Parsing Chinese with a Generalized Categorial Grammar

Manjuan Duan **William Schuler**
Department of Linguistics
The Ohio State University
{duan,schuler}@ling.osu.edu

Abstract

Categorial grammars are attractive because they have a clear account of unbounded dependencies. This accounting is especially important in Mandarin Chinese which makes extensive usage of unbounded dependencies. However, parsers trained on existing categorial grammar annotations (Tse and Curran, 2010) extracted from the Penn Chinese Treebank (Xue et al., 2005) are not as accurate as those trained on the original treebank, possibly because enforcing a small set of inference rules in these grammars leads to large sets of categories, which cause sparse data problems. This work reannotates the Penn Chinese Treebank into a generalized categorial grammar which uses a larger rule set and a substantially smaller category set while retaining the capacity to model unbounded dependencies. Experimental results show a statistically significant improvement in parsing accuracy with this categorial grammar.

1 Introduction

Categorial grammar annotations are attractive because they have a transparent syntactic-semantic interface and provide a natural account of traces (Rimell et al., 2009; Nguyen et al., 2012). This is especially important in parsing Chinese, which generates 1.5 times as many traces as English and makes heavy use of unbounded dependencies (Kummerfeld et al., 2013). Unfortunately, the accuracy of parsers trained on existing categorial grammar reannotations (Chinese CCGbank; Tse and Curran, 2010) of the Penn Chinese Treebank (Xue et al., 2005) is much lower than that of parsers trained on the original Treebank (Tse and Curran, 2012). This may be because previous attempts used Combinatory Categorial Grammar (CCG; Steedman, 2000; Steedman, 2012), which is strongly lexicalized (Karttunen, 1989), using a small set of language-independent rules and consequently a large set of language-dependent categories. This strong lexicalization may contribute to sparse data problems.

This work reannotates the Penn Chinese Treebank into a 'moderately lexicalized' generalized categorial grammar, similar to that defined for English by Nguyen et al (2012), which uses a larger set of language-specific inference rules and a substantially smaller category set. Experimental results show a statistically significant gain in parsing accuracy from this moderately lexicalized grammar over parsing with a strongly lexicalized CCG.

2 Grammar Framework

A generalized categorial grammar (GCG; Bach, 1981; Nguyen et al., 2012)[1] is a tuple $\langle P, O, R, W, M \rangle$ (Oehrle, 1994) consisting of a set P of primitive category types, a set O of type-constructing operators, a set R of inference rules, a set W of vocabulary items, and a mapping M from vocabulary items to complex category types. A set of complex category types C may then be defined as: $P \subset C$; $C \times O \times C \subset C$; nothing else is in C.

The mapping M in a GCG defines a category type c and a constraint function g encoded by each lexeme $w \in W$, using the notation $w \mapsto c : g$. Encoded constraints are expressed using dependency functions,[2] labeled with dependency types or argument position numbers: $\mathbf{f}_0$, $\mathbf{f}_1$, $\mathbf{f}_2$, etc. For example, a constraint function g may consist of a single '0'-labeled dependency to a constant 'people':

[1] Nguyen et al (2012) notate the '//' and '\\' operators of Bach (1981) as **-g** and **-h**, mnemonic for 'gap' and 'heavy shift'.

[2] Dependencies shown here can be interpreted as a shorthand for distributed representations of sentence meanings compatible with cognitive computational neuroscientific models of episodic memory (Schuler and Wheeler, 2014).

Proceedings of the Grammar Engineering Across Frameworks (GEAF) Workshop, 53rd Annual Meeting of the ACL and 7th IJCNLP, pages 25–32,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

$\lambda_x\,(\mathbf{f}_0\,x)$=people.

3 Chinese Syntax in GCG

Chinese is typically an SVO language, but it also has several SOV constructions, such as the focus constructions triggered by *lian* ‘even,’ or the *ba* construction, where the affected patient is moved to the preverbal position. Most adverbial modifiers are pre-verbal and most nominal modifiers, including relative clauses, are pre-nominal.

The set of primitive category types for Mandarin Chinese, P, $P \subset C$, contains the following primitive categories, generally labeled with the part of speech of the head of the category:

V: verb-headed clause
N: noun-headed phrase or clause
D: *de*-clause
C: cardinal number
Q: quantificational phrase
A: adjectival phrase or nominal modifier
R: adverbial phrase or verbal modifier
B: verbal complement of *ba*
E: verbal complement of *bei*

The set of type-constructing operators O for Mandarin Chinese includes **-a** and **-b** operators for unsatisfied requirements of preceding or succeeding arguments, **-c** and **-d** operators for unsatisfied requirements of preceding or succeeding conjuncts, and a **-g** operator for unsatisfied requirements of gap categories.[3] A GCG category consists of a primitive category followed by one or more unsatisfied dependencies, each consisting of an operator followed by another category.

The set of inference rules R is described below.

3.1 Argument composition

The basic operation of most categorial grammars is argument composition. However, unlike most categorial grammars, the GCG described in this paper defines composition rules to explicitly encode dependencies between lexical items. Specifically, inference rules for argument composition are defined as follows, where $c \in C$, $p \in P$ and each $\varphi \in \{\textbf{-a}, \textbf{-b}\} \times C$:

$$c{:}g \;\; p\varphi_{1..n\text{-}1}\textbf{-a}c{:}h \Rightarrow p\varphi_{1..n\text{-}1}{:}\lambda_x\, g\,(\mathbf{f}_n\,x) \wedge (h\,x) \quad \text{(Aa)}$$

$$p\varphi_{1..n\text{-}1}\textbf{-b}c{:}g \;\; c{:}h \Rightarrow p\varphi_{1..n\text{-}1}{:}\lambda_x\,(g\,x) \wedge h\,(\mathbf{f}_n\,x) \quad \text{(Ab)}$$

The first composition rule Aa stipulates that when a predicate h of category $p\varphi_{1..n\text{-}1}\textbf{-a}c$ takes a preceding argument g of category c as its n-th argument, the syntactic dependency that g is h’s n-th argument is added. The second composition rule Ab is an argument composition rule taking a succeeding argument.

3.2 Modifier composition

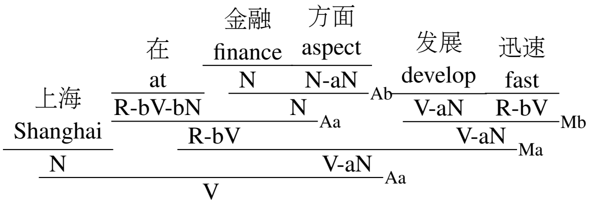
Inference rules for modifier composition apply preceding or succeeding modifiers of category $p\textbf{-b}d$ to modificands of category c, where $p \in \{\mathbf{A}, \mathbf{R}\}$, $d \in \{\mathbf{N}, \mathbf{V}\}$:

$$p\textbf{-b}d{:}g \;\; c{:}h \Rightarrow c{:}\lambda_x\,\exists_y\,(g\,y) \wedge (h\,x) \wedge (\mathbf{f}_1\,y){=}x \quad \text{(Ma)}$$

$$c{:}g \;\; p\textbf{-b}d{:}h \Rightarrow c{:}\lambda_x\,\exists_y\,(g\,x) \wedge (h\,y) \wedge (\mathbf{f}_1\,y){=}x \quad \text{(Mb)}$$

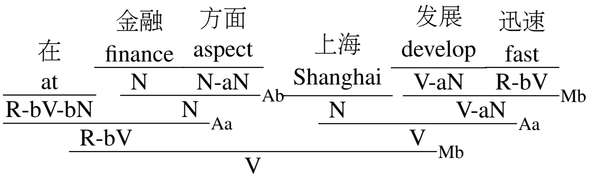
The modifier composition rules Ma and Mb establish a ‘1’-labeled dependency from the modifier to the modificand. With argument and modifier composition rules, we can derive the Chinese sentence shown in (1).

(1) ‘Shanghai, in the aspect of finance, develops fast.’



The separate modifier composition rules in GCG make it possible to reuse modifier categories across different contexts. For example, in (1), 在金融方面 ‘in the aspect of finance’ is an adverbial modifier, having the category **R-bV**. It has the same category when the phrase is a sentential modifier as shown in (2). Consequently, 在 ‘at’ in both (1) and (2) has the same category **R-bV-bN**, which means it takes a succeeding nominal argument to become an adverbial modifier.

(2) ‘In the aspect of finance, Shanghai develops fast.’



In contrast, since CCG enforces a restricted set of inference rules, it needs to provide two different categories, (S\NP)/(S\NP)/NP and S/S/NP for 在 in (1) and (2). In total, Chinese CCGbank

[3] Following (Nguyen et al., 2012), directional operators such as forward and backward slashes (‘\’ and ‘/’) are not used because some operators, such as gap operators in tough constructions, are undirected.

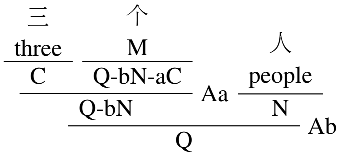
has 91 different categories for 在, since the prepositional phrase headed by 在 can modify constituents of various syntactic categories. In contrast, Chinese GCG annotations only have 9 different categories for 在.

Another example of differing lexicalization is the category of the tense aspect 了 in Chinese, which can either occur immediately after a verb or after the whole verb phrase to indicate past tense. Although generalized backward crossed composition (Steedman, 2000) helps aspect/tense particles in Chinese usually retain their canonical category (S\NP)\(S\NP), there are still 59 different categories for 了 in Chinese CCGbank (Tse and Curran, 2010), and most of them are semantically indistinguishable.

3.3 Nominal and quantificational expressions

Mandarin Chinese does not have determiners such as ‘the’ or ‘a’ in English, so there is no empirical motivation to distinguish NP and N categories. However, *classifiers* or *measure words*, glossed as ‘M’ in (3), are obligatory when a noun is quantified by a number. This unit of measure is needed for quantification of nouns because the lack of number morphology in Chinese makes all nouns behave as mass nouns (Allan, 1977; Borer, 2005).

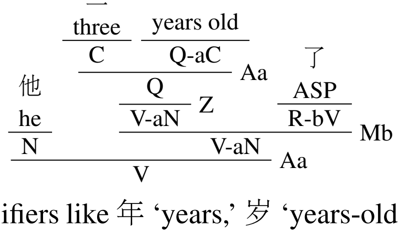
(3) ‘three people’



We propose a separate category Q for quantificational expressions because they can be predicative, as in (4), which makes them different from common nouns. A zero-head rule Z, where $c, d, e \in C$, converts the Q category to V-aN to make the quantificational expression predicative.

$$e{:}g \Rightarrow c\text{-}\mathbf{a}d{:}\lambda_x\,(\mathbf{f}_0\,x){=}\mathsf{pred} \wedge g\,(\mathbf{f}_2\,x) \qquad \text{(Z)}$$

(4) ‘He is three.’

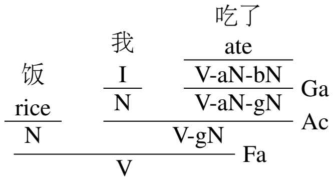


Classifiers like 年 ‘years,’ 岁 ‘years-old’ and 天 ‘day’ already contain the nominal information, so they do not require nominal arguments like other classifiers. Classifiers of this type have a different category ‘Q-aC’ to reflect this combinational difference. By doing so, the numbers receive the same category C in both 三天 ‘three days,’ and 三个人 ‘three people.’ However, in both Chinese Treebank and CCGbank, the category ‘M’ is used for both types of classifiers, which results in numbers like 三 ‘three’ having the category QP/M in ‘three days’ and the category (NP/NP)/M in ‘three people’ in Chinese CCGbank. This is not desirable because it expends training examples on an artificial distinction between the numbers 三 ‘three’ in each of these expressions, which are semantically the same.

3.4 Topicalization

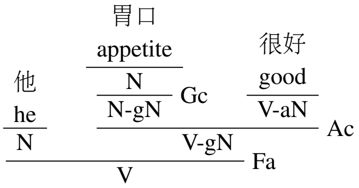
Topicalization in Mandarin can involve either movement of a topicalized constituent or not. The topicalization which involves movement is similar to that of English, in which the object is usually moved to the sentence initial position and a gap is left behind, as shown in (5).[4]

(5) ‘The rice, I ate.’



The non-movement topicalization occurs much more frequently in Mandarin, in which the subject of the sentence usually has an ‘association’ relation to the topic, as shown in (6).

(6) ‘Of him, the appetite is good.’



The referent of the subject in the non-movement topicalization needs to be further specified by the topic. Although topics are seen to be associated with other constituents of the sentence, especially in colloquial expressions, only associations with subjects are observed in the Treebank data. Therefore in our analysis of this type of topicalization, the subject undergoes a unary type conversion from N to N-gN to introduce a gap, which is

[4] The verb 吃 and the tense particle 了 are separate tokens, shown together here to simplify the derivation. We apply the same simplification to 很好 in following examples.

later discharged by the topic to capture the 'association' relation between the subject and the topic.

Inference rules for gap composition are:

$$p\varphi_{1..n\text{-}1}\,o\,c \Rightarrow p\varphi_{1..n\text{-}1}\textbf{-g}c{:}\lambda_{vx}\,(g\,x) \wedge (\mathbf{f}_n\,x){=}v \quad \text{(Ga)}$$

$$c{:}g \Rightarrow c\textbf{-g}d{:}\lambda_{vx}\,(g\,x) \wedge (\mathbf{f}_1\,v){=}x \quad \text{(Gb)}$$

$$\text{N}{:}g \Rightarrow \textbf{N-gN}{:}\lambda_{vx}\,(g\,x) \wedge \exists_e\,(\textbf{de-asso}\,e\,x\,v) \quad \text{(Gc)}$$

where $p \in P$, $o \in \{\textbf{-a}, \textbf{-b}\}$, $c \in C$, $d \in \{\textbf{A-bN}, \textbf{R-bV}\}$ and $\varphi \in \{\textbf{-a}, \textbf{-b}\} \times C$. Rule Ga hypothesizes a gap as a preceding or succeeding argument, rule Gb hypothesizes a nominal or adverbial modifier gap and rule Gc hypothesizes a gap which is associated with the subject in topicalization.

Non-local arguments, each consisting of a non-local operator and argument category $\psi \in \{\textbf{-g}\} \times C$, are then propagated to consequents from all possible combinations of antecedents. For $d{:}g\;\; e{:}h \Rightarrow c{:}(f\,g\,h) \in \{\text{Aa–b, Ma–b}\}$:

$$d\psi_{1..m}{:}g \;\; e\psi_{m+1..n}{:}h \Rightarrow$$
$$c\psi_{1..n}{:}\lambda_{v_{1..n}}\,f\,(g\,v_{1..m})\,(h\,v_{m+1..n}) \quad \text{(Ac–d, Mc–d)}$$

Rules Ac–d and Mc–d stipulate non-local propagation through argument and modifier composition.

Inference rules for filler attachment apply gapped clauses to topicalized phrases as fillers. For $c \in C$, and $p \in P$:

$$p{:}g \;\; c\textbf{-g}p{:}h \Rightarrow c{:}\lambda_x\,\exists_y\,(g\,y) \wedge (h\,y\,x) \quad \text{(Fa)}$$

In contrast, Tse and Curran (2010) analyze the topic in non-movement topicalization as a sentential modifier, which gives 他 'he' in (6) the category S/S, serving as a sentential modifier for the sentence 胃口很好 'appetite is good.' This analysis conflates sentential adverbial modifiers such as 'today' with topics such as 'he' in (6), yielding incorrect dependencies and expending probability mass on ungrammatical derivations (e.g.with topics conjoined with adverbs).

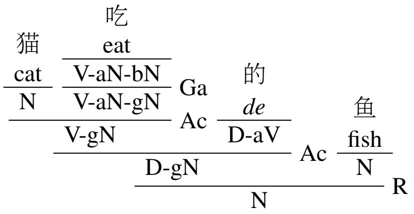
3.5 Relative and appositive clauses

In Mandarin relative clauses, the particle 的 'de' takes a preceding clause containing a gap to form a relative clause modifying a succeeding noun. The modified noun is the filler of the gap in the relative clause. The inference rules for relative clauses apply the gapped *de*-clause to the modificand as a filler. For $c \in C$:

$$\textbf{D-g}c{:}g \;\; \textbf{N}{:}h \Rightarrow \textbf{N}{:}\lambda_x\,(h\,x) \wedge \exists_y\,(g\,x\,y) \quad \text{(R)}$$

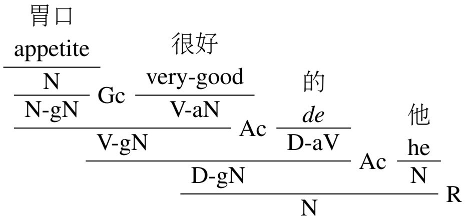
A GCG analysis of a relative clause with an object gap is shown in (7).

(7) 'fish that cats eat'



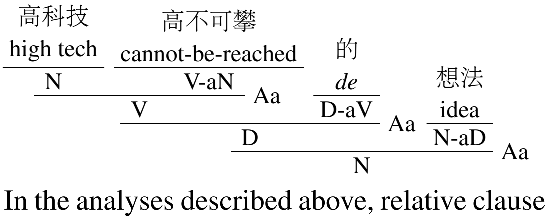
Our analysis of topicalization in (6) makes it easy to account for a relative clause which relativizes a topic. In (6) for example, relativizing the topic 他 'he' yields a nominal phrase containing a non-restrictive relative clause 胃口很好的他, 'he whose appetite is good.' A GCG analysis of this nominal phrase is shown in (8).

(8) 'he whose appetite is very good'



Appositive clauses in Mandarin Chinese are formed with the same 的 '*de*' particle used in relative clauses. However, unlike relative clauses, appositive clauses do not involve any gap constituent. In this GCG analysis of appositive clauses, 的 '*de*' receives the same category as it does in relative clauses. But the noun which takes an appositive clause as complement has the category **N-aD** to take a preceeding *de*-clause to further specify the content of the noun. An appositive clause in this grammar is shown in (9).

(9) 'the idea that high tech cannot be reached'



In the analyses described above, relative clauses with different types of gaps are differentiated, and relative clauses in general receive different analyses than appositive clauses. In the analysis of Tse and Curran (2010), a relative clause can only have either a subject or object gap in Chinese. Relative clauses that relativize topics receive the same categories as appositive clauses. This analysis blurs the distributional difference between certain types of relative clauses and appositive clauses, decreasing PCFG estimates of both types of relative clauses given the same (conflated) category.

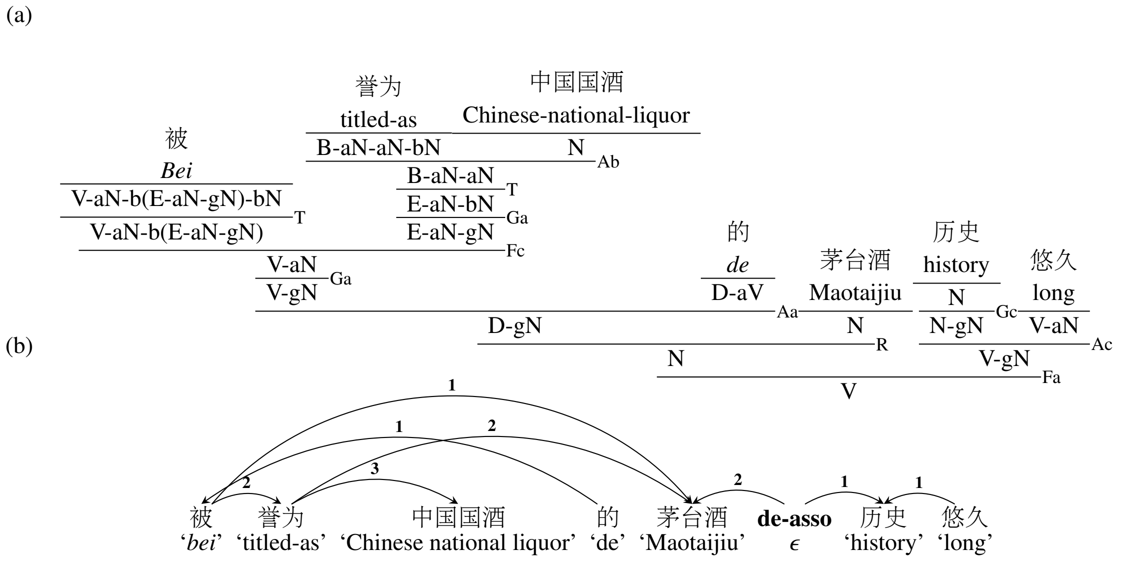
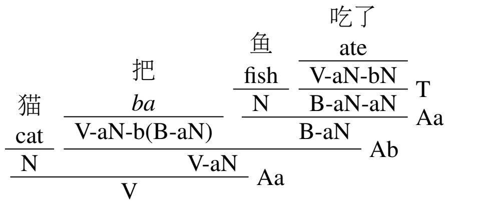


Figure 1: GCG derivation:"Maotaijiu, which is titled as the Chinese national liquor, has a long history" (a) and its associated dependencies (b)

3.6 *Ba* and *bei* constructions

Ba constructions in Mandarin Chinese require the affected patients of certain verbs to occur before the verb, instead of after the verb. For example, 鱼 'fish' in (10) is the object of 吃 'eat' and it occurs before the verb 'eat.' In the Penn Treebank, 把 *ba* takes a clause as argument. Therefore, 鱼吃了 'fish ate' in (10) is analyzed as a clausal complement of *ba*. This analysis makes 'fish' the subject of the verb 'eat,' instead of the object. Consequently, for example, Stanford dependencies extracted from Treebank annotations of this sentence have both 'nsubj (吃'eat,' 猫'cat')' and 'nsubj (吃'eat,' 鱼'fish'),' which is not correct.

(10) 'the cat ate the fish'



In our analysis, we propose that the particle *ba* takes a *ba*-verb as its complement. *Ba*-verbs are derived from transitive verbs with the type conversion rule given below.[5]

$$c{:}g \Rightarrow d{:}g \qquad \text{(T)}$$

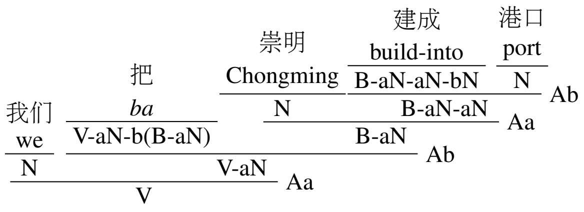
Using the type conversion rule T, we change a transitive verb **V-aN-bN** to **B-aN-aN** to capture the fact that the verb that occurs within a *ba* construction takes a preceeding second argument.

The particle 把 *ba* is assigned the category **V-aN-b(B-aN)**, with coindexation between the referent of its subject ($\mathbf{f}_1\ x$) and the referent of the subject of its complement ($\mathbf{f}_1$ ($\mathbf{f}_2\ x$)).

$$\text{把 'ba'} \mapsto \textbf{V-aN-b(B-aN)}{:}\lambda_x\,(\mathbf{f}_0\ x){=}\mathsf{ba} \;\wedge\; (\mathbf{f}_1\ x){=}(\mathbf{f}_1\ (\mathbf{f}_2\ x))$$

Usually, the affected patient is the direct object of a transitive verb, as shown in (10), but there are cases where some verbs can only occur in *ba* constructions or *bei* constructions. These types of verbs are *ba*-verbs to begin with and do not need to be changed from transitive verbs. They are given the general category **B-aN-aN-bN**. Many resultative verbs (VRD, in treebank annotation) have this category. An example is given in (11).

(11) 'We built Chongming into a port.'



Mandarin Chinese uses the particle 被 *bei* to construct passive sentences. In *bei* constructions,

[5] This rule is constrained by fact that the function *g* is preserved, and its usage is constrained by parsing probabilities for particular categories. Following Featherston (2005) and Crocker and Keller (2005), the model described in this paper assumes that grammaticality judgements are gradient and determined by probabilities of compositional inferences occurring in the experience of a particular language user.

the patient argument of a verb, usually the second argument of a transitive verb or a *ba*-verb, is moved to the subject position of the clause. We propose the particle 被 *bei* takes a *bei*-verb as its complement. *Bei*-verbs, which are of the category **E-aN-gN**, are derived from **E-aN-bN** by introducing a gap by rule Ga. **E-aN-bN** is derived by the type conversion rule T from **V-aN-bN** or **B-aN-aN**, transitive verbs or *ba*-verbs. Here is the lexical entry we propose for the *bei* particle.

$$\text{被 `bei'} \mapsto \textbf{V-aN-b(E-aN-gN)-bN}: \quad \lambda_x\,(\mathbf{f}_0\ x){=}\textsf{bei} \wedge (\mathbf{f}_3\ x){=}(\mathbf{f}_1\ (\mathbf{f}_2\ x))$$

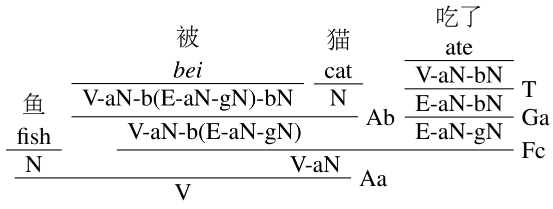
The lexical entry of 被 *bei* stipulates that the first argument of *bei* is the subject of its second argument, the VP complement, **E-aN-gN**. Since the agent in the passive voice construction is optional (as it is in passive voice in English), the category of the *bei* particle can have a type change from **V-aN-b(E-aN-gN)-bN** to **V-aN-b(E-aN-gN)**. The inference rule (Fc) is proposed for the composition of gap dependencies contained within succeeding arguments, where $p \in P$, $\varphi \in \{\textbf{-a}, \textbf{-b}\}{\times}C$, and $\psi \in \{\textbf{-g}\}{\times}C$.

$$p\varphi_{1...n-1}\textbf{-b}(d\psi){:}g \;\; d\psi{:}h \Rightarrow p\varphi{:}\lambda_x\, h\,(\mathbf{f}_1\,x)\,(\mathbf{f}_n\,x) \wedge g\,x \quad \text{(Fc)}$$

Using rule Fc, the first argument of the *bei* particle becomes the filler of the gap in the *bei* verb. This rule also supports an analysis of tough constructions in Chinese.

An example *bei*-construction which contains a transitive verb is shown in (12).

(12) 'The fish was eaten by the cat.'



The Penn Treebank uses the category 'LB' for the *bei* particle where the optional agent argument occurs, and 'SB' for the *bei* particle where it is elided. Tse and Curran (2010) follow the Treebank annotation, proposing two different categories for the *bei* particle. For example the CCG category for 'LB' is $(\text{S}\backslash\text{NP}_y)/(\text{S}\backslash\text{NP}_x/\text{NP}_y)/\text{NP}_x$, in which a coindexation scheme is used to ensure that the subject of *bei* is coindexed with the object of its verbal complement. The *ba* particle, with the category $(\text{S}\backslash\text{NP}_y)/(\text{S}\backslash\text{NP}_y/\text{NP}_x)/\text{NP}_x$, is different from *bei* only in the coindexing scheme. However, if the passivized verb is not a transitive verb, such as 誉为 'titled-as' in Figure 1, it is hard to infer what the coindexing scheme should be like in the CCG analysis.

Figure 1 shows a GCG derivation of a sentence from the Chinese Treebank. We use this sentence to illustrate how topicalization, passive voice, and relative clauses are analyzed in the GCG framework and what kind of dependencies we can extract from GCG derivations.

4 Experiments

We use a set of reannotation rules similar to those described by Nguyen et al. (2012) to reannotate the Penn Chinese Treebank into GCG trees. These reannotation rules work within a perl script that traverses each bracketed sentence in the Penn Chinese Treebank by selecting each pair of matching brackets from the top of the tree to the bottom, then running a sed-like pattern substitution rule on each selection. With around 200 annotation rules, we currently fully annotate 71% of sentences (18,505 sentences out of 26,062) from the Penn Chinese Treebank 5 and 6.

In order to evaluate the Chinese GCG annotations in terms of parsing accuracy, we compare the parsing performance of a latent-variable parser trained on Chinese GCG annotations with that of the same parser trained on Chinese CCG annotations. The Chinese CCGbank is obtained by converting the Penn Chinese Treebank into CCG annotations according to Tse and Curran (2012).[6] We divided the fully annotated sentences in both grammars into training, development and test sections according to the section divisions suggested by Tse and Curran (2012). In order to have a better understanding of how the parsing performance changes with the size of the training data, we trained the Chinese CCG parser on both the full training set (ccg.full) and the same training set used for training the Chinese GCG parser (ccg.same). The detailed section divisions are shown in Table 1.

For the two CCG parsers, ccg.full and ccg.same, we use the Petrov and Klein (2007) latent variable PCFG trainer, with 5 split-merge cycles, which is the best setting indicated by Tse and Curran (2012). As with CCG, we ran the Petrov et al.

[6] `https://github.com/jogloran/cnccgbank`

Model	Train	Dev	Test
ccg.full	22680	689	1986
ccg.same	13677	689	1986
gcg	13677	689	1986

Table 1: Train/Dev/Test Split

	R	P	F	tag
ccg.same	78.64	78.96	78.80	85.62
ccg.full	80.69	81.13	80.91	87.24
gcg	82.70	83.86	83.28	93.65

Table 2: Parsing results on the development set

	R	P	F	tag
ccg.same	78.39	78.55	78.47	85.02
ccg.full	79.77	79.93	79.85	86.33
gcg	82.19	83.07	82.63	93.66

Table 3: Parsing results on the test set

		% Err. Reduct. vs.		p-value vs.	
	F1	ccg.same	ccg.full	ccg.same	ccg.full
ccg.same	88.76	–	–	–	–
ccg.full	89.39	–	–	–	–
gcg	**90.07**	11.65	6.409	0.0007	0.04

Table 4: Parsing results, error reduction ratios and significance testing results on the common test set of NoUnary+NoLab trees.

(2006) latent-variable PCFG trainer on the GCG-reannotated training corpus. The PCFG trainer was used 'off the shelf' and run with its default parameters, only varying the number of split-merge iterations on the development section. We found 5 split-merge iterations yielded the best parsing performance in the development section.

Tables 2 and 3 show the parsing performance of the parsers on the development and test sets. The parsing results show that a larger training set is beneficial to the parsing performance of the Chinese CCG parer; the parsing performance of the CCG parser trained on the full training set performs consistently better than the parser trained on 71% of the training set. The GCG parser, trained on 71% of the training set, seems to parse reasonably well even compared with the CCG parser trained on the full training set. It is worth noting that the GCG parser is much higher in tagging accuracy than the CCG parser, which supports our hypothesis that the CCG parser might suffer from sparse data problems.

However, direct comparison of the parsing performance of these two parsers is not fair because these two grammars define different categories and different tree structures. In order to ensure a fair comparison between these grammars, it is necessary to have them produce exactly the same target representation. In this experiment, we test the parsing performance of these two grammars on a common test set of sentences to which the two grammars assign the same tree structure when syntactic labels and unary branches are removed, see Figure 2. We found 984 sentences in the test set which have exactly the same unlabeled binary structures (Figure 2c) in both grammars.

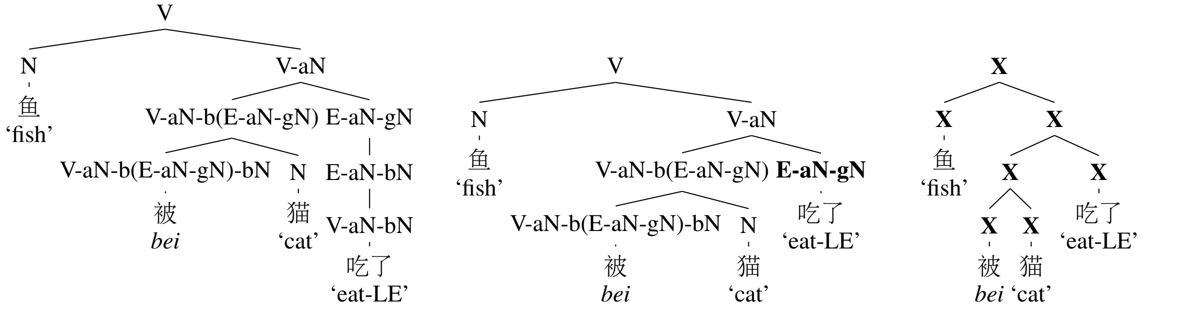
Table 4 shows the parsing results (F1) on parses with both syntactic category labels and unary branches removed (NoUnary+NoLab). After removing unary branches, the parses have exclusively binary tree structures and have identical results for precision, recall and F1 in parsing evaluations. Since both grammars predict exactly the same binary tree structures with exactly the same ('X') categories, significance testing is performed on these predictions using bootstrap resampling.

Results in Table 4 show that the parsing performance of the Petrov and Klein (2007) parser trained on the GCG-reannotated corpus is more accurate with strong significance ($p < 0.001$) than the same parser trained on the CCG-reannotated corpus of the same size. We observe a significant improvement ($p < 0.05$) of the GCG parser over the CCG parser trained on the full training set.

We believe that the Chinese CCG parser suffers from data sparsity effects. Excluding those words which are only associated with one preterminal category, the lexical-categorial confusion rate is 3.45 for the Chinese CCG annotations and 2.59 for the Chinese GCG annotations, which is also reflected in the large gap (more than 5 points) between their tagging accuracy. Enforcing a small set of language-independent inference rules in the Chinese CCG-annotations might have some formal appeal, but it leads to a large set of syntactic categories, many of which, such as nominal or adverbial modifiers, are syntactically or semantically indistinguishable. Since the GCG described in this paper uses a larger set of inference rules and consequently fewer category labels, it suffers fewer sparse data effects.

5 Conclusion and discussion

This paper has described a generalized categorial grammar for Mandarin Chinese, reannotated



(a) A full tree (b) A NoUnary Tree (c) A NoUnary+Nolab Tree

Figure 2: Constructing common test set

from the Penn Chinese Treebank. Unlike previous efforts using strongly lexicalized CCG (Tse and Curran, 2010), the reannotated corpus described in this paper adopts a policy of moderate lexicalization, allowing both inference rules and lexical categories to be language-specific. This moderation offers considerable representational freedom, particularly in modeling Chinese *ba*-, *bei*-, and *de*-constructions, which make substantial use of unbounded dependencies. Experimental results appear to show that, while there may be some formal appeal to a small universal set of language-independent combinators (Steedman, 2000; Steedman, 2002; Steedman, 2012), the large category set resulting from it might impose an empirical cost for parsing tasks.

The reannotation rules are available at `http://www.sourceforge.net/projects/modelblocks`.

References

Keith Allan. 1977. Classfiers. *Language*, 53:285–311.

Emmon Bach. 1981. Discontinuous constituents in generalized categorial grammars. *Proceedings of the Annual Meeting of the Northeast Linguistic Society (NELS)*, 11:1–12.

Hagit Borer. 2005. *Structure sense*. Oxford.

Matthew W. Crocker and Frank Keller. 2005. Probabilistic grammars as models of gradience in language processing. In Gisbert Fanselow, Caroline Féry, Ralph Vogel, and Matthias Schlesewsky, editors, *GRADIENCE IN GRAMMAR: GENERATIVE PERSPECTIVES*. University Press.

Sam Featherston. 2005. Universals and grammaticality: wh-constraints in German and English. *Linguistics*, 43(4):667–711.

Lauri Karttunen. 1989. Radical lexicalism. In M. R. Baltin and A. S. Kroch, editors, *Alternative Conceptions of Phrase Structure*, pages 43–65. University of Chicago Press, Chicago.

Jonathan K. Kummerfeld, Daniel Tse, James R. Curran, and Dan Klein. 2013. An empirical examination of challenges in chinese parsing. In *Proceedings of ACL'13*, pages 98–103, Sofia, Bulgaria.

Luan Nguyen, Marten van Schijndel, and William Schuler. 2012. Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of COLING '12*, pages 2125–2140, Mumbai, India.

Richard T. Oehrle. 1994. Term-labeled categorial type systems. *Linguistics and Philosophy*, 17(6):633–678.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of NAACL HLT 2007*, pages 404–411, Rochester, New York, April. Association for Computational Linguistics.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of COLING/ACL'06*.

Laura Rimell, Stephen Clark, and Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of EMNLP 2009*, volume 2, pages 813–821.

William Schuler and Adam Wheeler. 2014. Cognitive compositional semantics using continuation dependencies. In *Third Joint Conference on Lexical and Computational Semantics (*SEM'14)*.

Mark Steedman. 2000. *The syntactic process*. MIT Press/Bradford Books, Cambridge, MA.

Mark Steedman. 2002. Plans, affordances, and combinatory grammar. *Linguistics and Philosophy*, 25.

Mark Steedman. 2012. *Taking Scope - The Natural Semantics of Quantifiers*. MIT Press.

Daniel Tse and James R. Curran. 2010. Chinese CCGbank: extracting CCG derivations from the penn chinese treebank. In *Proceedings of COLING '10*, pages 1083–1091.

Daniel Tse and James R. Curran. 2012. The Challenges of Parsing Chinese with Combinatory Categorial Grammar. In *Proceedings of NAACL-HLT '12*, pages 295–304, Montréal, Canada.

Nianwen Xue, Fei Xian, Fu-Dong Chiou, and Martha Palmer. 2005. The Penn Chinese Treebank: Phrase Structure annotation of a large corpus. *Natural Language Engineering*, 11:207–238.