

Research Article

Parsing Chinese with Combinatory Categorical Grammar: A Linguistic and Computational Study

Haixia Man ¹, Manyi Wan ², Yunbao Shi ³ and Peng Chen ⁴

¹School of Foreign Studies, University of Science and Technology Beijing, Beijing 100083, China

²Biquan Academy, Xiangtan University, Xiangtan 411105, China

³School of Marxism, West Anhui University, Lu'an 237012, China

⁴School of Information, Beijing Language and Culture University, Beijing 100081, China

Correspondence should be addressed to Manyi Wan; 925154476@qq.com

Received 28 January 2022; Revised 28 July 2022; Accepted 20 August 2022; Published 30 September 2022

Academic Editor: Ning Cai

Copyright © 2022 Haixia Man et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Parsing Chinese language with CCG is very difficult because the architecture and assumptions of CCG do not fit well with facts from Chinese. Based on the concept of “realization” proposed by Zhu Dexi (1920–1992), this study sheds light on the discrepancy between CCG and Chinese syntax and puts forward a refined schema for Chinese compositionality. The discussion is supported by the data of Chinese CCGbank (CASS). Furthermore, by activating a function-based category setting and a noun/verb disambiguating tagging mechanism, we develop a rule-based mini-Chinese CCG parser without deep learning. The new NVN parser surpasses existing Chinese CCG parser C&C in parsing effect (LF 85.9 vs. LF 74.6) on a partial PCTB 6.0 test set of 500 sentences.

1. Introduction

Combinatory categorical grammar is a mildly context-sensitive grammar formalism that links syntactic derivation with semantic composition in the closest possible relation [1, 2]. Through the development of efficient and accurate broad-coverage parsers [3–5], CCG has become one of the most widely used grammar-based formalisms in the field of computational linguistics. However, few works on Chinese CCG parsing could be found. There is only one attempt to train CCG parsers on Chinese CCGbank so far [6] possibly because of the not-so-well-understood nature of Chinese syntax.

In order to identify the challenges of Chinese CCG parsing, Tse and Curran manipulate parser architecture and annotation decisions of fixed corpus, discovering that collapsing categorial distinctions in Chinese CCGbanks, like bare/nonbare NP and NP/localizer, can yield less ambiguous corpora and thus increase parsing accuracy [6]. Nevertheless, major challenges of Chinese parsing in general [7, 8], noun/verb ambiguity and argument-drop (We adopt “argument-drop” in present work, rather than “pro-drop” in [6–8] because we do not want to make any commitment in

generative sense that the dropped arguments are agreeing pronouns or NPs. Chinese has no agreement, and it allows arguments in almost all positions to be dropped. The dropped arguments are more likely to be the topics.) in particular, still linger. What is worth noting is that the parsing ambiguities invoke arbitrary label-rewriting choices in CCG derivations (as the NP is rewritten into S/S in Figure 1).

In Figure 1, the NP category of the topicalized constituent does not coincide with its syntactic function of a sentential premodifier (S/S), thus enforcing a category conversion manifested by unary phrase structure rule $NP \rightarrow S/S$. Resorting to such unary rules is an effective way to prevent overgenerations caused by form-function distinctions [9], a phenomenon distinctive and substantial to Chinese CCG parsing.

Man and Zou calculate the rules used in Chinese CCGBank (CASS) (<https://www.ccgbank.net/>), uncovering a surprisingly high percentage of unary rules that even outnumbers the three sets of compositional rules (composition, type-raising, and substitution) in total [10]. Though it is not unfamiliar that form-function-distinction-caused

tagging errors are major challenges in Chinese parsing [6–8], no one has explored the reasons behind.

Based on Zhu’s idea of “realization” [11] essential to comprehending form-function distinctions in Chinese, this paper endeavors to explore the niche that CCG and Chinese syntax assume on surface structure, hopefully shedding light on the nature of Chinese-English parsing gap. It then proposes a rule-based CCG parser that integrates a noun/verb disambiguating tagging mechanism and a syntactic-function-based category setting, enabling a higher LF score at 85.9. The new parser surpasses C&C parser (74.6) that uses supertagger [6], with the potentiality to significantly save the cost of supertagger training.

2. Combinatory Categorial Grammar

A CCG is a deductive system that contains two components: its categorial lexicon and a set of combinatory rules.

2.1. Two Components. The categorial lexicon defines lexical items of the language into triplets in the form of $\sigma|\phi: \lambda$, where σ is the phonological form, ϕ is the syntactic categories, and λ its semantic formula, as shown in (1):

- (1) John | NP: j' ; go to the Arctic | S\NP: λx .
go_arctic'(x) eats | (S\NP)/NP: $\lambda x \lambda y$.eat'xy; apples |
NP: a'

The categories in CCG are either atomic (for example, NP for noun phrase and S for sentence) or functional (for example, S\NP for intransitive verb and (S\NP)/NP for transitive). For functional categories, CCG uses a notation in which the result of the range category always appears to the left of the slash and the argument category to the right [1, 2]. Thus, the category (S\NP)/NP of “eats” specifies that it will only result in a grammatical sentence S when it first associates with an NP to its right (indicated by the direction of the slash), obtaining an S\NP category and then another NP to its left.

The second component is combinatory rules for combining functions and arguments and also functions and functions. Functional applications (F) in (2) are core rules for basic categorial grammar (BCG) of Ajdukiewicz-Bar-Hillel tradition [12, 13]. They can combine functional categories with their argument categories as the CCG derivation of “John eats apples” as exhibited in Figure 2:

- (2) Forward application ($>$): $A/B: \lambda x.Fx$ $B: a \rightarrow A: Fa$
Backward application ($<$): $B: a$ $A \setminus B: \lambda x.Fx \rightarrow A: Fa$

The arrows $>$ and $<$ mnemonically indicate which versions of functional application is applied by pointing at the directionality of its argument combination. The underlines in the categorial derivation, coupled with semantic interpretation, indicate combination via two functional application rules, first resulting the intransitive VP “eats apples” and then the whole sentence (The agreement values and the like will be ignored for present purpose. It will not be

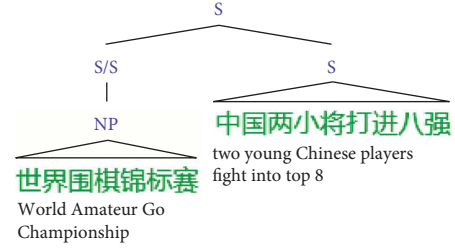


FIGURE 1: Derivational ambiguity.

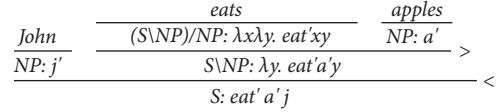


FIGURE 2: CCG derivation of “John eats apples.”

used in the case of our target language Chinese because of its lack of morphological features).

In addition to F, CCG includes type-raising (T), composition (B), and substitution (S) into BCG (we call them CCG rules hereinafter), increasing the expressiveness to mildly context-sensitive while preserving syntax-semantic transparency in the meantime ([14] (p. 74)). The three sets of rules differ from F in that they can operate on functional categories.

- (3) Forward type-raising ($>T$): $X: x \rightarrow T/(T \setminus X): \lambda f.fx$
Backward type-raising ($<T$): $X: x \rightarrow T \setminus (T/X): \lambda f.fx$
(4) Forward composition ($>B$): $X/Y: \lambda x.fx$ $Y/Z: \lambda x.gx \rightarrow X/Z: \lambda x.f(gx)$
Backward composition ($<B$): $Y \setminus Z: \lambda x.gx$ $X \setminus Y: \lambda x.fx \rightarrow X \setminus Z: \lambda x.f(gx)$
Forward crossed composition ($>\times B$): $X/Y: \lambda x.fx$ $Y \setminus Z: \lambda x.gx \rightarrow X \setminus Z: \lambda x.f(gx)$
Backward crossed composition ($<\times B$): $Y \setminus Z: \lambda x.gx$ $X \setminus Y: \lambda x.fx \rightarrow X \setminus Z: \lambda x.f(gx)$
(5) Forward substitution ($>S$): $(X/Y)/Z: \lambda x \lambda y.fxy$ $Y/Z: \lambda x.gx \rightarrow X/Z: \lambda x.fx(gx)$
Forward crossed substitution ($>S \times$): $(X/Y) \setminus Z: \lambda x \lambda y.fxy$ $Y \setminus Z: \lambda x.gx \rightarrow X \setminus Z: \lambda x.fx(gx)$
Backward substitution ($<S$): $Y \setminus Z: \lambda x.gx$ $(X \setminus Y) \setminus Z: \lambda x \lambda y.fxy \rightarrow X \setminus Z: \lambda x.fx(gx)$
Backward crossed substitution ($<S \times$): $Y \setminus Z: \lambda x.gx$ $(X \setminus Y) \setminus Z: \lambda x \lambda y.fxy \rightarrow X \setminus Z: \lambda x.fx(gx)$

CCG follows traditional categorial practice here in assuming that categories are short for categories of parts of speech ([12, 13, 15] etc.). To be specific, it assumes that there are enough parts of speech (pos for short) descriptions to categories and there is a one-one correspondence between category and pos. Thus, we can penetrate through existing studies over Chinese pos, or word classes in many works, to uncover the mysteries and challenges behind Chinese CCG derivation.

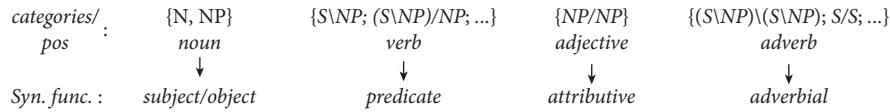


FIGURE 3: Mapping between categories/pos and syntactic functions.

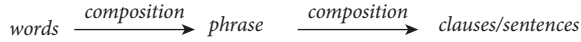


FIGURE 4: Sketch of CCG’s compositionality.

2.2. *Some Insights.* Before moving on to a discussion over complications of Chinese syntax, some insights on categories and compositional rules may help us to locate the discrepancies in CCG’s assumption and assumptions made in studies of Chinese syntax.

When a category participates in a syntactic derivation in lieu of the lexical item it is assigned to, the category is expected to play the syntactic function its corresponding pos is supposed to play within that syntactic construction. Let us notice immediately the systematic mappings assumed in CCG among categories/pos and syntactic functions (Figure 3). There is an into function from the set of categories or pos to the set of syntactic functions, outlawing one-to-many mapping from pos set to the set of syntactic functions, meaning, for example, a verb cannot play any syntactic roles other than a predicate.

Another crucial assumption of CCG is that all syntactic rules in CCG are syntactization of combinators in combinatory logic [16, 17]. They are basically functions manifesting expressible semantic dependencies observable through syntactic derivations among the categories of constituents. The syntax-semantics transparency is made clear by *The Principle of Combinatory Type Transparency* in (6):

(6) *All syntactic combinatory rules are type-transparent versions of one of a small number of simple semantic operation over functions.* [2].

Bozşahin [18] concludes (6) as a narrow claim of CCG that natural grammars are combinatory type-dependent, a property manifested in three aspects simultaneously—being a constituent, being derivable, and being immediately interpretable. This is neat. However, it should be understood that the neatness relies on three default settings behind:

- (7) (a) Syntactic categories reveal the syntactic dependencies (or syntactic relation) between constituents
- (b) A syntactic category will play the required syntactic function as shown in Figure 3 when participating in syntactic derivations
- (c) All syntactic rules are compositional in Frege’s sense, and they scaffold a part-to-whole derivation syntactically in a way exhibited in Figure 4

Among the three clauses in (7), a violation of (b) will result in form-function distinction, which may in turn affect the derivational approach in (c). Considering these three steps, Chinese syntax presents a picture more complicated

than those in Figures 3 and 4. In the section to come, we will outline more intricately a categorial mechanism in light of the findings made by Chinese grammarians. We will argue that those observable from surface structure of Chinese tell a different story, and the complications of Chinese syntax can be accounted for by a form-function unification strategy called “realization” and a refined differentiation of “composition” and “realization”.

3. Complications of Chinese Syntax under CCG Lens

3.1. *Category Ambiguities: A Story between Pos and Syntactic Functions.* The story of category ambiguities begins with the mysteries of Chinese pos. Tse and Curran [6], as well as other works on Chinese parsing [7, 8], find pos ambiguities (especially verb/noun ambiguity), a distinctive error type rare in English parsing. It directs to the fact that a linguistic element in Chinese is usually ambiguous between different parts of speech, in which case the actual pos of the element is determined only when it enters into an actual construction. This process is referred to as “realization” by Zhu ([11] (p. 74–5)).

Take the intransitive verb phrase “去北极” (go to the Arctic) as an example. In Chinese, it can form a larger VP with another intransitive (“探险”) as in (8) or be a predicate in a complete sentence as in (9), where its verbal nature is preserved:

(8) 去北极 探险
go the Arctic explore
‘to go to the Arctic and explore’

(9) 我 能 去北极
I can go to the Arctic
‘I can go to the Arctic’

Yet, such a verb phrase can also be a subject, a modifier in *de* construction, or even stand alone as an individual answer as shown in (10)–(12). Bearing no morphological changes. It seems that the same verbal phrase is capable of taking the functions that are usually performed by nouns, adjectives, and sentences. Thus, pos does not bear a one-one correspondence with syntactic functions in Mandarin Chinese. According to Zhu, the process that a VP functions as a predicate, a nominal modifier, or a sentence is a process called “realization”, in which a word or phrase of particular part of speech realizes into an actual part of sentence (or even a stand-alone sentence):

(10) 去北极 探险
go the Arctic explore

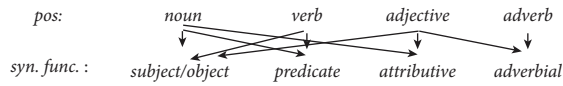


FIGURE 5: Mapping between pos and syntactic functions in Chinese.

(11) ‘To go to the Arctic is my dream’

我 能 去北极
I can go to the Arctic

(12) ‘All those who go to the Arctic are brave’

去北极 是 我 de 梦想
go to the Arctic is I de dream

‘To go to the Arctic’

In order to explain the dilemma, Zhu constructs a many-to-many mapping between pos and syntactic functions in Chinese (Figure 5), in contrast to the mapping for morphology languages such as English (Figure 4), which is one-to-one.

As the mapping indicates in Figure 5, there is no one-to-one correspondence between pos and syntactic functions in Chinese. We are stuck here, because acknowledging that an element belongs to a certain pos, for example to nouns, will result in flexibility of syntactic functions the element can play when it occurs in other positions other than subject/object, whereas endorsing a uni-functionality of a certain pos, for example believing that nouns can only function as subjects/objects, will engender a flexible pos system since verbs and adjectives should also be nouns when they appear as subjects/objects ([19, 20] etc.). This is how Chinese pos gets its fame of being flexible, as well as how form-function distinction in CCG arises.

Likewise, category ambiguity is inevitable in CCG when, for any element, its category assigned in the lexicon differs from the syntactic function it is expected to play in an actual derivation. In the case of (10) for instance, “去北极” of an S\NP category is expected to be an NP as it is in the position of subject, giving rise to a categorial version of noun/verb ambiguity:

(13) [去北极: S\NP]*_{NP}[是我的梦想]_{S\NP}

In order to maintain the strength of traditional pos system in conducting syntactic analysis, Zhu proposes the idea of “realization” to save pos-based syntactic architecture.

3.2. Two Kinds of Derivation

3.2.1. Composition versus Realization. Zhu distinguishes in [11] two different operations, namely, *composition* and *realization*, that are utilized to derive Chinese sentences. According to Zhu, a derivation of any Chinese sentence is composed of two phases (Figure 6), where words *compose* to get phrases and phrases *realize* as sentences, in contrast to a derivation in CCG that is compositional throughout

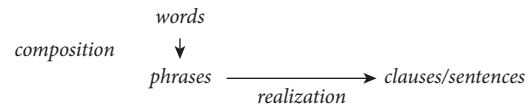


FIGURE 6: Zhu's compositionality of Chinese sentences.

both phases. According to Zhu, realization differs from composition in that it does not render a larger whole, but only bridges an abstract syntactic structure with an actual output when it is used in a real utterance.

3.2.2. A Refined Model for Zhu's Compositionality. Zhu's design is insightful but rough, constrained by lack of pragmatic studies during 1970s. It had been the mainstream idea that the surface structure of a language is propositional, with subject-predicate distinction being the basic binary structure. However, Chinese shows otherwise. It is often very tough to anchor the subject in Chinese because the preverbal constituents are not always the agent of the predicate verb, for example, “台上” in (14) and “一锅饭” in (15), and sometimes they do not bear certain semantic relations with the predicates, like “不下雨” in (16):

(14) 去北极 de 人 都 很勇敢
go to the Arctic de people all very brave

‘on the stage, there sits the presidium’

(15) Q: Where will you go?

A: 去北极。

‘A pot of rice can feed ten people’

(16) 台上 坐着 主席团
stage-on sit-prog presidium

‘It has not rained for three months’

With decades of heated discussion over subject-predicate distinction in Chinese, a consensus is reached—the surface structure of Chinese demonstrates an information structure (IS) based on topic-comment distinction, which, however, does not cling to the sentence's predicate-argument (PA) structure as languages with morphology do [21–25]. Thus, it is utterances, instead of sentences or clauses, that we see at surface structure. An utterance is more tolerant than a sentence in the structure it allows. PA structured sentences can be realized as topic-comment (TC) structured utterances, and in this case, an utterance is PA structured. When a phrase (or even a word if it has concrete meaning) is realized as a part of an utterance (a topic/comment) instead of forming the PA structure in the first place, the resulted TC structure may not comply with the PA structure (cf. (17)). Also, a phrase (or a content word), as pointed out by Zhu, can stand alone as an utterance by itself (cf. (13)). Hence, an utterance can but not necessarily be a function of its PA structure. Accordingly, we refine Chinese compositionality into Figure 7.

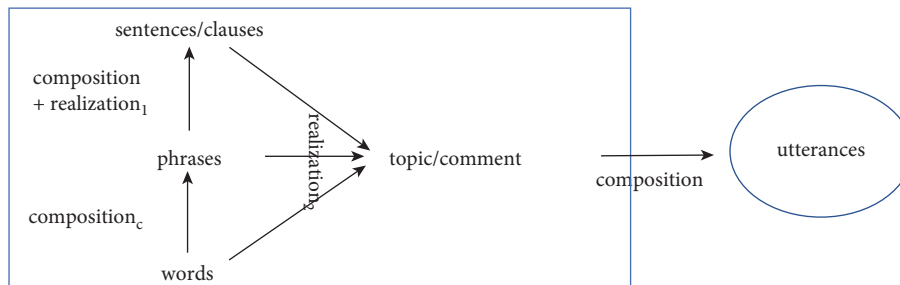


FIGURE 7: A refined compositionality of Chinese.

Compared with Figure 6, Figure 7 carefully peels utterances off from sentences/clauses and fills in the derivation details in between. In Chinese, it is the rightmost utterance that we see and hear, rather than the left-top sentences or clauses which relies on the PA structure. The processes in the left-hand box display how an utterance is derived in Chinese setting.

All arrows herein are transitive, so Figure 7 exhibits a mechanism with different paths to utterance derivation. We can get a multipath schema here because Chinese, being a language without case-marking or other forms of morphology, loosens the semantic restriction that predicate verbs impose on their arguments. Hence, instead of moving up to compose into a sentence, words or phrases can turn right directly and be realized as parts of utterances and then compose for a larger whole as long as the two constituents bearing an “aboutness” relation. This proffers a better explanation to the disparity of PA structure and IS structure in Chinese.

Besides, we put subscripts C, 1, and 2 on composition and realization to differentiate them from composition and realization in general. According to our discussion in Section 3.1, composition_C makes no commitment to the function, the linguistic element plays. The two functions, syntactic and pragmatic ones, are taken care of by realization₁ and realization₂, respectively. To put it in another way, for any linguistic item, a grammar of such kind indeed splits the role of its pos and the function it plays in actual utterances, either syntactic or pragmatic. Thus, realization provides a theoretical foundation to such form-function unification that takes effect in the form of unary rules.

Coming to the last part of this section, we will pay a visit to the data from Chinese CCGbank (CASS). The statistics will support the two conclusions we get through our discussion in Sections 3.1 and 3.2: (1) categories in Chinese ambiguate among three roles, namely, categories of pos, of syntactic function, and of pragmatic function; (2) realization is not an accidental phenomenon in Chinese CCG derivation, but an essential way to bridge the gaps of its category ambiguities.

3.3. *Chinese CCGbank (CASS) Data.* Chinese CCGbank (CASS) is converted automatically from Penn Chinese

Treebank (PCTB) 6.0 by following the algorithms in [26]. The derived corpus contains CCG derivations of 25,946 sentences and a lexicon of 46,085 words coupled with their syntactic categories. It inherits 7 primitive categories (Table 1) from PCTB.

Altogether, 2,483 CCG rules are applied by 577,668 times (this number is smaller than the frequency shown on the website (722,492 times) because the number there includes punctuation-absorbing rules and coordination rules, both of which are eliminated for present purposes in that the former is a technical operation bearing little relation with syntactic concatenation, and the latter can be reduced to application in two steps) (Table 2) for successful parsing. Among the totality of all rules being used, functional application takes the lead at nearly 92%, and the rest 8% is divided almost equally among CCG rules (B, T, and S altogether) and non-CCG rules (NCR). The fact that NCR outnumbers CCG rules, though slightly, indicates NCR’s significance to Chinese CCG derivation.

NCR can be subcategorized further into four kinds (Table 3). The rule of highest frequency under each subtype is instantiated in Table 4, and their use during derivation is shown in Figures 8–11. We are particularly interested in unary ones that take 90% of the total NCR usage not only because of their prominent status in Chinese CCG derivation but also their correspondence to Zhu’s realization (though unary rules can also be found in the construction of CCGbank in connection with English and other languages (especially for topicalization), the range they cover in Chinese (4 subtypes) is broader, and the proportion higher (M. Steedman, personal communication, April 21, 2019)).

As noted in [6, 9, 27], unary rules in CCGbank transit the connotation of categories from category of pos on the left end of the operation to category of function on the right end. For example, in Figure 8, the S\NP category of the topic-drop constituent is redirected into a stand-alone sentence (S); in Figure 10, the NP category of the topicalized constituent into a sentential premodifier (S/S) (similar to in Figure 1), and in Figure 11, the S\NP category of the verb phrase into a noun modifier (NP/NP) to legitimate the larger NP to be used in another clause as an object. However, a nuanced difference present on the output side is yet to be differentiated, with some output categories representing

TABLE 1: Primitive categories in Chinese CCGbank (CASS).

Primitive categories	NP	S	M	Qp	Conj	LCP	PP
Frequency	540,614	291,867	20,572	20,798	18,466	15,403	6,555

M, QP, Conj, LCP, and PP are categories for classifier, quantification phrase, conjunct, location phrase, and propositional phrase, respectively.

TABLE 2: Frequencies and percentages of rules utilized in Chinese CCGbank (CASS).

	F	$B + T + S$	NCR
Frequency	531,326	21,568	24,774
% of total	91.9777%	3.7337%	4.2886%

TABLE 3: NCR subtypes in Chinese CCGbank (CASS).

	Argument-drop	POS-absorption ⁶	Topicalization	POS-shift
Frequency	16,680	2761	2675	2658
% of total	67.33%	11.14%	10.80%	10.73%
Unary	Yes	No	Yes	Yes

⁶ A pos-absorption rule allows one of the two categories participating in the derivation to absorb the other and result in itself. For example, $S \backslash NP \ NP \rightarrow NP$, in which the right NP absorbs the left $S \backslash NP$. This pos-absorption rule violates CCG’s projection principle because $S \backslash NP$ finds its argument NP to its right (instead of left), disregarding the functionality stipulated in the structure head. However, as we have argued in Sections 3.1 and 3.2, this rule $S \backslash NP \ NP \rightarrow NP$ indeed incorporates two subprocesses: $S \backslash NP \rightarrow NP / NP$ and $NP / NP \ NP \rightarrow NP$, in which the first step realizes an intransitive verb phrase of $S \backslash NP$ category as an attributive (Zhu’s realization), and then the second step combines the resulted attributive NP / NP with its noun head. Thus, the pos-absorption rules actually add more proof to our refined model for Chinese compositionality.

TABLE 4: NCR of highest frequency under each subtype.

		Frequency	Instances
Argument-drop	Rule 1 $S \backslash NP \rightarrow S$	13,870	(13)
Pos-absorption	Rule 2 $NP \ NP \rightarrow NP$	1995	(14)
Topicalization	Rule 3 $NP \rightarrow S / S$	1838	(15)
Pos-shift	Rule 4 $S \backslash NP \rightarrow NP / NP$	733	(16)

达到目标后, <i>after-achieving-a-goal</i> $(S \backslash NP) / (S \backslash NP)$	不断提出更高的目标 <i>put-forward-higher-aims-consistently</i> $S \backslash NP$	$\xrightarrow{\text{rule.1}}$
$\frac{S \backslash NP}{S}$		

FIGURE 8: A CCG derivation with rule 1.

中国小将 <i>young-Chinese-player</i> NP	常昊 <i>CHANG-Hao</i> NP	$\xrightarrow{\text{rule.2}}$
$\frac{NP}{NP}$		

FIGURE 9: A CCG derivation with rule 2.

syntactic functions (local dependency shown in Figure 11) and others pragmatic functions (topic-related phenomena shown in Figures 8 and 10), corresponding to realization₁ and realization₂ respectively.

3.4. Some Thought. To summarize, Chinese syntax assumes differently on both components of CCG: one, *categories in Chinese equal not to categories of pos, they could also be categories of syntactic or pragmatic functions*; two, *derivations in Chinese not only feature composition but also*

realization. The picture is further complicated by the multipath derivation of utterances that Chinese’s surface structure demonstrates because the information structure expressed by utterances is manipulated by word order, which thus plays a dual role in determining category composition as well. It should be aware that the flat surface structure we see for Chinese is one that sandwiches both derivations by composition and realization in the left box in Figure 7.

Based on the discussion in this section, we put forward two proposals for our design of Chinese CCG parser:

- (1) Give up category division based on pos and activate a simple function-based category system put up in [28, 29] in order to solve the form-function distinction distinctive in Chinese CCG derivation (Section 4.2 for detail);
- (2) Anchor the constructions that causes wrong taggings on verbs because both challenges of noun/verb ambiguity and argument-drop for Chinese parsing (as well as Chinese CCG parsing) are essentially verbal. Both challenges may resort to realization₁ that takes over from phrasal derivation, thus is construction-restricted, whereas argument-drop may also occur at realization₂ when dropping the subject of a predicate which is also the topic of the discourse. For the latter case, we include Tse [30]’s number 1 strategy for argument-drop—adding S / NP category to the lexicon, as is shown by clause (4d) in Table 5. The remaining question now is which constructions cause verb tagging ambiguities in Chinese. Hence, it is our task in Section 4 to probe into the constructions that causes verb tagging ambiguities first thing of all.

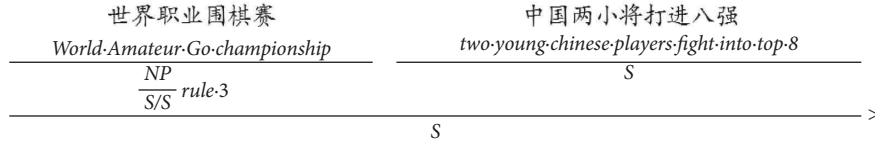


FIGURE 10: A CCG derivation with rule 3.

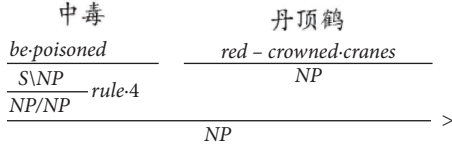


FIGURE 11: A partial CCG derivation of “又发现了三只中毒丹顶鹤” (find three more poisoned red-crowned cranes) with rule 4.

TABLE 5: NVN parsing process in detail.

Parsing process	Related rules
Step 1: tokenization and simple category assignment	—
Step 2: achieving larger SC and AC	Rule 1–2
Step 3: dealing with MP constructions	
(a) Tag-disambiguation within MP constructions	Rule 3,4,13
(b) Dependence analysis of MPs	Rule 1,2,5,6,9
Step 4: left-to-right parsing	
(a) Determine whether an adj. is the predicate	Rule 13
(b) VC or SC absorbs AC	Rule 6–8
(c) Parse object clause and NP internal	Rule 10–12 and(18)
(d) Generate S	Rule 14–16

4. Some Assumptions concerning Chinese CCG Parsing

4.1. Constructions in Association with Incorrect Verb Tagging. We choose 500 sentences randomly from PCTB 6.0 and parse them with C&C parser [6]. Taking whether all verbs in a sentence are tagged correctly as the basic standard, we eventually single out 89 sentences with 108 incorrect local parsing errors in relation to verbs. Table 6 presents the constructions related to these verb parsing errors. We instantiate a case under each construction in Figures 12–16. In each figure, the derivation tree on the left is incorrect C&C parsing structure, and the right is the parsing hypothesis.

Among the five C&C parsing trees above, three out of five (Figures 12–14) deny their final identities as sentences of category S (if we ignore the categorial differences between sentences and utterances for the moment). Figure 12 considers the whole structure as a topicalized constituent of category S/(S\NP) where the actual predicate verb “非常重视” is parsed into the head of the argument of PP, Figure 15 wrongly identifies sentence-initial noun phrase “对外开放” as a control verb, and for Figure 16, the subject clause “抢滩高科技市场” is split into two halves, in which the verbal half “抢滩” acting as the predicate verb and the nominal half “高科技市场” being the head of the postverbal argument

structure, resulting in an intransitive VP eventually. Though the other two figures (Figures 13 and 14) luckily obtain sentence category S in the end, both misidentify the predicate verbs as well.

When investigating into these parsing trees, it should be noticed that all tagging errors arise due to misrecognition over certain constructions, whose composition and realization usually accomplish prior to the decision of predicate verb. The most distinctive ones are the top three kinds (PP, *de* construction, and coordination) in Table 6 because each of them contains syntactic markers to help identify those constructions syntactically. Inspired by observation above, we propose a “maximum projection dynamic pos tagging” mechanism (MP tagger) to tag the three kinds of maximal projected constructions first before anchoring the predicate in similar spirit with the Stanojević and Steedman’s incremental parsing algorithm [31, 32]. Before moving on to the working hypothesis of MP tagger, we need to be prepared with another category system tailored to the needs of Chinese.

4.2. Simple Categories. Traditional pos distinction encounter great challenges when it is used to analyze Chinese because of the ubiquitous form-function distinctions. Thus, Chinese grammarians propose a *substantive-predicates-auxiliaries* distinction based on the functions the words can play [21, 28, 29]. *Substantives* function as subjects and objects; *predicates*, as the name suggests, the predicates of the structure; and *auxiliaries* umbrella the rest that do not make predicate or argument contributions to the structure. Following Chao and Zhu’s discussion over the three-way distinction in [21, 29], we elaborate in Table 7 the three function classes with the word classes that Chao and Zhu list under each one of them (note that the word classes here bear some differences from traditional pos because they are the commonly used ones in Chinese studies), as well as corresponding CCG categories in terms of the primitive categories in Table 1.

We put forward a simple category system for CCG concatenation in line with the three-way distinction accordingly (Table 8). Besides three basic categories SC (substantive component), VC (predicative component), and AC (auxiliary component) that corresponds to substantive, predicate and auxiliary, respectively, the system includes an additional U category for deductive purpose, standing for final utterance. VC and AC can be deductively defined by SC and U as in (17).

- (17) (a) $VC \in \{(U\backslash SC)/SC, U\backslash SC, U/SC\}$
 (b) $AC \in \{SC/SC, VC_1/VC_1, VC_2/VC_2, VC_3/VC_3, VC_1\backslash VC_1, VC_2\backslash VC_2, VC_3\backslash VC_3, (VC_1/VC_1)/SC, (VC_2/$

TABLE 6: Constructions containing verb parsing errors.

Constructions	PP	<i>de</i> construction	Coordination	NP-internal	Subject clause	Others
Number	31	29	22	18	3	5
% of total	28.7%	26.9%	20.4%	16.7%	2.8%	4.6%

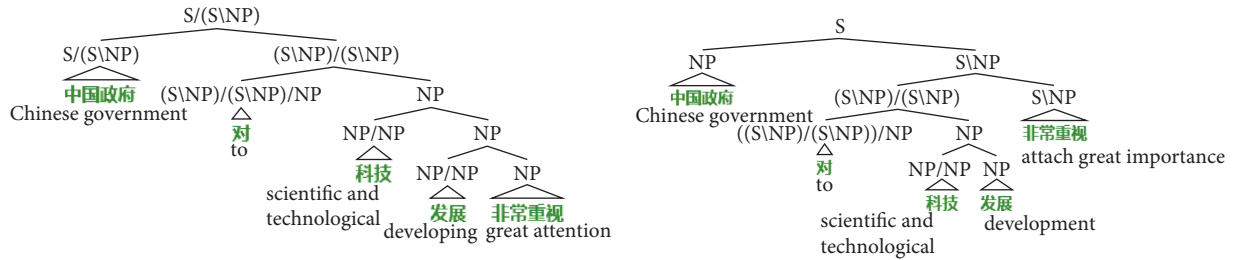


FIGURE 12: PP structure.

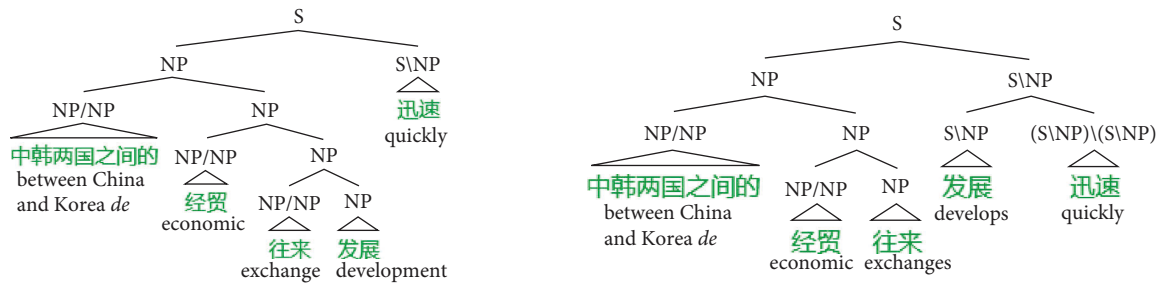


FIGURE 13: *de* construction.

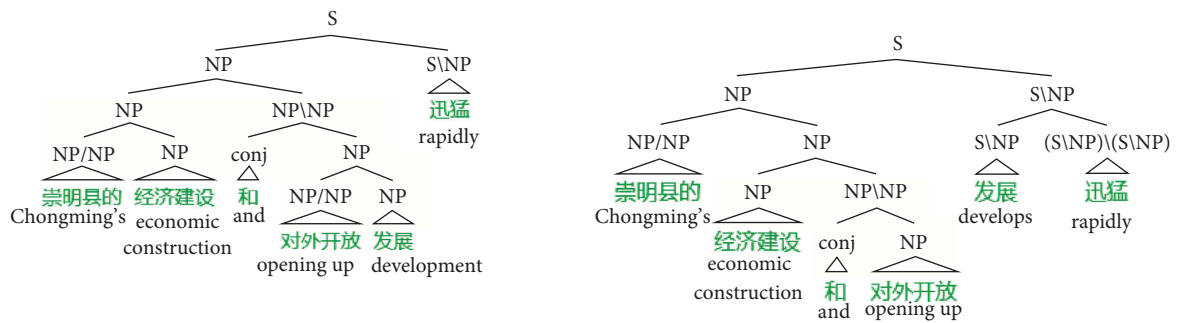


FIGURE 14: Coordination.

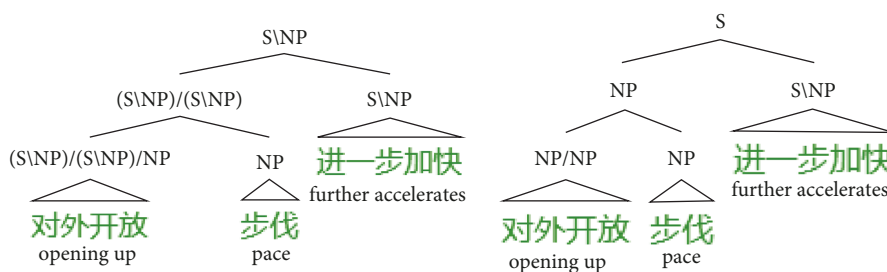


FIGURE 15: NP internal.

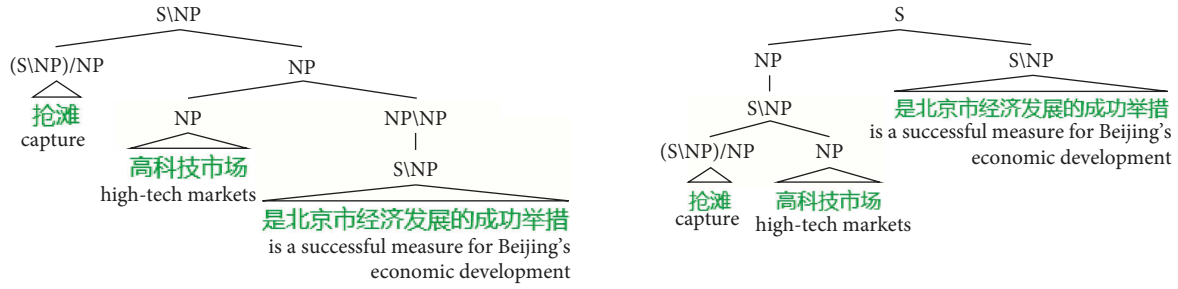


FIGURE 16: Subject clause.

TABLE 7: Three-way function class distinction.

	Function classes	Word class in [23]	Sample words	CCG category
Content words	Substantives	Nouns	水 'water', 道德 'morality'	NP
		Localizers	北京 'beijing', 图书馆 'library'	NP
		Place words	上 'top', 东边 'east'	LCP\NP
		Time words	今天 'today', 从前 'before'	NP/NP
		Numerals	一 'one', 百 'hundred'	QP/M
	Predicates	Classifiers	张 'piece', 条 'slip'	M
		Pronouns	谁 'who', 这 'this'	(NP/NP)/M
		Verbs	来 'come', 吃 'eat'	S\NP, (S\NP)/NP, etc.
		Adjectives	红 'red', 大 'big'	NP/NP
		Function words	Auxiliaries	Adverbs
Prepositions	把 'cause', 被 'passive'			(S\NP)/S, etc.
Conjunctives	和 'and', 如果 'if'			Conj
Particles	的 'of', 所 'possessive'			(NP/NP)\(S\NP), etc.
Interjunctions	啊 'ah'			S/S, S\S, etc.
Conjunctives	和 'and', 如果 'if'			Conj

TABLE 8: Simple category system.

Simple category	Word class in Jieba	Word class in [23]	Function class
SC	Common nouns; names of place/ person/ organization/ works; other proper names	Nouns	Substantives
		Localizers	
		Place words	
		Time words;	
		Numerals	
		Classifiers	
VC	Common verbs; modals; nominal verbs	Verbs	Predicate
	adj.; adverbial adj.; noun adjective	Adjectives	
AC		Adverbs	Predicate
		Prepositions	
		Conjunctives	
		Particles	Auxiliaries
U	Other auxiliaries, punctuations	Interjunctions	
U	Utterance		

$$VC_2/SC, (VC_1VC_1)/VC_1, (VC_2VC_2)/VC_2, (VC_3VC_3)/VC_3, \text{ where } VC_1 = (U/SC)/SC, VC_2 = U/SC, \text{ and } VC_3 = U/SC$$

It can be seen from Table 8 and (17) that we modify the connotation of VC slightly relative to predicates in two ways. First, adjectives are excluded from VC to be AC because adjectives are mainly nominal modifiers that makes no

contribution to PA structure according to our calculation upon the 500-sentence test set (among the 500 sentences, 207 contains adjectives, among which 189 are modifiers and 18 are predicates). The few cases where adjectives are predicates can be taken care of by rule 13 (AC converted into VC) at parsing process 4a as shown in Table 5. Second, a U/SC category (number 1 approach in [30]) is included in VC to deal with subject position topic-drop. From the

TABLE 9: Computational rules guided by simple categories (coordination rules in the form of $A \text{ Conj } A \rightarrow A$ are also used, where A can stand for SC, AC, or VC).

(1) AC AC \rightarrow AC	(5) VC AC VC \rightarrow VC	(9) AC SC \rightarrow AC	(13) AC NULL \rightarrow VC
(2) SC SC \rightarrow SC	(6) AC SC \rightarrow SC	(10) VC SC \rightarrow SC	(14) SC VC SC \rightarrow U
(3) VC NULL \rightarrow SC ¹¹	(7) AC VC \rightarrow VC	(11) SC VC \rightarrow SC	(15) SC VC \rightarrow U
(4) AC NULL \rightarrow SC	(8) VC AC \rightarrow VC	(12) SC VC SC \rightarrow SC	(16) VC SC \rightarrow U

¹¹ Null here is only a placeholder and the rules with NULL here are indeed unary rules.

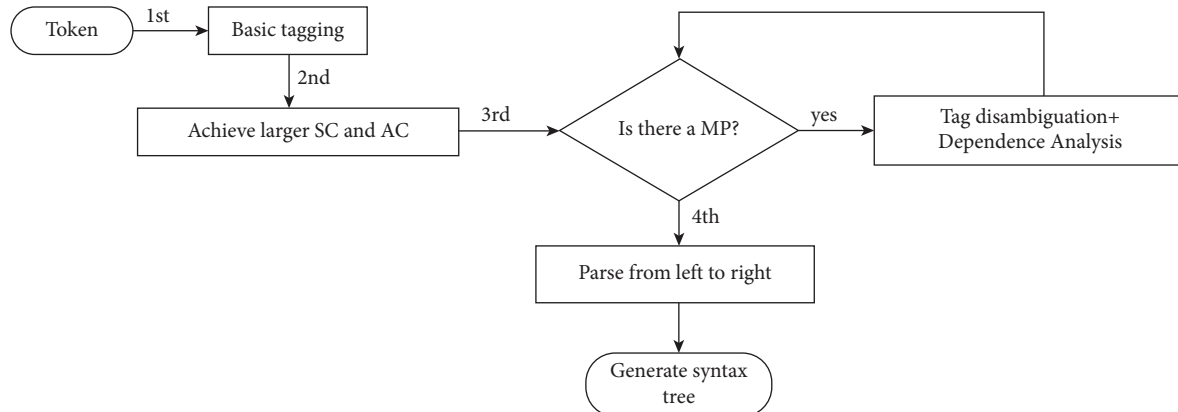


FIGURE 17: NVN parsing process in general.

perspective of parsing, we conduct sentence segmentation with Jieba segmentation tool developed specifically for Chinese by Chinese Academy of Sciences (Available at <https://github.com/fxsjy/jieba>). The 24 word classes adopted by Jieba are listed in the fourth column of Table 8, matching approximately to those in the second column. An MP Tagger then labels the tokens with one of the simple categories each of their Jieba word class belongs to and thus is capable of parsing Chinese sentences without deep learning. It is our originality to consider function and structural requirements when parsing.

4.3. *MP Tagger*. MP tagger tags the verbs within an utterance by the procedures as follows:

(18)

- (a) Mark all verbs as VC in the first round of pos tagging
- (b) Recognize the syntactic markers of MP constructions (PP, *de* construction, coordination, NP internal, and subject clause as in Table 6) and convert VCs within MP structures into NC or AC
- (c) Determine the unique VC
- (d) Finalize the main predicate according to the Predicate Rule (19) when the parser cannot recognize the MP structure

(19) *The Predicate Rule: the predicate of a sentence is the leftmost verb in the sentence.*

In MP structures with syntactic markers (PP, *de* construction, and coordination), the pos tagging of their internal components in both PP and *de* construction can be

determined instantly after locating the markers. It is either the case that the category of the construction is fixed (*de* construction is an SC), or the category of its internal constituent is fixed (the argument in PP is always an SC). Since coordination can coordinate SCs, VCs, ACs, and even Us, we also need to rely on the pos tagging distribution around coordination to determine the pos of the coordination.

As for NP internal and subject clause that possess no fixed syntactic markers, it is not easy to identify the structure with nonstatistical methods. We propose the “leftmost verb” strategy (19) to assist the processing over NP internal based on our investigation over the top 100 sentences in PCTB 6.0 where we find that the main verbs in 73% of the sentences are the leftmost ones. For now, the subject clauses remain unparseable under our mechanism.

Synthesizing our reflections over MP constructions, simple categories, and MP tagger, we then design a NVN parser tailored for Chinese CCG parsing without deep learning model.

5. NVN Parser Based on MP Tagging

5.1. *NVN Parser*. NVN parser offers a rule-based parsing model over Chinese. The core parsing ideas are based on our MP tagger and further materialized by simple categories and 16 phrase-structural kind of computational rules (Table 9). Compared with earlier CCG parsers, mainly C&C in [6], NVN Parser can resolve more reasonably the parsing errors of noun/verb ambiguity and argument-drop. The parsing procedure consists of 4 steps in general (Figure 17):

Step 1: segment sentences into tokens with Jieba segmentation tool and then label the tokens with simple

categories SC, VC, or AC according to Jieba-simple category correspondence in Table 8

Step 2: achieve larger SCs and ACs

Step 3: deal with the possible MP constructions with syntactic markers

Step 4: parse the structures without syntactic markers from left to right until only one VC is left

5.2. *NVN Parsing Processes.* The parsing details are shown in Table 5 together with the rules (Table 9) used in each step or substep.

Step 1: tokenization and simple category assignment. We start with segmenting sentences into tokens with Jieba segmentation tool and then label the tokens with simple categories SC, VC, or AC according to the pos they get from Jieba according to Table 8.

Step 2: achieving larger SC and AC. Our parser absorbs neighboring ACs and neighboring SCs according to rules 1 and 2. AC absorption rule 1 in Table 5 allows AC components not in three typical MP constructions to absorb its adjacent AC, forming a larger AC component. For example, token “very” absorbs token “good” when the two cooccur to form a larger AC token “very good”. Likewise, SC absorption rule 2 can achieve larger noun phrases, some of which are long proper nouns not recognizable by word segmentation.

Step 3: dealing with MP constructions. There are three MP constructions to be treated during this step—*de* construction, coordination, and PP, whose syntactic markers are “的”(‘*de*’), “和”(‘*and*’), “或”(‘*or*’), “在”(‘*at*’), “比”(‘*compare with*’), “宁可”(‘*would rather*’) etc. The parser first detects abovementioned syntactic markers. If no syntactic markers are detected, the parser will skip Step 3 directly.

If a syntactic marker is detected, the parsing should determine the scope of MP constructions and parse according to Table 10.

We apply 6 bit array “choose” to record the pos distribution near the coordination, use 3 bit array “left,” “right” for *de* construction, and 3 bit array “right” to record the right pos distribution of a PP. In these arrays, we mark null element with -1, AC with 0, SC with 1, VC with 2, *de* with 3, syntactic marker of coordination with 4. EOVS is short for “existence of other verbs outside the array.” It will be adopted when disambiguation listed above fails:

Step 4: left-to-right parsing. At this step, the parser processes all structures without syntactic markers: decide whether an adjective is the predicate as in 4a; if the adjective is the predicate, parse NP internal that contains VC as in 4b and absorb other ACs as in 4c and generate U by rule 14–16; otherwise, parse NP-internals and object clause, absorb all ACs, and generate U.

5.3. *Evaluation.* We adopt a ready testing model for CCG parsers—dependency tuple—proposed by Clark et al. [33].

TABLE 10: Partial projection requirements for MP constructions.

MP constructions	
(1) <i>de</i> construction	
VC	$\rightarrow_{\text{Right}[2,-1,-1]}$ SC
AC VC	$\rightarrow_{\text{Right}[1,2,-1]_{\text{EOV}}}$ AC SC
VC	$\rightarrow_{\text{Left}[2,-1,-1]}$ SC
VC SC	$\rightarrow_{\text{Left}[2,0,-1]_{\text{EOV}}}$ SC SC
(2) Coordination	
VC SC 和 VC SC VC	$\rightarrow_{\text{Choose}[-1,2,0,4,2,0,2]}$ SC SC 和 SC SC VC
AC VC 和 VC SC	$\rightarrow_{\text{Choose}[-1,1,2,4,2,0,-1]}$ AC VC SC
VC AC VC 和 AC VC	$\rightarrow_{\text{Choose}[0,1,2,4,2,2,2]}$ VC AC SC 和 AC SC
AC VC SC 和 SC AC AC	$\rightarrow_{\text{Choose}[1,2,0,0,1,1,3]_{\text{EOV}}}$ AC SC SC 和 SC AC AC
(3) PP	
对 VC AC VC	$\rightarrow_{\text{Right}[2,1,2]}$ 对 SC AC VC
对 AC SC AC	$\rightarrow_{\text{Right}[1,0,1]}$ 对 AC SC VC
对 AC VC VC	$\rightarrow_{\text{Right}[1,2,2]}$ 对 AC SC VC

TABLE 11: Evaluation for NVN parser and C&C.

	NVN parser	C&C [6]
(A) The general indicators		
Coverage	99.40	95.01
UF	89.20	79.27
LF	85.93	74.65
(B) Challenges of Chinese parsing in [6]		
Prodrop (LF)	86.2	76.8
Noun/verb ambiguity (recall)	88.4	79.3
(C) The LF of MP constructions in parsing		
Subject clause	53.1	81.1
NP internal	85.6	69.2
Coordination	83.2	73.0
<i>de</i> construction	88.4	82.4
PP structure	74.1	79.6

Later standard in [4] is similar. We test existing C&C parser and our NVN parser on 500 sentences randomly chosen from PCTB 6.0, calculating the two parser’s F-score over unlabeled dependencies (UF), F-score over labelled dependencies (LF), coverage and the LF of five MP constructions by manually comparing the candidate results with gold standard, as shown in Table 11.

Our evaluation deviates from [33] in that it takes chunks, instead of words, as the minimum units because the two parsers cannot be compared by words for two reasons. One, NVN parser bears no dependency relation in the sense of traditional CCG parsing due to the adoption of simple categories. Two, NVN parser with a smaller tagging set achieves correct tagging more easily than C&C parser if both are under correct dependency. These problems can be avoided by calculating UF\LF by chunks because it only cares about the labeling of the chunk as a whole without peeking into its internal structure. We choose the predicate as the first lexical chunk, the subject component on its left as

the second, the object component on its right the third, other ingredients being accessories attaching to them. UF, LF, and other indicators are calculated by the dependence tuple in relation to the three chunks. It can be seen that NVN parser surmounts at both indicators A and B, proving “maximum projection dynamic pos tagging” to some extent. Yet, indicators in C find the deficiency of NVN parser over subject clause, which only appears in 15 sentences. It was serious but did not cause a devastating blow to the overall LF.

6. Conclusion and Limitation

Chinese parsing has always been perplexing because of its flexible pos and lack of strict inflections. Analyzing how Chinese syntax assumes differently on categories and combinatory rules from that by CCG, a clearer picture is unfolded in front of us to help explain why and how pos ambiguities challenge Chinese parsing with CCG and also Chinese parsing in general. We propose a simple category system that is based on syntactic functions proposed earlier by Chao, Zhu, and Lv and design an NVN parser with simple categories and an MP tagger. Admittedly, despite of its high LF and UF score compared with C&C parser, NVN parser still has some shortcomings to overcome: first, simple category label might be oversimplified to dwell in the clear pos orientation of traditional CCG category; second, reasonable non-deep learning mechanisms are needed for parsing prepositional MP, NP internal, subject clauses, as well as asymmetric coordination; third, large-scale data set measurement is not yet carried out. The present work is an initial attempt to ponder over the Chinese parsing from theoretical perspective. It is our hope that it could shed light on more works on Chinese and CCG parsing.

Data Availability

The data used to support the analysis of the study are available from the first author and the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was funded by the Major Program of National Social Science Foundation of China (grant no. 17ZDA027) and Fundamental Research Funds for the Central Universities, USTB (grant no. FRF-BR-20-13BA).

References

- [1] M. Steedman, *Surface Structure and Interpretation*, MIT Press, Cambridge, CA, USA, 1996.
- [2] M. Steedman, *The Syntactic Process*, MIT Press, Cambridge, CA, USA, 2000.
- [3] S. Clark and J. Hockenmaier, “Evaluating a wide-coverage CCG parser,” in *Proceedings of the LREC 2002 beyond Parseval Workshop*, pp. 60–66, Las Palmas, Spain, 2002.
- [4] S. Curran and J. R. Curran, “Wide-coverage efficient statistical parsing with CCG and log-linear models,” *Computational Linguistics*, vol. 33, no. 4, pp. 493–552, 2007.
- [5] S. Clark and J. R. Curran, “Formalism-independent parser evaluation with CCG and DepBank,” in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pp. 248–255, Prague, Czech Republic, June 2007.
- [6] D. Tse and J. R. Curran, “The challenges of parsing Chinese with combinatory categorial grammar,” in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 295–304, Montreal, Canada, 2012.
- [7] R. Levy and C. Manning, “Is it harder to parse Chinese, or the Chinese Treebank?” in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pp. 439–446, Sapporo, Japan, July 2003.
- [8] J. Kummerfeld, D. Tse, J. R. Curran, and D. Klein, “An empirical examination of challenges in Chinese parsing,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 98–103, Sofia, Bulgaria, August 2013.
- [9] J. Hockenmaier, *Data and Models for Statistical Parsing with Combinatory Categorial Grammar*, University of Edinburgh, PhD. Dissertation, Edinburgh, Scotland, 2003.
- [10] H.-X. Man and C.-L. Zou, “On the hierarchical pyramid structure of performance grammar,” *Journal of Hubei University (Philosophy and Social Science)*, vol. 46, no. 5, pp. 77–84, 2019, in Chinese.
- [11] D.-X. Zhu, *Questions and Answers on Grammar*, The Commercial Press, Beijing, China, 1985, in Chinese.
- [12] K. Ajdukiewicz, “Die syntaktische Konnexität,” *Studia Philosophica*, vol. 1, pp. 1–27, 1935.
- [13] Y. Bar-Hillel, “A quasi-arithmetical notation for syntactic description,” *Language*, vol. 29, no. 1, pp. 47–58, 1953.
- [14] R. Çakırcı, “Automatic induction of a CCG grammar for Turkish,” in *Proceedings of the ACL Student Research Workshop*, pp. 73–78, Ann Arbor, Michigan, June 2005.
- [15] R. Montague, “The proper treatment of quantification in ordinary English,” in *Formal Philosophy: Selected Papers of Richard Montague*, R. H. Thomason, Ed., pp. 247–270, Yale University Press, London, UK, 1974.
- [16] M. Schönfinkel, “On the building blocks of mathematical logic,” in *From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931*, van Heijenoort, Ed., Harvard University Press, Cambridge, CA, USA, 1967.
- [17] H. Curry, R. Feys, and W. Cragge, *Combinatory Logic*, Vol. 1, North-Holland Publishing company, Amsterdam, Holland, 1958.
- [18] C. Bozşahin, *Combinatory Linguistics*, De Gruyter Mouton, Berlin, Germany, 2012.
- [19] M.-Y. Hu, “On word classes in modern Chinese,” *Studies of Chinese Language*, vol. 44, no. 5, pp. 381–389, 1995, in Chinese.
- [20] J.-X. Shen, “My view of word classes in Chinese,” *Linguistic Sciences*, vol. 8, no. 1, pp. 1–12, 2017, in Chinese.
- [21] Y.-R. Chao, *A Grammar of Spoken Chinese*, University of California Press, Berkeley and Los Angeles, USA, 1968.

- [22] C. Li and S. Thompson, “Subject and topic: a new typology of language,” in *Subject and Topic*, C. Li, Ed., pp. 457–590, Academic Press, New York, NY, USA, 1976.
- [23] R. J. LaPolla, “Pragmatic relations and word order in Chinese,” in *Word Order in Discourse*, P. Downing and M. Noonan, Eds., John Benjamins Publishing Company, Amsterdam, Holland, pp. 297–329, 1995.
- [24] R. J. LaPolla, “Chinese as a topic-comment (not topic-prominent and not SVO) language,” in *Studies of Chinese Linguistics: Functional Approaches*, J. Xing, Ed., pp. 9–22, Hong Kong University Press, Hong Kong, China, 2009.
- [25] J.-X. Shen, “Does Chinese have the subject-predicate structure?” *Modern Foreign Languages*, vol. 40, no. 1, pp. 1–13, 2017, in Chinese.
- [26] J. Hockenmaier and M. Steedman, *CCGbank: Users’ Manual, Technical Report MS-CIS-05-09*, University of Pennsylvania, Pennsylvania, 2005.
- [27] M. Honnibal, *Hat Categories: Representating Form and Function Simultaneously in Combinatory Categorical Grammar from the Penn Treebank*, University of Sydney, PhD. Dissertation, Sydney, 2010.
- [28] S.-X. Lv, *Issues of Modern Chinese Grammar Analysis*, The Commercial Press, Beijing, China, 1979, in Chinese.
- [29] D.-X. Zhu, *A Course in Chinese Grammar*, The Commercial Press, Beijing, China, 1982, in Chinese.
- [30] D. G.-S. Tse, *Chinese CCGbank: Deep Derivations and Dependencies for Chinese CCG Parsing*, PhD. Dissertation, The University of Sydney, Sydney, 2013.
- [31] M. Stanojević and M. Steedman, “CCG parsing algorithm with incremental tree rotation,” in *Proceedings of the NAACL-HLT2019*, pp. 228–239, Minneapolis, Minnesota, 2019.
- [32] M. Stanojević, J. Hale, and M. Steedman, “Predictive processing of coordination in CCG,” in *Proceedings of the 33rd Annual CUNY Conference on Human Sentence Processing*, Amherst, MA, USA, March 2020, <https://osf.io/4qh28/>.
- [33] S. Clark, J. Hockenmaier, and M. Steedman, “Building deep dependency structures with a wide-coverage CCG parser,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 327–334, Philadelphia, July 2002.