 2005. Minimal recursion semantics: an introduction. *Research on Language and Computation*, 3(4):281–332.

Ann Copestake. 2002. *Implementing Typed Feature Structure Grammars*. CSLI, Stanford, USA.

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

Yuqing Guo, Josef van Genabith, and Haifeng Wang. 2007. Treebank-based acquisition of lfg resources for chinese. In *Proceedings of LFG07 Conference*, pages 214–232.

Yuqing Guo. 2009. *Treebank-Based Acquisition of Chinese LFG Resources for Parsing and Generation*. Ph.D. thesis, School of Computing, Dublin City University, July.

Julia Hockenmaier and Mark Steedman. 2005. Ccgbank: User's manual. Technical Report MS-CIS-05-09, Department of Computer and Information Science, University of Pennsylvania.

Hans-Ulrich Krieger and Ulrich Schäfer. 1994. Tdl - a type description language for constraint-based grammars. In *Proceedings of the15th International Conference on Computational Linguistics (COLING '94), August 5-9*, pages 893–899.

Charles N. Li and Sandra A. Thompson. 1989. *Mandarin Chinese: A functional reference grammar*. University of California Press, London, England.

Yusuke Miyao, Takashi Ninomiya, and Jun'ichi Tsujii. 2004. Corpus-oriented grammar development for acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP 2004)*, pages 684–693, Hainan Island, China.

Stefan Müller and Janna Lipenkova. 2009. Serial verb constructions in chinese: A hpsg account. In *Proceedings of the 16th International Conference on Head-Driven Phrase Structure Grammar*, pages 234–254, Germany.

Say Kiat Ng. 1997. A double-specifier account of chinese nps using head-driven phrase structure grammar. Master's thesis, Department of Linguistics. University of Edinburgh.

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

Daniel Tse and James R. Curran. 2010. Chinese ccgbank: extracting ccg derivations from the penn chinese treebank. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1083–1091, Beijing, China.

Xiangli Wang, Shunya Iwasawa, Yusuke Miyao, Takuya Matsuzaki, and Jun'ichi Tsujii. 2009. Design of chinese hpsg framework for data-driven parsing. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, December.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(02):207–238.

Kun Yu, Miyao Yusuke, Xiangli Wang, Takuya Matsuzaki, and 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 (Coling 2010)*, pages 1417–1425, Beijing, China.