Building Zhong [||], a Chinese HPSG Shared-Grammar

Zhenzhen Fan◇♠   **Sanghoun Song**♠   **Francis Bond**♠

◇Institute of Systems Science, National University of Singapore

♠Division of Linguistics and Multilingual Studies, Nanyang Technological University

Singapore

zhenzhen@nus.edu.sg, {sanghoun,fcbond}@ntu.edu.sg

1 Introduction

Head-Driven Phrase Structure Grammar (HPSG: Pollard & Sag, 1994) is a lexicalized generative grammar theory developed by Carl Pollard and Ivan Sag at Stanford University. An HPSG-based grammar includes constraint-based grammar rules and a lexicon containing syntactic and semantic information about words, which makes it very useful as a grammar framework in natural language processing for deep linguistic analysis of human language aiming at content level understanding.

Computational linguists from different research centers worldwide have been collaborating to develop broad coverage HPSG grammars of different languages in a consortium called Deep Linguistic Processing with HPSG (DELPH-IN, *http://www.delph-in.net*). Broad coverage HPSGs for English (LinGO English Resource Grammar, ERG: Flickinger, 2000), German (GG: Müeller & Kasper, 2000; Crysmann, 2005), Japanese (Jacy: Siegel & Bender, 2002), Korean (KRG: Kim et al., 2011), Spanish (SRG: Marimon, 2012), Portuguese (LXGram: Branco & Costa, 2010), Norwegian (NorSource: Hellan, 2005), and several other languages have been developed and used in various applications.

We have been developing a Chinese HPSG shared-grammar named Zhong (Fan et al., 2015) based upon the existing work on Mandarin Chinese from the HPSG community. The objective is to build a broad-coverage resource grammar that can be used for applications such as machine translation and computer aided language learning. We take a corpus-driven approach to improving its coverage through grammar rule enhancement and lexicon expansion. This paper describes how we approach two Chinese phenomena, reduplicated adjectives and SUO structure, and implement them in our grammar.

2 Previous Work on Chinese HPSG

Since 1990s, linguistic analysis of specific Chinese phenomena in HPSG framework started to appear (Xue et al., 1994; Xue & McFetridge, 1996; Gao, 1994; Xue & McFetridge, 1995; Ng, 1997). Subsequently, two PhD theses (Gao, 2000; Li, 2001) documented the efforts towards a more comprehensive analysis of Chinese, covering major phenomena such as topic sentences, valence alternations (including BA, ZAI, and other constructions), as well as separable verbs and Chinese derivation and affixes.

More recent works accompany linguistic analysis with computational implementation, leading to several independently developed HPSG grammars on Mandarin Chinese: MCG (Zhang et al., 2011), ManGO (Yang, 2007), and ChinGram (Müller & Lipenkova, 2013), all adopting Minimal Recursion Semantics (MRS) (Copestake et al., 2005) as the semantic representation format. These grammars focus on a variety of linguistic phenomena in Chinese, but typically only cover the words appearing in their testsuites.

3 Zhong

There are many varieties of Chinese, historically related but now separate languages. Zhong models some of these varieties in a single hierarchy. The different Chinese grammars share some elements, such as basic word order, and have other elements distinct, such as lexemes and specific grammar rules (e.g., classifier constructions).

Taking the original implementation of ManGO, we restructured it as follows: All grammars build upon the common constraints and inherit from `zhong.tdl`, `zhong-lextypyes.tdl`, and `zhong-letypes.tdl`. The differences between Mandarin and Cantonese, such as NP structures, are reflected in `cmn.tdl` and `yue.tdl`, respectively. The Mandarin Chinese grammars are

further divided into `zhs` and `zht` depending on whether simplified characters or traditional characters are used. Further distinction between the two are modeled in `zhs.tdl` and `zht.tdl`, respectively.

The official webpage of Zhong, with demo and test results, is *http://wiki.delph-in.net/moin/ZhongTop*. And the entire data set can be freely downloaded from *https://github.com/delph-in/zhong*.

4 Chinese-specific Phenomena

As part of the efforts to enhance the grammar's coverage of Chinese phenomena, we have since analysed and implemented several Chinese-specific phenomena such as VV resultative compounds, A-no-A questions, sentence end particles, interjections and fragments. The details of these analyses can be found in several related presentations at this conference. Here we present how we handled another two phenomena, reduplicated adjectives and the SUO structure.

4.1 Reduplicated Adjectives

According to Li & Thompson (1989), reduplication is a morphological process of repeating a morpheme to form a new word, which mainly applies to verbs and adjectives in Chinese. When applied to disyllabic adjectives, reduplication repeats each character independently, from AB to AABB, as in (1).
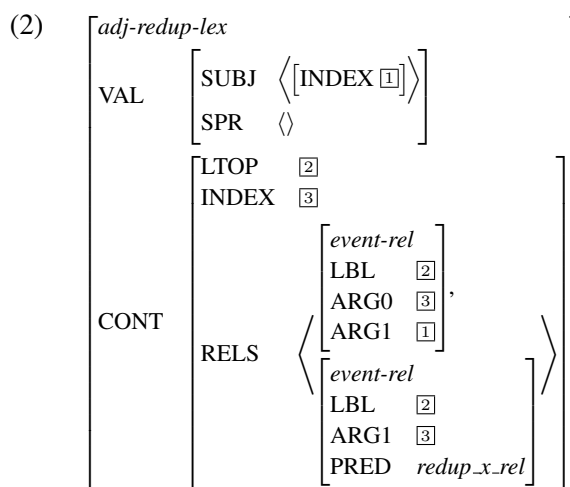
(1)　张三　　　干干净净
　　　zhāngsān　gāngānjìngjìng
　　　PRN　　　ADJ:AABB

　　　"Zhangsan is **clean**"

The meaning of the reduplicated adjective AABB is more vivid or intensified than its original form AB (Li & Thompson, 1989). Based on our position that sentences with similar meaning should have similar semantic representations, we model the semantic representation of AABB adjectives with the predicate of the AB adjective and a predicate that acts as an intensifier.

Since AABB reduplication of AB adjective is not very productive in Chinese (i.e., there are few AB adjectives that can be reduplicated this way), instead of using lexical rules to produce AABB from AB, we directly created lexical entries for the AABB adjectives (76 so far). Their lexical

type is defined to contain the predicate of AB and *redup_x_rel*, as shown in (2).

(2)
$$
\begin{bmatrix}
\textit{adj-redup-lex} \\
\text{VAL} \begin{bmatrix} \text{SUBJ} & \langle [\text{INDEX}\ \boxed{1}] \rangle \\ \text{SPR} & \langle \rangle \end{bmatrix} \\
\text{CONT} \begin{bmatrix} \text{LTOP} & \boxed{2} \\ \text{INDEX} & \boxed{3} \\ \text{RELS} & \left\langle \begin{bmatrix} \textit{event-rel} \\ \text{LBL} & \boxed{2} \\ \text{ARG0} & \boxed{3} \\ \text{ARG1} & \boxed{1} \end{bmatrix}, \begin{bmatrix} \textit{event-rel} \\ \text{LBL} & \boxed{2} \\ \text{ARG1} & \boxed{3} \\ \text{PRED} & \textit{redup\_x\_rel} \end{bmatrix} \right\rangle \end{bmatrix}
\end{bmatrix}
$$

The dependency graph representing the MRS structure of (1) is provided in (3), which basically neans "Something called "张三" is redup clean".

(3)
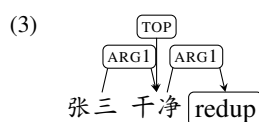
redup_x_rel is one of the three predicates inheriting from a common parent *intensifier_x_rel*, as illustrated in (4)

(4)
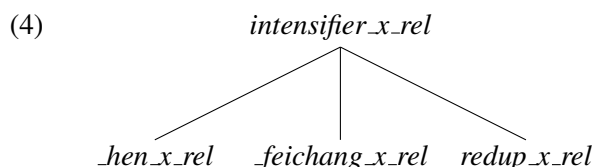
With this definition, if we generate from an MRS representation "Something called "张三" is intensifier clean", we can get three possible surface forms:

(5)　a.　张三　　　很　　干净
　　　　　zhāngsān　hěn　gānjìng
　　　　　Zhangsan　very　clean

　　　　　"Zhangsan is very clean"

　　　b.　张三　　　非常　　　干净
　　　　　zhāngsān　fēicháng　gānjìng
　　　　　Zhangsan　extremely　clean

　　　　　"Zhangsan is extremely clean"

　　　c.　张三　　　干干净净
　　　　　zhāngsān　gāngānjìngjìng
　　　　　Zhangsan　REDUP-clean

　　　　　"Zhangsan is **clean**"

Other specific redupliation patterns like AAB, ABB, ABAC, etc., will also be added as lexical entries. More productive reduplication patterns, such as AA for monosyllabic adjectives and verbs, and ABAB for disyllabic verbs, will be handled using lexical rules.
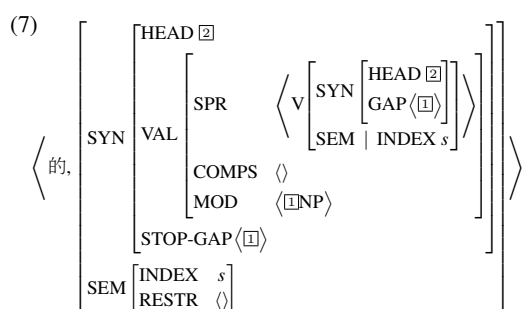
4.2 SUO structure

In Mandarin Chinese, 所 *suǒ* is a particle used before a transitive verb to nominalize the structure "SUO+V" into a noun phrase (Lǚ, 1999). According to Lu & Ma (1985), in modern Chinese, SUO is used most commonly in the structure "(NP$_1$+)SUO+V+DE", either to modify a noun following it (NP$_2$) or to act as a noun phrase itself. One such usage, "NP$_1$+SUO+V+DE+NP$_2$", is shown in example (6).

(6) 他   所   写   的   书
     tā suǒ xiě   DE   shū
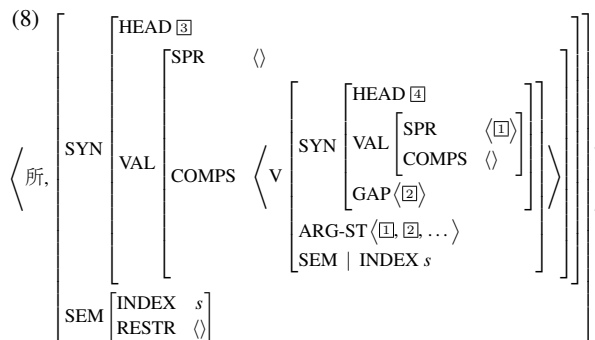   he SUO write DE   book

"the book written by him"

We take the view of Deng (2009) that in "NP$_1$+SUO+V+DE", DE plays the key role of nominalizing the phrase, so that it can be a prenominal adjunct (relative clause) to NP$_2$. The role of SUO in the construction is to indicate that the missing argument of the verb is its patient or direct object.

The lexical entry for the relativizing DE is presented in (7). The feature SPR of DE selects a preceding verbal clause. DE heads the resulting relative clause which is expected to contain a gap coreferential with the noun it modifies. The GAP value of DE's selected clause is defined to be identical to the NP in DE's MOD. DE's non-empty STOP-GAP feature ensures that it performs the gap-filling required.

DE also shares its HEAD feature with that of the selected clause. Semantically, DE does not introduce any information, so its RESTR list is empty, and its INDEX is the same as that of its selected clause.

(7)
$$\left\langle 的, \begin{bmatrix} \text{SYN} \begin{bmatrix} \text{HEAD } \boxed{2} \\ \text{VAL} \begin{bmatrix} \text{SPR} & \left\langle V \begin{bmatrix} \text{SYN} \begin{bmatrix} \text{HEAD } \boxed{2} \\ \text{GAP} \langle \boxed{1} \rangle \end{bmatrix} \\ \text{SEM | INDEX } s \end{bmatrix} \right\rangle \\ \text{COMPS} & \langle \rangle \\ \text{MOD} & \langle \boxed{1} \text{NP} \rangle \end{bmatrix} \\ \text{STOP-GAP} \langle \boxed{1} \rangle \end{bmatrix} \\ \text{SEM} \begin{bmatrix} \text{INDEX} & s \\ \text{RESTR} & \langle \rangle \end{bmatrix} \end{bmatrix} \right\rangle$$

The lexical entry for SUO is shown in (8). SUO selects a transitive verb which has an unrealized subject and a GAP value referring to its direct object (2nd item on ARG-ST list). As a non-head marker marking the missing object, SUO has nothing to add on semantically.

(8)
$$\left\langle 所, \begin{bmatrix} \text{SYN} \begin{bmatrix} \text{HEAD } \boxed{3} \\ \text{VAL} \begin{bmatrix} \text{SPR} & \langle \rangle \\ \text{COMPS} & \left\langle V \begin{bmatrix} \text{SYN} \begin{bmatrix} \text{HEAD } \boxed{4} \\ \text{VAL} \begin{bmatrix} \text{SPR} & \langle \boxed{1} \rangle \\ \text{COMPS} & \langle \rangle \end{bmatrix} \\ \text{GAP} \langle \boxed{2} \rangle \end{bmatrix} \\ \text{ARG-ST} \langle \boxed{1}, \boxed{2}, \dots \rangle \\ \text{SEM | INDEX } s \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix} \\ \text{SEM} \begin{bmatrix} \text{INDEX} & s \\ \text{RESTR} & \langle \rangle \end{bmatrix} \end{bmatrix} \right\rangle$$

(7) and (8) interact to produce the noun phrase structure for (6) in (9). In the tree, SUO constrains the missing argument of the verb to be the direct object. This information, contained in feature GAP, is passed up the tree, until the S or VP combines with DE to form a relative clause.

(9)
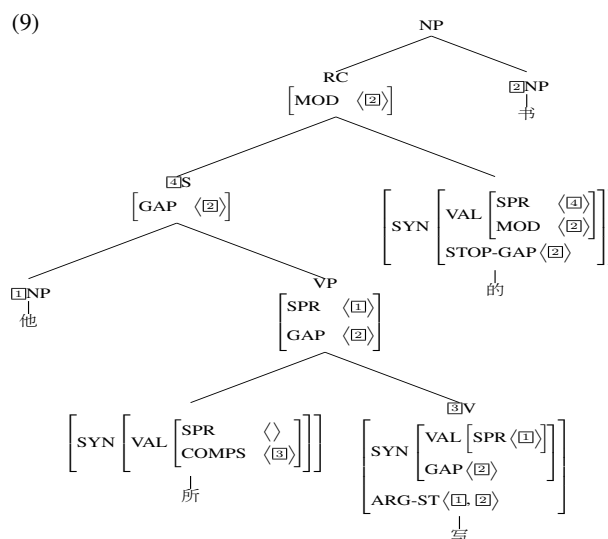
We have implemented SUO into our grammar and the implementation for semantic gap-filling of DE is currently in progress.

5 Future Work

Treebanking using the current version of Zhong has revealed many gaps, expecially in dealing with longer sentences found in real text, where different phenomena tend to interact to make constraint specification challenging . We plan to focus our subsequent efforts on phenomena that would help parse such longer sentences. Some of the tasks on

the immediate agenda are: relative clauses, variations of nominalisation, serial verb construction, conjunctions, other forms of VV compounds, etc. Lexical acquisition for zht and yue will also be performed to expand their lexical coverage.

References

Branco, António & Francisco Costa. 2010. A Deep Linguistic Processing Grammar for Portuguese. In *Computational Processing of the Portuguese Language*, vol. LNAI6001 Lecture Notes in Artificial Intelligence, 86–89. Berlin: Springer.

Copestake, Ann, Dan Flickinger, Carl Pollard & Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation* 3(2-3). 281–332.

Crysmann, Berthold. 2005. Relative Clause Extraposition in German: An Efficient and Portable Implementation. *Research on Language and Computation* 3(1). 61–82.

Deng, Dun. 2009. xiàndài hànyǔ suǒ jí suǒ zì jiégòu de chóngxīn shěnshì 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 Mordern Chinese. *Hànyǔ xuéxí*《汉语学习》（*Chinese Language Learning*）2. 106–112.

Fan, Zhenzhen, Sanghoun Song & Francis Bond. 2015. An HPSG-based Shared-Grammar for the Chinese Languages: ZHONG [—]. In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop*, 17–24. Beijing, China: Association for Computational Linguistics. http://www.aclweb.org/anthology/W15-3303.

Flickinger, Dan. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering* 6(1). 15–28. (Special Issue on Efficient Processing with HPSG).

Gao, Qian. 1994. Chinese NP Structure. *Linguistics* 32. 475–510.

Gao, Qian. 2000. Argument structure, HPSG, and Chinese grammar: Ohio State University dissertation.

Hellan, Lars. 2005. Implementing Norwegian Reflexives in an HPSG Grammar. In *Proceedings of the 12th International Conference on Head-driven Phrase Structure Grammar (HPSG)*, 519–539. Stanford: CSLI Publications.

Kim, Jong-Bok, Jaehyung Yang, Sanghoun Song & Francis Bond. 2011. Deep Processing of Korean and the Development of the Korean Resource Grammar. *Linguistic Research* 28(3). 635–672.

Li, Charles N & Sandra A Thompson. 1989. Mandarin Chinese: A functional reference grammar. Univ of California Press.

Li, Wei. 2001. The Morpho-Syntactic Interface in a Chinese Phrase Structure: Simon Fraser University dissertation.

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 sǎnlùn*《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.

Marimon, Montserrat. 2012. The Spanish DELPH-IN Grammar. *Language Resources and Evaluation* 47(2). 371–397.

Müeller, 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 & Janna Lipenkova. 2013. ChinGram: A TRALE Implementation of an HPSG Fragment of Mandarin Chinese. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, 240–249. Taipei, Taiwan: Department of English, National Chengchi University.

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.

Pollard, Carl & Ivan A. Sag. 1994. Head Driven Phrase Structure Grammar. Chicago: University of Chicago Press.

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.

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

Xue, Ping & Paul McFetridge. 1996. Complement Structure in Chinese. In *Proceedings of the 32nd Annual Meeting of the Chicago Lingusitic Society (CLS '96)*, Chicago, IL.

Xue, Ping, Carl Pollard & Ivan A Sag. 1994. A new perspective on Chinese ziji. In *the Proceedings of the Thirteenth West Coast Conference on Formal Linguistics*, .

Yang, Chunlei. 2007. Expert Systems for Pragmatic Interpretations of *ziji* and Quantified Noun Phrases in HPSG: Shanghai International Studies University dissertation.

Zhang, Yi, Rui Wang & Yu Chen. 2011. Engineering a Deep HPSG for Mandarin Chinese. In *Proceedings of the 9th Workshop On Asian Language Resources*, Chiang Mai, Thailand.