Language, Information and Computation(PACLIC 11), 1996, 363-371

Principle-based Parsing for Chinese

Charles D. Yang and Robert C. Berwick
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
545 Technology Square, NE43-769
{charles, berwick}@ai.mit.edu

Abstract

This paper describes the implementation of Mandarin Chinese in the Pappi system, a principle-based multi-lingual parser. We show that substantive linguistic coverage for new and linguistically diverse languages such as Chinese can be achieved, conveniently and efficiently, through parameterization and minimal modifications to a core system. In particular, we focus on two problems that have posed hurdles for Chinese linguistic theories. A a novel analysis is proposed for the so-called BA-construction, along with a principled computer implementation. For scoping ambiguity, we developed a simple algorithm based on Jim Huang's Isomorphic Principle. The implementation can parse fairly sophisticated sentences in a couple of seconds, with minimal addition (less than 100 lines of Prolog code) to the core parser. This study suggests that principle-based parsing systems are useful tools for theoretical and computational analysis of linguistic problems.

1 Introduction

Natural languages are complex and their syntactic properties seem to differ from one another quite dramatically. Traditional parsing technologies utilize language-particular, rule-based formalisms, which usually result in large and inflexible systems (Marcus 1980). For a recent example of the rule-based approach to Chinese parsing, see (Lee et al., 1991).

In recent years, computational linguistics has seen the development of the Principle-based Parsing (Berwick et al 1991). A principle-based parser transparently reflects the structure of the contemporary linguistic theory, the Principles and Parameters framework (Chomsky 1981). It is believed that languages are constrained by a small number of universal principles, with linguistic variations largely specified by parametric settings.

The merit of principle-based parsing is two-fold. As a tool for linguists, it is directly rooted in grammatical theories. Therefore, linguistic problems, particularly those that involve complex interactions among linguistic principles, can be cast in a computational framework and extensively studied by drawing directly on an already-substantiated linguistic platform. It is designed from the start to accommodate a wide range of languages — not just 'Eurocenteric'

Romance or Germanic languages. Japanese, Korean, Hindi and Bangla have all been relatively easily modeled in PAPPI (Berwick and Fong 1991, Berwick forthcoming). As a tool for engineering, it inherits the useful design principle of *modularity* among at least some of the principles. Differences among languages reduce to distinct dictionaries, required in any case, plus parametric variation in the principles.

To show how this project may be concretely executed for Chinese, we first review the basic phrase structures and parameters for Chinese. Our principle source is Jim Huang's seminal dissertation (1982). Then we turn to two particular problems that have generated much interest and productive work in the literature, the BA-construction and the scoping non-ambiguity in Chinese. Both theoretical and computational analyses are presented. The paper concludes with some general observations on principle-based parsing, in relation to linguistics and computation.

2 Structures and Parameters in Chinese

"Chinese exhibits a full range of head-final constructions, but allows only a limited range of head-initial construction" (Huang 1982). Using the X-bar schemata for phrase structure, we have:

(1) (a) $XP \rightarrow Spec\ \overline{X}$

(b) $\overline{X} \rightarrow X\ \ Comp, X \neq noun$

The basic $\overline{X}$ parameters indicate that Chinese has roughly the same word order as English, except that in the noun phrase case the modifier always *precedes* the noun in Chinese. These parameters are specified in a separate file for the parser:

```
%% X-Bar Parameters
specInitial.
specFinal :- \+ specInitial.

headFinal(n).
headFinal(c).
headInitial(X) :- \+ headFinal(X).
```

In Chinese, Complex noun phrases consist of a sequence of modifiers (e.g. possessives, adjectives, relative clauses, as in (2.a-c) respectively), followed by a marker **de**:

(2) (a) ta de che (his car)

(b) meili de yanjing (beautiful eyes)

(c) ta xihuan de che (the car he likes)

Chinese possessives, such as *ta* in (2.a), are *not* reflected in surface morphological features, contrary to English (e.g. *she* → *her*). Therefore, one must be cautious because the parser can have no *visible* information to identify possessives, which then must nevertheless still receive case as in English (*she/her*). To deal with this problem, we assume that the morphological marker **de** assigns genitive case to the noun or clausal phrase on its left, to prevent a case filter violation.

```
lex(de,mrkr,[left(np,case(gen),[])]).
lex(de,mrkr,[left(c2,case(gen),[])]).
```

In Chinese, adverbial adjunctions take particular orders (unlike English, which is rather free): time precedes location, which in turn precedes manner:

(3) John zuotian zai xuexiao kanjian-le Mary.
John yesterday at school saw Mary.

We specify adverbial properties with a **predicate** feature in the lexicons, for example:

```
lex(zai,p,[grid([],[location]),predicate(location)]).
lex(chang,adv,[adjoin(left),predicate(manner)]).
lex(zuotian,adv,[adjoin(left),predicate(time)]).
```

A simple procedure is written to constrain adjunction orderings.

```
%% Check VP adjunct order: TIME > LOCATION > MANNER
rhs [pp(P),vp(X)] add_goals [checkVPadj(P,X)].
rhs [adv(A),vp(X)] add_goals [checkVPadj(A,X)].

checkVPadj(Adj1,VP) :-
        \+ adjoined(VP).
checkVPadj(Adj1,VP) :-
        adjoined(VP,Adj2,VP1),
        precedenceInOrder(Adj1,Adj2),
        checkVPadj(Adj2,VP1).

precedenceInOrder(Adj1,Adj2) :-
        Adj1 has_feature predicate(time),
        Adj2 has_feature predicate(location).
precedenceInOrder(Adj1,Adj2) :-
        Adj1 has_feature predicate(time),
        Adj2 has_feature predicate(manner).
precedenceInOrder(Adj1,Adj2) :-
        Adj1 has_feature predicate(location),
        Adj2 has_feature predicate(manner).
```

Like Japanese but unlike English, Wh-movement does not occur overtly in Chinese:

(4) (a) Ni da-le shui? (You beat who?)

(b) Who did you beat *t*?

This parametric distinction constitutes an important module in the principle-based theory. It has been directly implemented in Pappi. We simply set the parameter to:

```
%% Wh In Syntax Parameter

no whInSyntax.
```

Conventionally, *Bounding Nodes* represent a constraint on how far elements can be displaced from their canonical argument positions, like *who* in *Who did you see*? Following standard linguistic analyses, in Chinese the bounding nodes are the inflection phrase (IP) and NP, similar to Dutch and Italian. Again, this is straightforwardly implemented in the current system; we simply set the bounding node parameters for Chinese:

```
%% Subjacency Bounding Nodes
boundingNode(i2).
boundingNode(np).
```

These central aspects of Chinese phrase structure are hence readily handled by parametric settings and simple goal-adding mechanisms. We next turn to specific case studies — beginning with the BA-construction.

3 The BA-construction: Theory and Computation

The BA-construction in (5) has been a perennial problem for Chinese linguists.

(5) Wo ba John da-le. (I had John beat)

It is perhaps impossible to give a purely syntactic analysis to account for all the rich semantic and pragmatic factors (Ding 1991) in BA-constructions. Huang's analysis (1982) relies on phrase structure constituency. Li (1990) suggests that BA-constructions as base-generated. We propose a different approach. We try to derive BA-construction from syntactic theories on the basis of some semantic considerations. The aim is to show that a complex type of construction can be derived from syntactic principles, unless theoretically proven otherwise. Stipulations such as base-generation rules are taken as a "last-resort", which involves language-particular peripheral mechanisms, an effort we try to avoid or postpone in principle-based syntactic analysis.

The first step is to determine the categorical nature of **ba**. There are two obvious possibilities: **ba** as a preposition or as a verb. It is not difficult to eliminate the former, on the basis of the so-called Theta Criterion (Chomsky 1981), that every thematic role like Agent, Patient, etc. must be uniquely assigned and every argument must be receive one and only one theta role (a

standard assumption also used many current theories, such as LFG, HPSG, and the like). If **ba** were a preposition, its thematic role (oblique) must be discharged. The only possible recipient of its thematic role is *John*, which has already received the patient role from *da* – a contradiction. Thus **ba** must be a verb.

The second step draws motivations from semantic considerations. We note that in BA-constructions, there is an observable *semantic predicate shift* of main (sentential) action emphasis from the transitive verb phrase in (6a) to the verbal phrased headed by **ba** (dubbed BaP henceafter) in (6b)

(6) (a) Wo [$_{VP}$ da-le John].

(b) Wo [$_{BaP}$ ba John da-le].

On surface, *John* superficially looks like the direct object of **ba**. However, although it is reasonable to assume BaP becomes the main action predicate of the sentence (as the result of *semantic predicate shift*), *John* still receives its thematic role from *da* (beat) in semantic interpretation, contrary to the superficial intuition. Therefore, we conclude that *John* must have moved from its original thematic position in (5).

The third step answers the following question: If *John* is displaced in (5), what **triggers** this movement from its original thematic position? This dislocation cannot be driven to place *John* in a thematic-role receiving position, since *John* is understood as the recipient of the patient-role from **da** Therefore, the only syntactic parsing option we are left with for must be case-related.

The question then becomes, Why cannot *John* receive case from its thematic role assigner, *da*? This leads to the fourth and last step in our analysis. We propose that in the BA-construction, the original action verb (e.g. *da* in (5)) loses its case-assigning ability, much like **BEI** passivized verbs or passive verbs in English. We draw this conclusion from the well-known **BA** and **BEI** construction parallelism, as follows.

As usually manifested in routine Chinese grammar exercises, there exists a one-to-one mapping between **BA** and **BEI** constructions. That is, any **BA** construction has a **BEI** counterpart – passivization in Chinese, and vice versa:

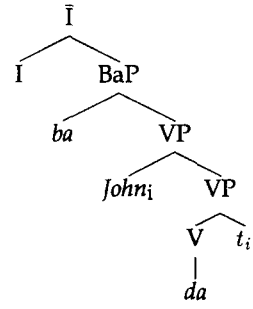
(7) (a) Wo **BA** John da-le. (I beat John.)

(b) John **BEI** wo da-le. (John was beaten by me.)

Given this parallel contrast, we suppose that the BA-construction is in fact analogous to the BEI-construction, i.e. passivization. As quite standardly assumed, passivized verbs lose their case-marking ability; thus, the verb in the BA-construction loses its case-marking ability as well.

Therefore, a clear analysis is in sight. In BA constructions, the action verb (*da* in (5)) loses case-assigning ability as in passivization; its direct object is forced to move to the object position of **BA** to receive case. **BA**, on the other hand, is a light verb that has no thematic role to assign, but is able to Exceptionally Case Mark (ECM) into a small clause, i.e. the action verbal phrase:

(8)



Having worked out a plausible theory, implementation becomes straightforward. We take the following steps:

(9) (a) Mark the lexical features of **ba** with the properties:
- take a verb phrase small clause
- have 'exceptional' case marking ability
- blocks the verb's case-marking ability

(b) Allows overt NP (the object) to adjoin VP, creating the subject position for the small clause to facilitate ECM to take place.

4 Scoping Unambiguity

The well-known quantifier scoping ambiguities in English do not exist in their Chinese counterparts:

(10) (a) Every man loves a woman.
i. For every man x, there exists **one** woman y such that x loves y.
ii. For every man x, there exists **a** woman y, such that x loves y.

(b) Meige nanren dou xihuan yige nuren.
"Everyman likes **one** woman."

A valuable analysis is suggested by Aoun and Li (1992). They assume that the subject is generated within the verbal projection (VP), and that in English, the subject raises the Specifier position (i.e., the traditional 'subject' position; 'specifier' is the familiar X-bar terminology for the same) of the inflection phrase (IP) but not in Chinese, because Chinese has no subject-verb agreement and thus is "inflectionally impoverished". This (parametric) optionality allows multiple landing sites in English, but not in Chinese, for Quantifier Raising (May 1985) that occurs at LF – thus, ambiguity in English and non-ambiguity in Chinese (see (Aoun and Li 1992) for details).

This analysis, although reasonable and principled, is fairly complex and requires a few more crucial syntactic principles that have yet been implemented in PAPPI. Instead, in order to maintain system integrity and computational efficiency, we develop a practical solution that attributes the scoping contrast to Huang's Isomorphic Principle (1982):

(11) **The Isomorphic Principle**
If a quantifier phrase QP A *c-commands* QP B at phrase structure (S-structure), then it does so at logical form (LF) as well.

Crucially, we view the Isomorphic Principle as a **parametric** variation between Chinese and English. In particular, we assume that

(12) (a) The Isomorphic Principle is applicable to QR in Chinese.
(b) The Isomorphic Principle is *not* applicable to QR in English.

We introduce a language parameter, `isomorphicPrinciple` and set it **TRUE** for Chinese and **FALSE** for English, in their parameter files, respectively. This parameter is then called upon as a LF condition on QR.

```
lfMovement(SS,LF) :-
        qr(SS,SS1),
        checkIsomorphic(SS1) if isomorphicPrinciple,
        moveWh(SS1,LF).

checkIsomorphic(SS1) :-
        qrT(SS1,TIs),
        qrQ(SS1,QIs),
        TIs == QIs.
qrT collect I in_all_configurations CF where CF has_feature ec(qr),
        CF has_feature index(I).
qrQ collect I in_all_configurations CF where CF has_feature moved(qr),
        CF has_feature index(I).
```

In Chinese, we observe that QP α c-commands β if and only if α precedes β at LF and S-Structure. In the implementation above, we collect two lists of item: one with QPs after QR (precedence at LF), the other with QPs at base positions (precedence at S-Structure). If the Isomorphic Principle holds, the two lists must have identical orderings. In other words, we reconstruct S-Structure information from LF representations. The computational effort is minimal.

5 Results and conclusions

Overall, the implementation is quite simple, summarized as follows:

1. A dictionary is constructed, which contains lexical entries for words used in the implementation.

2. Chinese-particular parametric values are determined through linguistic literature, for the X-bar theory, the Bounding theory, the movement theory, etc.

3. Some Chinese-particular, linguistically interesting problems are considered, e.g. the BA-construction and scoping unambiguity. Theoretical solutions are proposed and implemented as language peripheral component of the parsing system

We are able to parse sentences with the range of structures including Wh movement, the Binding Theory, Quantifier Scoping, the BA-construction to complex NP (clausal, possessive, and numeral/classifier). All testing sentences are correctly analyzed: LF logical form representations are computed for the grammatical sentences and the ungrammatical ones are ruled out ones with linguistic principle violation(s) shown. Each parse takes no more than 2 seconds on a Sparc 10 workstation. Overall, excluding a hand-wired dictionary, less than 100 additional lines of Prolog are required.

Because the Pappi system implements its model linguistic theory faithfully, adapting new languages is expected to be quite minimal, as our implementation shows. Additionally, it provides a platform on which linguists can experiment theoretical proposals extensively and also cross-linguistically, without having to know much about the internal design of the parser. Furthermore, principle-based systems output very rich and accurate structural descriptions, including logical form representations, that assist in more engineering-oriented NLP tasks that go beyond parsing (Lin 1995).

6 References

1. Aoun, Joseph and Audrey Li. (1992). *The Syntax of Scope.* MIT Press, Cambridge, MA.

2. Berwick, Robert C., editor. (forthcoming). *Principle-Based Parsing: From Theory to Practice.* Kluwer Academic Publishers, Norwell, MA.

3. Berwick, Robert C., Abney, Steven, and Carol Tenny., editors (1991). *Principle-Based Parsing: Computation and Psycholinguistics.* Kluwer Academic Publishers, Norwell, MA.

4. Berwick, Robert C., and Sandiway Fong. (1993). Madama Butterfly Redux: parsing English and Japanese with a principles and parameters approach. In R. Mazuka and N. Nagai (eds). *Japanese Sentence Processing.* Hillsdale, NJ: Erlbaum. 177–208.

5. Chomsky, Noam. (1981). *Lectures on Government and Binding.* Foris, Dordrecht, Holland.

6. Ding, Shizhi. (1991). Ba-constructions typology in mandarin Chinese: A government-binding approach. Master's thesis, University of British Columbia.

7. Fong, Sandiway. (1991). *Computational Properties of Principle-Based Grammatical Theories*. PhD thesis, MIT, Cambridge, MA.

8. Huang, James. (1982). *Logical Relations in Chinese and the Theory of Grammar*. PhD thesis, MIT, Cambridge, MA.

9. Lee, L. S., Chien, L. F., Lin, L. J., Huang, J., and Chen, K. J. (1991). An efficient natural language processing system specially designed for the Chinese language. *Computational Linguistics*, 17(4).

10. Li, Audrey. (1990). *Order and constituency in Mandarin Chinese*. Kluwer Academic Publishers, Norwell, MA.

11. Lin, Dekang. (1995). Talk given at the ARPA Message Understanding Conference (MUC).

12. Marcus, Mitch. (1980). *A Theory of Syntactic Recognition for Natural Language*. MIT Press, Cambridge, MA.

13. May, Robert. (1985). *Logical Form*. MIT Press, Cambridge, MA.