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: Müller & 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-lexypyes.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.


### 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).

\[
(1) \quad \text{张三 干净} \quad \text{zh¯angs¯an} \quad \text{g¯anj`ıng}
\]

“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 \( \text{redup}_\text{x-rel} \), as shown in (2).

\[
(2) \quad \begin{align*}
&\text{adj-redup-lex} \\
&\text{VAL} \\
&\text{SUBJ} \langle\text{INDEX}\rangle \\
&\text{SPR} \langle\rangle \\
&\text{LTOP} \langle\rangle \\
&\text{INDEX} \langle\rangle \\
&\text{CONT} \quad \text{LTOP} \langle\rangle \\
&\text{RELS} \langle\text{event-rel}\rangle \\
&\text{LBL} \langle\rangle \\
&\text{ARG0} \langle\rangle \\
&\text{ARG1} \langle\rangle \\
&\text{PRED} \quad \text{redup}_\text{x-rel}
\end{align*}
\]

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

\[
(3) \quad \begin{align*}
&\text{张三 干净 redup}
\end{align*}
\]

\( \text{redup}_\text{x-rel} \) is one of the three predicates inheriting from a common parent \( \text{intensifier}_\text{x-rel} \), as illustrated in (4).

\[
(4) \quad \begin{align*}
&\text{intensifier}_\text{x-rel} \\
&\_\text{hen}_\text{x-rel} \\
&\_\text{feichang}_\text{x-rel} \\
&\_\text{redup}_\text{x-rel}
\end{align*}
\]

With this definition, if we generate from an MRS representation “Something called “张三’ is intensifier clean”, we can get three possible surface forms:

\[
(5) \quad \begin{align*}
&\text{a. 张三 很 干净} \\
&\text{zh¯angs¯an hˇen g¯anj`ıng} \\
&\text{Zhangsan very clean} \\
&\text{“Zhangsan is very clean”}
\end{align*}
\]

\[
(5) \quad \begin{align*}
&\text{b. 张三 非常 干净} \\
&\text{zh¯angs¯an f¯eich´ang g¯anj`ıng} \\
&\text{Zhangsan extremely clean} \\
&\text{“Zhangsan is extremely clean”}
\end{align*}
\]

\[
(5) \quad \begin{align*}
&\text{c. 张三 干干净净} \\
&\text{zh¯angs¯an g¯ang¯anj`ıng} \\
&\text{Zhangsan REDUP-clean} \\
&\text{“Zhangsan is clean”}
\end{align*}
\]
Other specific reduplication 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 (Lü, 1999). 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 such usage, "NP1+SUO+V+DE+NP2", is shown in example (6).

(6) 他所写的书
    tā suǒ xiě DE shū
    "the book written by him"

We take the view of Deng (2009) that in "NP1+SUO+V+DE", DE plays the key role of nominalizing the phrase, so that it can 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.

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.

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.

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, especially 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


