Joint Grammar and Treebank Development for Mandarin Chinese with HPSG

Yi Zhang, Rui Wang, Yu Chen

LT-Lab, German Reserach Center for Artificial Intelligence, Saarbrücken, Germany
{yizhang,ruiwang,yuchen}@dfki.de

Abstract

We present the ongoing development of MCG, a linguistically deep and precise grammar for Mandarin Chinese together with its accompanying treebank, both based on the linguistic framework of HPSG, and using MRS as the semantic representation. We highlight some key features of our grammar design, and review a number of challenging phenomena, with comparisons to alternative linguistic treatments and implementations. One of the distinguishing characteristics of our approach is the tight integration of grammar and treebank development. The two-step treebank annotation procedure benefits from the efficiency of the discriminant-based annotation approach, while giving the annotators full freedom of producing extra-grammatical structures. This not only allows the creation of a precise and full-coverage treebank with an imperfect grammar, but also provides prompt feedback for grammarians to identify the errors in the grammar design and implementation. Preliminary evaluation and error analysis shows that the grammar already covers most of the core phenomena for Mandarin Chinese, and the treebank annotation procedure reaches a stable speed of 35 sentences per hour with satisfying quality.

Keywords: Grammar Engineering, Treebank Annotation, Syntax

1. Introduction

High-quality deep linguistic grammars and richly annotated treebanks are two types of language resources which are both extremely rare and valuable to high-quality NLP application development as well as linguistic studies. Due to the contrasting differences in the nature of grammars and treebanks, traditional development typically takes completely different approaches. Linguistic grammars are typically the generalizations of the language mechanism which are not only responsible for the interpretation of the naturally occurring sentences, but are also capable of producing well-formed linguistic expressions. Treebanks, on the other hand, are annotations documenting the instantiations of various linguistic structures.

The conventional approach to grammar development is based on the manual work of grammarians through painstaking retrospective thinking. The development of a new grammar normally takes years or even decades. And the task is intellectually demanding, requiring high expertise in both linguistics (for compentent analysis) and computational skills (for practical and creative implementation). The traditional approach to treebank annotation is established on top of the so-called *annotation guidelines* which are sets of semi-formalized descriptive protocols for human annotators. While the treebank annotation also takes a long time, the possibility of having multiple annotators working in parallel can speed up the progress, provided that a satisfying level of inter-annotator agreement can be established.

In more recent developments, we have seen that the engineering approach of grammars and treebanks moved closer and start to benefit from each other. On the one hand, Miyao et al. (2004), Hockenmaier and Steedman (2005), Cahill et al. (2005), Cramer and Zhang (2009) showed that the development of linguistic grammars can be greatly accelerated by automatically learning detailed lexical information from manually annotated treebanks. Oepen et al. (2002), on the other hand, shows that the precision oriented linguistic grammars can also help bootstrap detailed annotation of large-scale corpora with limited amount of human intervention.

In this paper, we report on the on-going development of a Mandarin Chinese grammar (MCG) and the accompanying treebank with rich annotation. The development starts from the design of a detailed HPSG analyses for Mandarin Chinese, which plays a central role for both grammar implementation and the treebank annotation. We will first outline the design of the grammar, and then describe the workflow for the treebank annotation. Some preliminary evaluation of the resources are also provided in the end. Further, we point out several important differences between our annotation scheme and the one adopted by the Penn Chinese Treebank (Xue et al., 2005).

2. Related Work

Other recent Chinese grammar development work mainly focus on the grammar induction from converted CTB treebanks. Guo et al. (2007) annotated the CTB trees with fine-grained *f-structures* of LFG and learned the grammar based on the enriched trees. By using the hand-crafted gold-standard f-structures for 200 sentences from the CTB 5.1, they achieved 96.34% precision and 96.46% recall for unseen texts (Guo, 2009). Yu et al. (2010) converted CTB 6.0 into HPSG style and added predicate-argument structures. They extract an HPSG lexicon with 97.24% accuracy, and achieved 98.51% lexical coverage and 76.51% sentential coverage on unseen texts. Wang et al. (2009) had another design of the Chinese HPSG, but no experiments were reported yet. Tse and Curran (2010) built a Chinese CCGbank on top of CTB with CCG derivations. The corpus contains 760,000 words and their process yields a corpus of 27,759 derivations, covering 98.1% of the CTB.

Most of these approaches starts from an existing treebank with coarse-grained annotation. By asserting certain core linguistic principles on the basic syntactic structure, richer annotations are derived from category compositions. The resulting grammar will contain detailed instantiations of lexical templates, but lacks the linguistic generalizations found typically in hand-crafted grammars.

On the other hand, introspective grammar development methodology has also advanced nowadays, especially with the aid of accumulated experience in multilingual grammar engineering through the last decades. For example, the LinGO Grammar Matrix (Bender et al., 2002) helps jump-start the new grammar development by offering pre-designed common solutions based on a customization step. Our approach to the Chinese grammar development will benefit from these previous experiences, as well as emphasize the tight integration with the treebank development, which brings feedbacks from the corpus analysis to the grammar revision process.

3. An HPSG Analysis of Mandarin

3.1. 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 $\begin{bmatrix}\text{RC} & \pm\end{bmatrix}$ to indicate whether the complement is to the right or to the left of the head. The grammar currently contains about 40 rule schemata, many of which are highly generalized and used to handle multiple constructions.

3.2. HEAD types

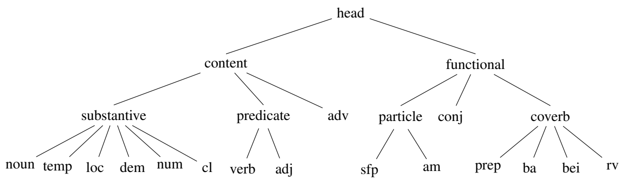


Figure 1: HEAD type hierarchy

The HEAD types in HPSG identify the major categories of parts-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 (e.g., 在, 用), BA (把), BEI (被), and resultative coverbs (e.g., 来, 开).

3.3. Principle Phenomena

3.3.1. Nominal Phrases & DE-Constructions

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 (Elephant likes eating apples).

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.

DE (的) is involved in two major types of phrases: i) *Associative* DE-*phrase* where a semantic relation is created to associate the NPs before and after DE; ii) *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: a) `subject gapping relative DE`, where the NP after DE serves as the subject to the predicative before DE; b) `complement gapping relative DE`, where the NP after DE serves as the complement to the predicative before DE; c) `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).

Mandarin Chinese is a topic-prominent language. 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: i) when the sentential topic equals the subject, the composition is done with SUBJ-HEAD, with no special treatment involved; ii) temporal or location topics are treated as modifiers with ADJ-HEAD; iii) 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.

3.3.2. Verbs and Co-verbs

BA-construction moves the direct object of a verb to the preverbal 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 (on a somehwat continuous spectrum between verbs and prepositions in Mandarin). It will be subcategorized by (instead of modifying) the verb phrase.

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 $\begin{bmatrix} \text{HEAD} & \textit{bei} \\ \text{RC} & - \\ \text{INDEX} & \boxed{1} \end{bmatrix}$. The current version of the grammar can also cover complex cases like 苹果 被 她 削 了 皮 (the apple is skinned by her).

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.

Several types of constructions were covered by the HEAD-MARKER rules, 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.

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 from 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 passes upward the head type from the first verb, and the complements from the resultative coverb.

Verbal modifiers are also quite tricky to handle, including manner adverbs (i.e., adjectives + 地), non-manner adverbs (e.g., 经常, 已经, etc.), (post-verbal) stative adverbs (i.e., 得), etc. The manner adverb 地 takes an adjective as its complement and form a normal adverb which can further modify a predicative phrase. The stative adverb 得 takes a predicative phrase as its complement and form an adverb as well, e.g., 高兴 得 跳 起来 ((someone) is so happy that (he/she) jumps up). Notice that adverbs can only occur after the subject or topic in Chinese.

3.3.3. Coordination & Serial Verbs

Coordination is infamously difficult to tackle. For the moment, we mainly take the solution provided by the Grammar Matrix, which combines coordinate items one by one. The issue of over-generation is still under investigation.

Serial verb construction refers to a group of complex phenomena in Mandarin Chinese where multiple verb phrases or clauses occur in a sentence without any markers 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.

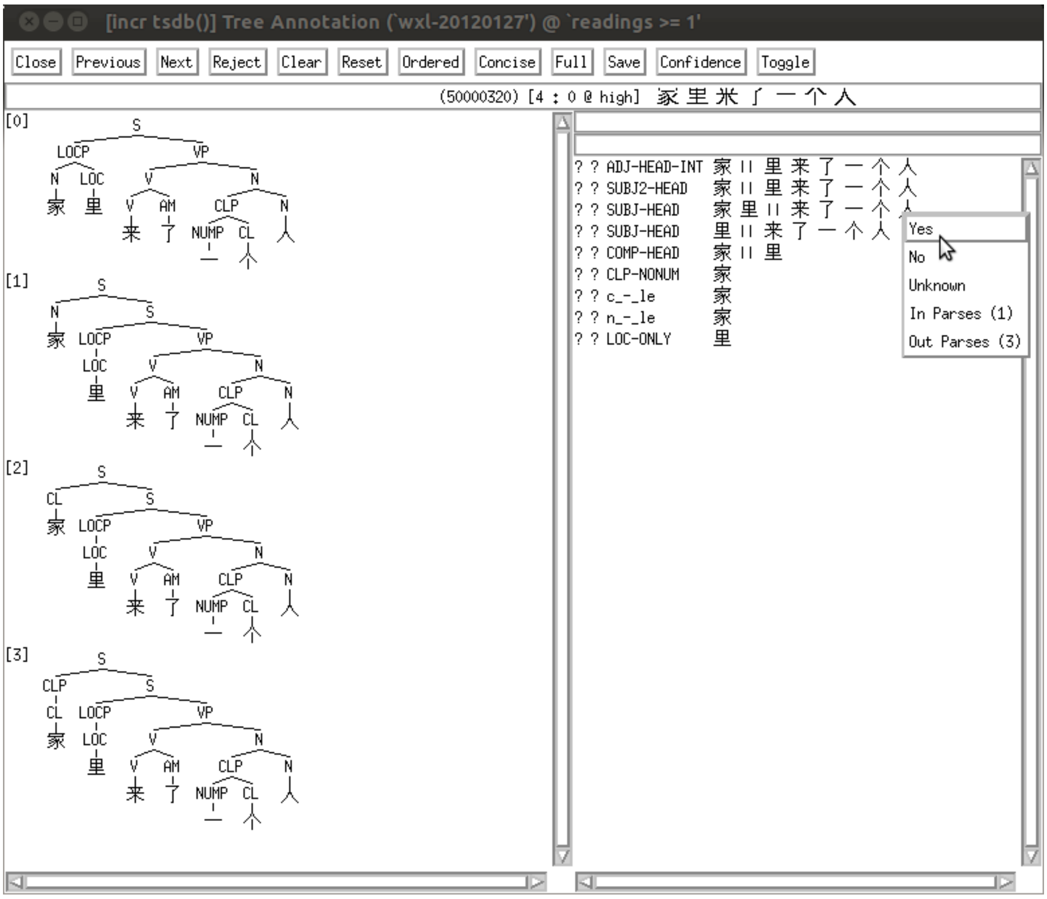
Yu et al. (2010) dealt mainly with the first case of the serial verb constructions. Two or more verbs were treated as coordinations, which can share subjects, topics or left-complements. Tse and Curran (2010) treated both serial verb constructions and resultative verb compound as *verbal compounding*. 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.

4. Treebank Development

The hand-crafted grammar achieves decent coverage with moderate overgeneration. The candidate analyses are recorded in the form of parse forests. However, to produce high quality treebank in the end, further manual annotation labor is required. First, although the readings in a parse forest are licensed by the grammar, not all of them are equally plausible. To record the most appropriate analysis for the sentence, human annotators must manually disambiguate the forest to arrive at the most desired reading. Furthermore, in case the best reading is still different from the ideal analysis which the grammar failed to produce, an additional editing step is required to make the necessary changes.

For the first step of the manual disambiguation, we use the discriminant-based approach (Carter, 1997; Oepen et al., 2002). The `[incr tsdb()]` system is used as the treebank annotation platform. A dynamic annotation cycle similar to the one presented in Kordoni and Zhang (2009) is used. To further increase the annotation speed, we use the blazing technique (Tanaka et al., 2005) to guide the selection of the desired reading by pruning the implausible ones with external annotations by either part-of-speech taggers or phrase structure annotations, and use the discriminant ranking mechanism (Zhang and Kordoni, 2010) for annotator-specific optimization. Figure 2 shows a screen shot of the disambiguation annotation interface, where the left panel shows a list of candidate derivation trees, and the right panel shows a list of discriminants identifying the atomic differences between readings in the parse forest.

Once the disambiguation step is done, each parsed sentence is paired with one candidate reading to be further adjudicate. The second step of annotation works on top of the derivations produced by the `MCG`. The annotators are allowed to further change the derivation tree where they see



```
(50000320) [4 : 0 @ high] 家 里 米 了 一 个 人

? ? ADJ-HEAD-INT 家 || 里 来 了 一 个 人
? ? SUBJ2-HEAD   家 || 里 来 了 一 个 人
? ? SUBJ-HEAD    家 里 || 来 了 一 个 人
? ? SUBJ-HEAD    里 || 来 了 一 个 人
? ? COMP-HEAD    家 || 里
? ? CLP-NONUM    家
? ? c_-_le       家
? ? n_-_le       家
? ? LOC-ONLY     里
```

Figure 2: A screenshot of the [incr tsdb()] discriminant-based disambiguation user interface

it appropriate. The incremental editing steps are recorded as further modifications to the original HPSG parses. The editing actions include constituent *insertion (I)*, *deletion (D)*, *renaming (R)*, and *reattachment (A)*. This set of operations allows arbitrary changes to the syntactic structure. Although this is not a minimal set (e.g., a *renaming* operation can be equally achieved by a pair of *deletion* and *insertion* operations), these operations are easy to perform. When assisted with the interactive graphical user interface (see Figure 3 for a screenshot), the annotators usually arrive at a satisfying tree within less than 10 operations. The edited derivations can be also used to produce semantics in the form of MRS. Due to the extra-grammaticality of such derivations, they will not form fully consistent HPSG analyses, i.e., TFS unification may fail on some of the constituents. However, here we only need unification as a mechanism for semantic composition, hence can ignore the (syntactic) constraints. Using a version of the LKB unifier that implements the default strategy, we can robustly reconstruct the MRS from the edited derivation trees. The resulting semantics are used as feedbacks to the grammarians for detecting flaws in the grammar design. With the intearctive unification debugging tool in LKB, we are able to quickly pinpoint the failures in large typed feature structures. Once an error is fixed in the grammar, we can also quickly update the annotations automatically with [incr tsdb()] without manually reannotating the sentences.

According to the dry-run, annotators can maintain a consistent disambiguation speed at around 50 sentences per hour for the first step. For the sentences that requires the post-editing, annotators can correct about 35-40 sentences per hour. The overall annotation throughput is about 35 sentences per annotator hour.

We use two sources of texts for the creation of the treebank. The first dataset contains a selective set of relatively short sentences targetting specific phenomena of Mandarin Chinese. This is also the test suite for our grammar development, and is highly interesting from the linguistic persepective. The second dataset contains raw sentences from the Penn Chinese Treebank, mostly newspaper texts. They are automatically parsed and manually disambiguated with our HPSG grammar. The resulting analysis is much finer-grained than the CTB annotation, including not only the phrase-structure trees, but also more detailed syntactic features and the semantic representation in the form of MRS. While the annotation progress will take many months to complete, we see already practical outcomes from the dry-run over the smaller datasets. The next section will report on some initial evaluation and comparison results.

5. Evaluation & Comparison

The MCG is currently developed on the LKB platform (Copestake, 2002), which implements the typed feature structure formalism. 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

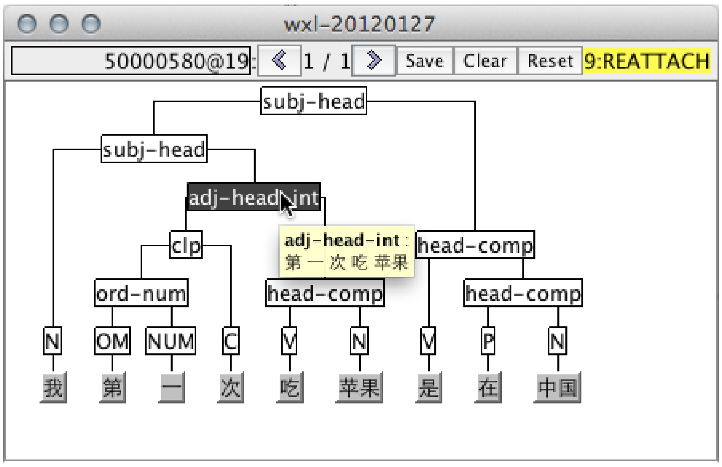


Figure 3: A screenshot of the derivation post-editor user interface

definition of the universal types for basic rule schemata and corresponding semantic compositions. Significant amount of development time was 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[1]. 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.

The phenomenon-oriented test suite contains 732 sentences (with both positive and negative test items, 632 and 100 ones respectively). In our previous work, we showed promising results on a randomly sampled set of 129 unseen sentences (Zhang et al., 2011) and here we show the results on the whole test suite with the latest version of MCG.

		Gold standard	
		Positive	Negative
System	Parsed	**305**	34
	Rejected	327	**66**
	Recall	48.3%	66.0%

Table 1: Test suite parsing performance of MCG

While the test set contains only short sentences[2], the phenomena are non-trivial from the linguistic view point. A sentence is considered to be successfully parsed when there is a reading that is both syntactically and semantically correct. We achieve a high sentential precision (305/(305+34)=90.0%) with an acceptable recall (48.3%). Among all the negative sentences, the grammar can also reject 66% of them, which is due to the fine-grained constraints in the grammar.

In comparison with the CTB annotation, we see several advantages of our constraint-basd approach. For instance, the prevalent noun/verb ambiguity in Mandarin (partially due to the lack of inflectional morphology) has led to inconsistency in annotation. A word with such ambiguity in a object position can sometimes be annotated as a noun and subcategorieze for by the main verb, while in other cases be annotated as a verb and projected to the sentential level (IP) with an empty traced subject. In our treatment, the second (verbal) interpretation only occurs when a predicative complement is permitted by the main verb.

6. Conclusion and Future Work

In this paper, we presented the on-going development of a Mandarin Chinese grammar in the framework of HPSG. Modern grammar engineering techniques were employed to jump-start the development with the help of the LinGO Grammar Matrix.

In the design of the grammar, we notice that many language specific phenomena can be treated in a uniform way that fits naturally into the framework of HPSG. Comparing to the previous studies, we see relatively less ad-hoc deviations in our analysis from the classical theory even when complex phenomena are concerned. As a result, we expect our design of the grammar to be clean and extensible.

Given that the grammar is still in its early stage of development, we are aware that it is still far from achieving broad coverage. Future work on extending the grammar with both manual grammar engineering and automated corpus-driven learning approaches are planned.

The most straightforward application of the grammar is parsing. The preliminary evaluation of the current version of our grammar has already shown promising results on the test suite (Section 5.). Another main advantage of deep grammars is to support bi-directional processing. When we design the grammar, we aim at a *precise* account of the language phenomena, in order to avoid over-generation. The hypothesis space of the analyses is clearly much smaller than that of the treebank-induced grammars. This makes it feasible to use the grammar as a sentence realization model (i.e., text generation from semantics). Further application of the grammar can be seen in enriching the annotations

[1] http://moin.delph-in.net/GrammarCatalogue

[2] The average length of Chinese sentences are in general shorter than the English ones, even if each Chinese character is counted as one word. Multiple shorter sentences are preferred over a long one with complex embeded subordinate clauses.

in existing treebanks. This will further allow us to acquire deeper grammars with corpus-driven approaches.

Last but not least, the grammar will keep in line with the open-source spirit of `DELPH-IN`, and be freely available for research purposes. At the time of this publication, a preview release of the grammar is available at `http://mcg.opendfki.de/`.

Acknowledgements

The first and second authors thank the *Deependance* project funded by the BMBF (contract 01IW11003) for its support of the work. The third author is supported by European Community through the EuroMatrixPlus project (ICT-231720) funded under the Seventh Framework Programme for Research and Technological Development. The authors would like to thank Xiangli Wang for providing the test suite and Dan Flickinger, Antske Fokkens, Janna Lipenkova, Takuya Matsuzaki, Yusuke Miyao, Stefan Müller, Chunlei Yang, and Kun Yu for precious discussions.

7. References

Emily Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-lit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation*.

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

Aoife Cahill, Michael Burke, Martin Forst, Ruth Odonovan, Christian Rohrer, Josef Genabith, and Andy Way. 2005. Treebank-based acquisition of multilingual unification grammar resources. *Research on Language and Computation*, 3(2):247–279.

David Carter. 1997. The treebanker: a tool for supervised training of parsed corpora. In *Proceedings of the ACL Workshop on Computational Environments for Grammar Development and Linguistic Engineering*, pages 9–15, Madrid, Spain.

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

Bart Cramer and Yi Zhang. 2009. Construction of a german hpsg grammar from a detailed treebank. In *Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks*.

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.

Valia Kordoni and Yi Zhang. 2009. Annotating wall street journal texts using a hand-crafted deep linguistic grammar. In *Proceedings of The Third Linguistic Annotation Workshop (LAW III)*, Singapore.

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 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.

Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger, and Thorsten Brants. 2002. The LinGO Redwoods treebank: motivation and preliminary applications. In *Proceedings of COLING 2002*, Taipei, Taiwan.

Takaaki Tanaka, Francis Bond, Stephan Oepen, and Sanae Fujita. 2005. High precision treebanking—blazing useful trees using POS information. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 330–337, Ann Arbor, Michigan. Association for Computational Linguistics.

Daniel Tse and James R. Curran. 2010. Chinese ccgbank: extracting ccg derivations from the penn chinese treebank. In *Proceedings of 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 PACLIC 2009*, 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 COLING 2010*, pages 1417–1425, Beijing, China.

Yi Zhang and Valia Kordoni. 2010. Discriminant ranking for efficient treebanking. In *Coling 2010: Posters*, pages 1453–1461, Beijing, China.

Yi Zhang, Rui Wang, and Yu Chen. 2011. Engineering a deep hpsg for mandarin chinese. In *Proceedings of the 9th Workshop On Asian Language Resources*, Chiang Mai, Thailand, November.