Chinese CCGbank: extracting CCG derivations from the Penn Chinese Treebank

Daniel Tse and **James R. Curran**
School of Information Technologies
University of Sydney
{dtse6695,james}@it.usyd.edu.au

Abstract

Automated conversion has allowed the development of wide-coverage corpora for a variety of grammar formalisms without the expense of manual annotation. Analysing new languages also tests formalisms, exposing their strengths and weaknesses.

We present Chinese CCGbank, a 760,000 word corpus annotated with Combinatory Categorial Grammar (CCG) derivations, induced automatically from the Penn Chinese Treebank (PCTB). We design parsimonious CCG analyses for a range of Chinese syntactic constructions, and transform the PCTB trees to produce them. Our process yields a corpus of 27,759 derivations, covering 98.1% of the PCTB.

1 Introduction

An annotated corpus is typically used to develop statistical parsers for a given formalism and language. An alternative to the enormous cost of hand-annotating a corpus for a specific formalism is to convert from an existing corpus.

The Penn Treebank (PTB; Marcus et al., 1994) has been converted to HPSG (Miyao et al., 2004), LFG (Cahill et al., 2002), LTAG (Xia, 1999), and CCG (Hockenmaier, 2003). Dependency corpora, e.g. the German Tiger corpus, have also been converted (Hockenmaier, 2006). The Penn Chinese Treebank (PCTB; Xue et al., 2005) provides analyses for 770,000 words of Chinese. Existing PCTB conversions have targeted TAG (Chen et al., 2005) and LFG (Burke and Lam, 2004; Guo et al., 2007).

We present Chinese CCGbank, a Chinese corpus of CCG derivations automatically induced from the PCTB. Combinatory Categorial Grammar (CCG; Steedman, 2000) is a lexicalised grammar formalism offering a unified account of local and non-local dependencies. We harness the facilities of CCG to provide analyses of Chinese syntax including topicalisation, pro-drop, zero copula, extraction, and the 把 *ba-* and 被 *bei*-constructions.

Pushing the boundaries of formalisms by subjecting them to unfamiliar syntax also tests their universality claims. The freer word order of Turkish (Hoffman, 1996) and the complex morphology of Korean (Cha et al., 2002) led to the development of extensions to the CCG formalism.

We present our analysis of Chinese syntax under CCG, and provide an algorithm, modelled after Hockenmaier and Steedman (2007), to incrementally transform PCTB trees into CCG derivations. The algorithm assigns CCG categories which directly encode head and subcategorisation information. Instances of Chinese syntax demanding special analysis, such as extraction, pro-drop or topicalisation, are pin-pointed and given elegant analyses which exploit the expressivity of CCG.

Our conversion yields CCG analyses for 27,759 PCTB trees (98.1%). Coverage on lexical items, evaluated by 10-fold cross-validation, is 94.46% (by token) and 73.38% (by type).

We present the first CCG analysis of Chinese syntax and obtain a wide-coverage CCG corpus of Chinese. Highly efficient statistical parsing using a CCGbank has recently been demonstrated for English (Clark and Curran, 2007). Our Chinese CCGbank will enable the development of similarly efficient wide-coverage CCG parsers for Chinese.

2 Combinatory Categorial Grammar

CCG (Steedman, 2000) is a lexicalised grammar formalism, with a transparent syntax-semantics interface, a flexible view of constituency enabling concise accounts of various phenomena, and a consistent account of local/non-local dependencies.

It consists of *categories*, which encode the type and number of arguments taken by lexical items, and *combinators*, which govern the possible interactions between categories.

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1083–1091,
Beijing, August 2010

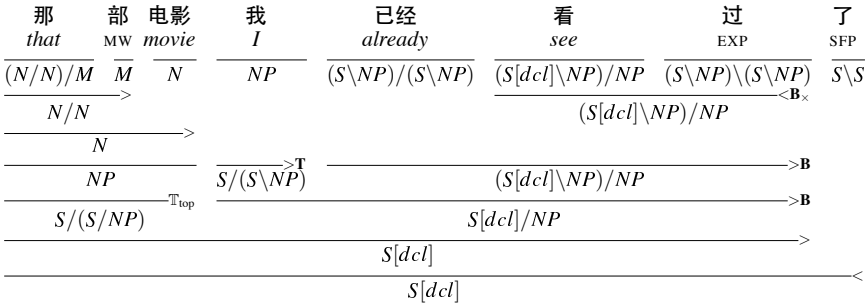


Figure 1: Chinese CCG derivation: "That movie, I've already seen."

A CCG grammar defines *atomic categories*, e.g. *NP* and *S*, which may be recursively constructed into *complex categories*, e.g. N/N and $S\backslash NP$.[1] Figure 1 shows how combinators govern the interaction of categories for lexical items, while slashes specify argument directionality.

The combinators allow us to reduce lexical ambiguity, by preserving a word's canonical category even when displaced from its canonical position. This facility is a strength of CCG, but elevates its generative power to mild context-sensitivity.

Some combinators may be disabled in a given language – the *multi-modal* CCG (Baldridge, 2002) allows these distinctions to be lexically specified.

Introducing non-CCG rules decrease categorial ambiguity at the expense of deviating from the formalism. Hockenmaier and Steedman (2002) show that these greatly improve lexical coverage. Their analysis of English employs non-CCG rules to coerce a verb phrase headed by a participle (category $S[ng]\backslash NP$) to a post-nominal modifier:

$$S[ng]\backslash NP \longrightarrow NP\backslash NP \quad (1)$$

This frees verbs from having to possess a distinct category in each position, thus trading off lexical ambiguity for derivational ambiguity. Honnibal and Curran (2009) extended CCG with *hat categories*, enabling the lexical specification of these unary type-change rules.

Hockenmaier and Steedman (2002, 2007) developed CCGbank, the first wide-coverage English CCG corpus, by converting 1.2 million words from the Wall Street Journal section of the PTB. CCGbank has made possible the development of wide-coverage statistical parsers for CCG in English, notably C&C (Clark and Curran, 2007).

[1] *Abbreviations in this paper:* The directionless slash | stands for one of $\{/, \backslash\}$. We also use the verbal category abbreviations $VP \equiv S\backslash NP$ and $TV \equiv (S\backslash NP)/NP$.

3 Penn Chinese Treebank

Xue et al. (2005) developed the Penn Chinese Treebank (PCTB), the first syntactically annotated corpus for Chinese. The corpus includes newswire text, magazine articles, and transcribed speech.[2]

Xue et al. establishes several principles for a more disciplined and consistent style of annotation compared to the original PTB. These principles include *complement/adjunct marking:* allowing the recovery of predicate-argument structure; *limited semantic role marking:* the annotation of modifier phrases with semantic roles; *covert argument marking:* the retention of traces of arguments deleted through pro-drop; and *NP internal structure:* bracketing of NP structure where the intended interpretation is clear.

The *one relation per bracketing* principle unambiguously encodes a grammatical relation (chiefly, predication, adjunction, or complementation) through the configuration of a node and its children. Xue et al. developed this principle to assist conversions from the PTB, e.g. Hockenmaier (2003), in resolving argument/adjunct distinctions.

PCTB derivations are pre-segmented, pre-tokenised, and POS tagged. Owing to the dearth of morphology in Chinese, the concept of *part of speech* is more fluid than that of English – the word 比较 *bijiao* 'compare' might be glossed as a verb, adjective, adverb, or noun depending on its context. Noun/verb mis-taggings are a frequent error case for PCFG parsing on PCTB data, compounded in Chinese by the lack of function words and morphology (Levy and Manning, 2003). This ambiguity is better handled by the adaptive multitagging approach used by Clark and Curran (2007) for CCG supertagging, in which each lexical item is tagged with a set of CCG categories.

We present our CCG analysis of Chinese syntax below, followed by our conversion algorithm.

[2] We use the Penn Chinese Treebank 6.0 (LDC2007T36).

4 The syntax of Chinese

4.1 Basic clause structure

Chinese is typologically SVO, with some OV elements (relative clauses, adjunct PPs and noun modifiers precede their heads). Numbers and determiners may not modify nouns directly; a *measure word* must intervene.

The category structure of the grammar may be inferred directly from headedness information. Heads subcategorise for the type, number and directionality of their arguments, while adjuncts receive modifier categories of the form $X \mid X$.

(2)

我	在		超市		买
I	at		supermarket		buy
NP	$(VP/VP)/NP$		*NP*		VP/NP
了	一		盒	鸡蛋	
PERF	one		box:MW	eggs	
$VP\backslash VP$	$(N/N)/M$		*M*	*N*	

I bought a box of eggs at the supermarket.

4.2 Topicalisation

In topic-prominent languages, the *topic* refers to information which the speaker assumes is known by the listener. In Mandarin, topicalisation manifests as left-dislocation of the topic phrase (Li and Thompson, 1989). We distinguish *gap* and *non-gap* topicalisation depending on whether the topic is co-referent with a gap in the sentence.[3]

For gapped topicalisation (cf. Figure 1), we adopt the Steedman (1987) topicalisation analysis:

$$T \rightarrow S/(S/T) \text{ for parametrically licensed } T \quad (3)$$

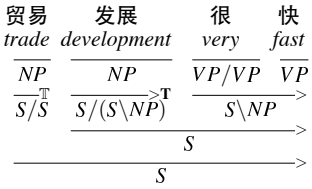
For non-gap topicalisation (Example 5), we use a variation of the analysis described in Hockenmaier and Steedman (2005), which treats the topicalised constituent as a sentential modifier. Under this analysis, the determiner in a topicalised *NP* receives $(S/S)/N$ instead of its canonical category NP/N. Instead, we propose a unary rule:

$$T \rightarrow S/S \text{ for topicalisation candidate } T \quad (4)$$

This delays the coercion to sentential modifier type (i.e. $NP \rightarrow S/S$) until after the NP has been consolidated, allowing the words under the topicalised NP to preserve their canonical categories.

[3]Non-gap topicalisation is also known as the *double subject construction* (Li and Thompson, 1989).

(5) (As for) trade, it has developed rapidly.



Topicalisation is far less marked in Chinese than in English, and the structure of topicalised constituents is potentially quite complex. The additional categorial ambiguity in Hockenmaier and Steedman (2005) compounds the data sparsity problem, leading us to prefer the unary rule.

4.3 Pro-drop

Since Chinese exhibits *radical pro-drop* (Neeleman and Szendrői, 2007), in which the viability of the pro-drop is not conditioned on the verb, the categorial ambiguity resulting from providing an additional argument-dropped category for every verb is prohibitive.

Rather than engendering sparsity on verbal categories, we prefer derivational ambiguity by choosing the unary rule analysis $S[dcl] \mid NP \rightarrow S[dcl]$ to capture Chinese pro-drop.

4.4 Zero copula

Although the Chinese copula 是 *shi* is obligatory when equating NPs, it may be omitted when equating an NP and a QP or PP (Tiee and Lance, 1986).[4]

(6)

她	今年	十八	岁
NP	*VP*/*VP*	$(S\backslash NP)/M$	*M*
3SG	this-year	18	years-old

She is 18 this year.

A solution involving a binary rule $NP\ QP \rightarrow S[dcl]$ is not properly headed, and thus violates the Principle of Lexical Head Government (Steedman, 2000). Conversely, a solution where, for example, 十八 '18' would have to receive the category $(S[dcl]\backslash NP)/M$ instead of its canonical category QP/M would lead to both data sparsity and over-generation, with VP modifiers becoming able to modify the QP directly. Tentatively, we ignore the data sparsity consequences, and have 十八 '18' receive the category $(S[dcl]\backslash NP)/M$ in this context.

[4]The copula is ungrammatical in predication on an adjectival verb, such as 高兴 'happy'. However, we analyse such words as verbs proper, with category $S[dcl]\backslash NP$.

4.5 把 *ba*- and 被 *bei*-constructions

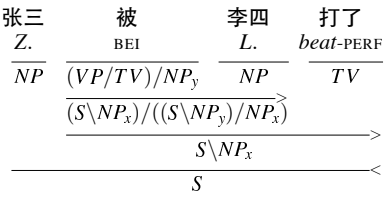
被 *bei* and 把 *ba* introduce a family of passive-like constructions in Chinese. Although superficially similar, the resulting constructions exhibit distinct syntax, as our CCG analysis reflects and clarifies.

In the 被 *bei*-construction, the patient argument of a verb moves to subject position, while the agent either becomes the complement of a particle 被 *bei* (the *long passive*), or disappears (the *short passive*; Yip and Rimmington, 1997). Although the two constructions are superficially similar (apparently differing only by the deletion of the agent NP), they behave differently in more complex contexts (Huang et al., 2008).

The long passive occurs with or without an object gap (deleted by identity with the subject of the matrix verb). We analyse this construction by assigning 被 *bei* a category which permutes the surface positions of the agent and patient. Coindexation of heads allows us to express long-distance dependencies.

Bei receives $((S\backslash NP_y)/((S\backslash NP_x)/NP_y))/NP_x$ in the gapped case (cf. Example 7) and $((S\backslash NP)/(S\backslash NP_x))/NP_x$ in the non-gapped case.

(7) Zhangsan was beaten by Lisi.



Short passives also occur with or without an object gap, receiving $(S\backslash NP_x)/((S\backslash NP)/NP_x)$ in the gapped case and $(S\backslash NP)\backslash(S\backslash NP)$ in the non-gapped case. Our analysis agrees with Huang et al. (2008)'s observation that short-*bei* is isomorphic to English *tough*-movement: our short-*bei* category is the same as Hockenmaier and Steedman (2005)'s category for English *tough*-adjectives.

In the 把 *ba* construction, a direct object becomes the complement of the morpheme 把 *ba*, and gains semantics related to "being affected, dealt with, or disposed of" (Huang et al., 2008). As for 被 *bei*, we distinguish two variants depending on whether the object is deleted under coreference with the complement of 把 *ba*.

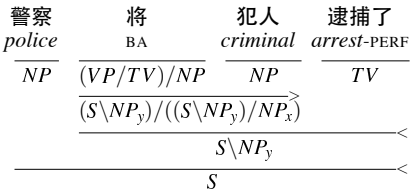
Ba receives $((S\backslash NP_y)/((S\backslash NP_y)/NP_x))/NP_x$ in the gapped case (cf. Example 8), and $((S\backslash NP_y)/(S\backslash NP_y))/NP$ in the non-gapped case.

As Levy and Manning (2003) suggest, we reshape the PCTB analysis of the *ba*-construction so that *ba* subcategorises for its NP and VP, rather than subcategorising for an IP sibling, which allows the NP to undergo extraction.

Tag	*Headedness*	*Example*
VSB	head-final	规划 建设 'plan [then] build'
VRD	right-adjunction	煮 熟 'cook done'
VCP	head-initial	确认 为 'confirm as'
VCD	appositive	投资 设厂 'invest [&] build-factory'
VNV	special	去 不 去 'go [or] not go'
VPT	special	离 得 开 'leave able away'

Table 1: Verb compounds in PCTB

(8) The criminals were arrested by the police.



4.6 Verbal compounding

Verbs resulting from compounding strategies are tagged and internally bracketed. Table 1 lists the types distinguished by the PCTB, and the headedness we assign to compounds of each type.

Modifier-head compounds (PCTB tag VSB) exhibit clear head-final semantics, with the first verb V_1 causally or temporally preceding V_2. Verb coordination compounds (VCD) project multiple heads, like ordinary lexical coordination.

In a resultative compound (VRD), the result or direction of V_1 is indicated by V_2, which we treat as a post-verbal modifier. The *V-not-V* construction (VNV) forms a yes/no question where $V_1 = V_2$. In the *V-bu/de-V* or potential verb construction (VPT), a disyllabic verb $V = V_1V_2$ receives the infix 得 *de* or 不 *bu* with the meaning *can/cannot* V. In both these cases, it is the infixed particle 得 *de* or 不 *bu* which collects its arguments on either side.

4.7 Extraction

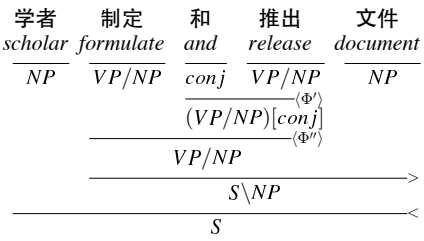
In the Chinese relative clause construction, the particle 的 *de* links a sentence with a subject or object gap with a NP to which that gap co-refers, in an analysis similar to the English construction described by Hockenmaier and Steedman (2005), mediated by the relative pronoun *that*.

As in the English object extraction case, forward type-raising on the subject argument, and forward composition into the verbal category allows us to obtain the correct object gap category S/NP.

4.8 Right node raising

Two coordinated verbs may share one or more contiguous arguments under right node raising. This analysis follows directly from the CCG definition of coordination, requiring no new lexical categories.

(9) Scholars have formulated and are releasing the documents.



4.9 Apposition

Apposition is the juxtaposition of two phrases referring to the same entity. Unlike noun modification, no clear modification relationship holds between the two phrases. The direct juxtaposition rules out Hockenmaier's (2003) analysis where a delimiting comma mediates the apposition. Chinese also allows full sentence/NP apposition:

(10) (用户 浪费 水)$_S$ 事件$_{NP}$
(users waste water)$_S$ incident$_{NP}$
incidents of users wasting water

This gives rise to the Chinese apposition binary rules $NP\ NP \rightarrow NP$ and $S[dcl]\ NP \rightarrow NP$.

5 The translation pipeline

5.1 Tagging

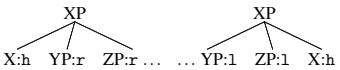
Each PCTB internal node structurally encodes a *configuration*, which lets us distinguish head-initial and head-final complementation from adjunction and predication (Xue et al., 2000).

The tagging mechanism annotates the PCTB tag of each internal node with a *marker*, which preserves this headedness information, even after the nodes are re-structured in the binarisation phase.

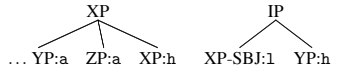
Hockenmaier's (2003) conversion algorithm uses the Magerman (1994) head-finding heuristics, a potential source of noise. Fortunately, the PCTB encodes gold standard headedness data.

The tagging algorithm is straightforward: if a node and its children unify with one of the schemata below, then the markers (e.g. `:l` or `:n`) are attached to its children. The markers `l` and `r` indicate complements *left*, or *right* of the *head* `h`; adjuncts are marked with `a`.

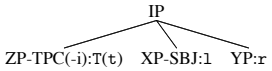
Head-initial, -final complementation



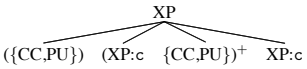
Adjunction, predication



Topicalisation (gap and non-gap)



Coordination



Others identify nodes with special syntax, such as topicalisation (`t`/`T`), apposition (`A`) or coordination (`c`), for special treatment in following phases.

NP internal structure

To speed annotation, NP internal structure is often left underspecified in PCTB (Xue et al., 2005), as in the Penn Treebank. As a result, 68% of non-trace `NP`s in PCTB have only a flat bracketing.

We assume that the internal structure of flat NPs is right-branching and head-final (Li and Thompson, 1989), following Hockenmaier and Steedman (2005), who assume this structure for English. A re-analysis of PCTB, like Vadas and Curran (2007) for the PTB, could restore this structure, and allow our conversion algorithm to yield the correct CCG analysis with no further modifications.

To obtain this default analysis, each node under NP internal structure receives the marker `n`, except the the final node, the head, which receives `N`.

5.2 Binarisation

CCG combinators take at most two categories, inducing binary derivation trees. As such, PCTB trees must be re-shaped to accommodate a CCG analysis.

Our markers control the shape of the binarised structure: head-initial complementation yields a left-branching tree, while head-final complementation, adjunction, predication, coordination, and NP internal structure all yield right-branching trees. Following Hockenmaier (2003), sentence-final punctuation is attached high.

Although the distinction between word-level tags (such as `NN`, `VA`) and phrasal tags (such as `NP`, `VP`, `LCP`) enables the configurational encoding of grammatical relations, it leaves a large number of

VP	←	VV,VE,VA,VRD	ADJP	←	JJ
ADVP	←	AD, CS	CLP	←	M
LCP	←	LC	DP	←	DT, OD
LST	←	OD	INTJ	←	IJ
FLR	←	*any node*	PP	←	P

Figure 2: Pruned unary projections

LCP	Localiser phrase	PP	Prepositional phrase
M	Measure word	QP	Quantifier phrase
N	Bare noun	S	Sentence
NP	Noun phrase	conj	Conjunction word

Table 2: Chinese CCGbank atomic category set

unary projections. While an intransitive verb (e.g. 睡觉 'sleep') would carry the verbal PCTB tag VV, and a transitive verb combined with its object (e.g. 吃了晚饭 'ate dinner') is annotated as VP, under CCG's freer concept of constituency, both receive the category $S\backslash NP$.

Pruning the unary projections in Fig. 2 prevents spurious category labellings in the next phase.

5.3 Labelling

We label each node of the binarised tree with CCG categories, respecting the headedness information encoded in the markers.

Atomic categories

The chosen mapping from PCTB tags to categories defines the *atomic category set* for the grammar. The richer representation in CCG categories permits some constituents to be expressed using a smaller set of atoms (e.g. an adjective is simply a noun modifier – N/N). Despite their critical importance in controlling the degree of under-/over-generation in the corpus, little guidance exists as to the selection of atomic categories in a CCG grammar. We observed the following principles:

Modifier proliferation: when two classes of words can be modified by the same class of modifiers, they should receive a single category;

Over-generation: the atom set should not overgeneralise to accept ungrammatical examples;

Efficiency: the representation may be motivated by the needs of applications such as parsers.

Table 2 shows the eight atomic categories chosen for our corpus. Two of these categories: *LCP* (localisers) and *M* (measure words) have variously been argued to be special sub-classes of nouns (Huang et al., 2008). However, based on our overgeneration criterion, we decided to represent these as atomic categories.

We adopt the bare/non-bare noun distinction from Hockenmaier and Steedman (2007) on parsing efficiency grounds. Although they roughly correspond to English *PP*s, the distributional differences between *PP*s, *LCP*s and *QP*s justify their inclusion as atoms in Chinese. Future work in training a wide-coverage parser on Chinese CCGbank will evaluate the impact of these choices.

Labelling algorithm

We developed a recursive algorithm which applies one of several labelling functions based on the markers on a node and its children.

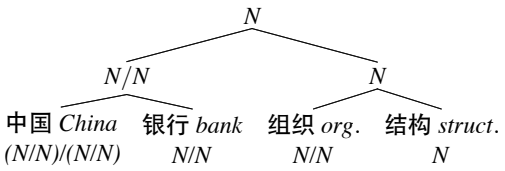
The algorithm proceeds top-down and assigns a CCG category to every node. The markers on a node's children are matched against the schema of Table 3, applying the categories of the matching schema to the children. The algorithm is then called recursively on each child. If the algorithm is called on an unlabelled node, the mapping from PCTB tags is used to assign a CCG category.

Predication	$C \to L\ C\backslash L$	Left absorption	$C \to p\ C$
Adjunction	$C \to C/C{:}\mathtt{a}\ C$	Right absorption	$C \to C\ p$
Right adjunction	$C \to C\ C\backslash C{:}\mathtt{a}$	Coordination	$C \to C{:}\mathtt{c}\ C[conj]$
Head-initial	$C \to C/R{:}\mathtt{h}\ R$	Partial coordination	$C[conj] \to conj\ C{:}\mathtt{c}$
Head-final	$C \to L\ C\backslash L{:}\mathtt{h}$	Apposition	$NP \to XP{:}\mathtt{A}\ NP$

Table 3: Category labelling schemata

Left- and right-absorption are non-CCG rules which functionally ignore punctuation, assuming that they project no dependencies and combine to yield the same category as their non-punctuation sibling (Hockenmaier and Steedman, 2007). In the schema, p represents a PCTB punctuation POS tag.

NPs receive a head-final bracketing (by our right-branching assumption), respecting NP internal structure where provided by PCTB:



6 Post-processing

A number of cases remain which are either not covered by the general translation algorithm, or otherwise could be improved in a post-processing step. The primary disharmony at this stage is the presence of *traces*, the empty categories which the PCTB annotation style uses to mark the canonical position of extraposed or deleted constituents. 19,781 PCTB derivations (69.9%) contain a trace. Since CCG aims to provide a transparent interface between surface string syntax and semantics, traces are expressly disallowed (Steedman, 2000). Hence, we eliminate traces from the annotation, by devising alternate analyses in terms of categories and combinatory rules.

Subject/object extraction

8966 PCTB derivations (31.7%) contain a subject extraction, while 3237 (11.4%) contain an object extraction. Figure 3 shows the canonical representation of subject extraction in the PCTB annotation style. The PCTB annotation follows the X' analysis of the relative clause construction as described by Wu (2004), which we transform into an equivalent, trace-free CCG analysis.

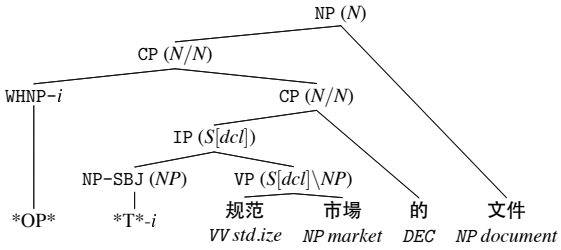


Figure 3: '*the document which standardises the market*'

First, the *Spec* trace, WHNP-i, coindexed with the extracted argument(s), is deleted. Next, the extracted argument(s) with matching indices are deleted, and category structure is adjusted to generate the correct gap category.

Modifier categories

Under our analysis, aspect particles such as 了 *le* (perfective) and 过 *guo* (experiential) are verbal post-modifiers, corresponding to *right adjunction* in Table 3. Accordingly, an aspect particle following a transitive verb VP/NP will receive the modifier category $(VP/NP)\backslash(VP/NP)$. Under this analysis, every verbal category gives rise to one possible modifier category for each aspect particle, leading to detrimental categorial ambiguity.

However, the generalised backward crossed composition combinator (Steedman, 2000) lets aspect particles retain their canonical category $(S\backslash NP)\backslash(S\backslash NP)$ regardless of the arity of the verb they modify.

Transformations

The PCTB annotation style posits traces to account for gapping, control/raising, argument sharing, pro-drop and topicalisation. To effect the parsimonious CCG analyses of Section 4, structural transformations on the original PCTB trees are necessary to accommodate the new analyses.

We developed a `tgrep`-like language which identifies instances of Chinese constructions, such as right node raising and pro-drop, whose PCTB annotation posits traces. The local trees are then reshaped to accommodate trace-free CCG analyses.

7 Evaluation

This section explores the coverage characteristics of Chinese CCGbank, in comparison with the English and German CCGbanks generated by Hockenmaier. Our analysis follows Hockenmaier (2006) in establishing *coverage* as the metric reflecting how well the target corpus has accounted for constructions in the source corpus.

7.1 Corpus coverage

The Chinese CCGbank conversion algorithm completes for 28,227 of the 28,295 (99.76%) PCTB trees. Annotation noise, and rare but legitimate syntax, such as ellipsis, account for the coverage lost in this phase. Following Hockenmaier and Steedman (2005), we adjust the PCTB annotation only for systematic tagging errors that lead to category mis-assignments, maintaining as far as possible the PCTB bracketing.

269 derivations (0.95%) contain unresolved traces, resulting from annotation noise and rare constructions (such as ellipsis) not currently handled by our translation algorithm. In 468 (1.66%) derivations, residues of PCTB tags not eliminated by the translation algorithm generate malformed categories outside the allowed set (Table 2). Excluding these cases, our conversion algorithm results in a corpus of 27,759 (98.1%) valid derivations.

7.2 Category set

The Chinese CCGbank category set is compared against existing CCG corpora derived from similar automatic corpus conversions, to determine how

well we have generalised over syntactic phenomena in the source corpus.

A total of 1197 categories appear in the final corpus, of which 329 occur at least ten times, and 478 are attested only once. By comparison, English CCGbank, contains 1286 categories, 425 of which occur at least ten times, and 440 only once, while German CCGbank has a category inventory of 2506 categories, with 1018 attested only once.[5]

7.3 Lexicon coverage

Lexical item coverage establishes the extent to which data sparsity due to unseen words is problematic in the source corpus, and hence in any corpus derived from it. Hockenmaier and Steedman (2001) showed that formalisms with rich tagsets, such as CCG, are particularly sensitive to this sparsity – while a lexical item may be attested in the training data, it may lack the necessary category.

We divided the 27,759 valid derivations into ten contiguous sections, performing ten-fold cross-validation to determine the coverage of lexical items and CCG categories in the resulting corpus.

Average coverage on lexical items is 73.38%, while average coverage on categories is 88.13%. 94.46% of token types from the held-out set are found in the training set. These figures compare to 86.7% lexical coverage (by type) and 92% (by token) in German CCGbank (Hockenmaier, 2006). Although lexical coverage by token is comparable to the German corpus, we observe a marked difference in coverage by type.

To explain this, we examine the most frequent POS tags among the missing tokens. These are NN (common nouns; 16,552 tokens), NR (proper noun; 8458), VV (verb; 6879), CD (numeral; 1814) and JJ (adjective; 1257). The 100 most frequent missing tokens across the ten folds comprise 48 NR tokens, 46 NR, 3 NT (temporal nouns), 2 JJ (adjectives) and one VA (verbal adjective). Personal names are also not tokenised into surnames and forenames in the PCTB, increasing unseen NR tokens.

The missing VVs (verbs) include 1342 *four-character compounds*, fossilised idiomatic expressions which are considered atomic verbs in the PCTB annotation. Another source of verb sparsity stems from the PCTB analysis of verbal infixation. Given a polysyllabic verb (e.g. 离开 *leave-away* "leave"), we can add the adverbial infix 不 *not* to form a potential verb 离不开 *leave-not-away* "unable to leave". In the PCTB annotation, however, this results in lexical items for the two cleaved parts, even though 离 *leave* can no longer stand alone as a verb in modern Chinese. In this case, a morphologically decomposed representation which does not split the lexical item could mitigate against this sparsity. Alternatively, candidate verbs for this construction could have the first verb fragment subcategorise for the second.

[5] All German verbs having at least two categories to account for German verbal syntax contributes to the greater size of the category set (Hockenmaier, 2006).

8 Conclusion

We have developed the first analysis of Chinese with Combinatory Categorial Grammar, crafting novel CCG analyses for a range of constructions including topicalisation, pro-drop, zero copula, verb compounding, and the long-range dependencies resulting from the 把 *ba-* and 被 *bei*-constructions.

We have presented an elegant and economical account of Chinese syntax that exploits the power of CCG combinatory rules, supporting Steedman's claim to its language-independence.

We have designed a conversion algorithm to extract this analysis from an existing treebank, avoiding the massive cost of hand re-annotation, creating a corpus of 27,759 CCG derivations, covering 98.1% of the PCTB. The corpus will be publicly released, together with the converter, providing the tools to create CCGbanks in new languages.

At release, Chinese CCGbank will include gold-standard head co-indexation data, as required for the training and evaluation of head-driven dependency parsers. Co-indexation analyses, like those provided for the 把 *ba-* and 被 *bei*-constructions, will be extended to all categories.

Future refinements which could be brought to bear on Chinese CCGbank include the integration of PropBank data into CCGbank (Honnibal and Curran, 2007; Boxwell and White, 2008) using Chinese PropBank (Xue, 2008). The *hat categories* of Honnibal and Curran (2009) may better handle form/function discrepancies such as the Chinese zero copula construction, leading to cleaner, more general analyses.

We have presented a wide-coverage Chinese corpus which exploits the strengths of CCG to analyse a range of challenging Chinese constructions. We are now ready to develop rich NLP tools, including efficient, wide-coverage CCG parsers, to address the ever-increasing volumes of Chinese text now available.

Acknowledgements

James Curran was supported by Australian Research Council (ARC) Discovery grant DP1097291 and the Capital Markets Cooperative Research Centre.

References

Jason Baldridge. 2002. *Lexically Specified Derivational Control in Combinatory Categorial Grammar*. Ph.D. thesis, University of Edinburgh.

Stephen Boxwell and Michael White. 2008. Projecting Propbank roles onto the CCGbank. *Proceedings of LREC 2008*.

Michael Burke and Olivia Lam. 2004. Treebank-based acquisition of a Chinese lexical-functional grammar. In *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation*, pages 161–172.

Aoife Cahill, Mairead McCarthy, Josef van Genabith, and Andy Way. 2002. Automatic annotation of the Penn Treebank with LFG F-structure information. In *LREC 2002 Workshop on Linguistic Knowledge Acquisition and Representation-Bootstrapping Annotated Language Data*, pages 8–15.

Jeongwon Cha, Geunbae Lee, and Jonghyeok Lee. 2002. Korean combinatory categorial grammar and statistical parsing. *Computers and the Humanities*, 36(4):431–453.

John Chen, Srinivas Bangalore, and K. Vijay-Shanker. 2005. Automated extraction of Tree-Adjoining Grammars from treebanks. *Natural Language Engineering*, 12(03):251–299.

Stephen Clark and James R. Curran. 2007. Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models. In *Computational Linguistics*, volume 33, pages 493–552.

Yuqing Guo, Josef van Genabith, and Haifeng Wang. 2007. Treebank-based acquisition of LFG resources for Chinese. In *Proceedings of LFG07 Conference*, pages 214–232.

Julia Hockenmaier. 2003. *Data and Models for Statistical Parsing with Combinatory Categorial Grammar*. Ph.D. thesis, University of Edinburgh.

Julia Hockenmaier. 2006. Creating a CCGbank and a wide-coverage CCG lexicon for German. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 505–512. Morristown, NJ, USA.

Julia Hockenmaier and Mark Steedman. 2001. Generative models for statistical parsing with combinatory categorial grammar. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 335–342. Association for Computational Linguistics, Morristown, NJ, USA.

Julia Hockenmaier and Mark Steedman. 2002. Acquiring compact lexicalized grammars from a cleaner treebank. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1974–1981.

Julia Hockenmaier and Mark Steedman. 2005. CCGbank: Users' manual. Technical report, MS-CIS-05-09, Computer and Information Science, University of Pennsylvania.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Beryl Hoffman. 1996. *The computational analysis of the syntax and interpretation of free word order in Turkish*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Matthew Honnibal and James R. Curran. 2007. Improving the complement/adjunct distinction in CCGbank. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING-07)*, pages 210–217.

Matthew Honnibal and James R. Curran. 2009. Fully Lexicalising CCGbank with Hat Categories. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1212–1221.

C.-T. James Huang, Y.-H. Audrey Li, and Yafei Li. 2008. *The syntax of Chinese*. Cambridge University Press.

Roger Levy and Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 439–446. Morristown, NJ, USA.

Charles N. Li and Sandra A. Thompson. 1989. *Mandarin Chinese: A functional reference grammar*. University of California Press.

David M. Magerman. 1994. *Natural language parsing as statistical pattern recognition*. Ph.D. thesis, Stanford University.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Yusuke Miyao, Takashi Ninomiya, and Jun'ichi Tsujii. 2004. Corpus-Oriented Grammar Development for Acquiring a Head-Driven Phrase Structure Grammar from the Penn Treebank. pages 684–693.

Ad Neeleman and Kriszta Szendrői. 2007. Radical pro drop and the morphology of pronouns. *Linguistic Inquiry*, 38(4):671–714.

Mark Steedman. 1987. Combinatory grammars and parasitic gaps. *Natural Language & Linguistic Theory*, 5(3):403–439.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press. Cambridge, MA, USA.

Henry H.Y. Tiee and Donald M. Lance. 1986. *A reference grammar of Chinese sentences with exercises*. University of Arizona Press.

David Vadas and James R. Curran. 2007. Adding noun phrase structure to the Penn Treebank. In *Association for Computational Linguistics*, volume 45, page 240.

Xiu-Zhi Zoe Wu. 2004. *Grammaticalization and language change in Chinese: A formal view*. Routledge.

Fei Xia. 1999. Extracting tree adjoining grammars from bracketed corpora. In *Proceedings of Natural Language Processing Pacific Rim Symposium '99*, pages 398–403.

Nianwen Xue. 2008. Labeling chinese predicates with semantic roles. *Computational Linguistics*, 34(2):225–255.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(02):207–238.

Nianwen Xue, Fei Xia, Shizhe Huang, and Anthony Kroch. 2000. The Bracketing Guidelines for the Penn Chinese Treebank (3.0). *IRCS Report 00-08, University of Pennsylvania*.

Po Ching Yip and Don Rimmington. 1997. *Chinese: An essential grammar*. Routledge.