Information-based Case Grammar

Keh-jiann CHEN
Institute of Information Science
Academia Sinica
Taipei, Taiwan
R.O.C.

Chu-Ren HUANG
Institute of History and Philology
Academia Sinica
Taipei, Taiwan
R.O.C.

KCHEN%IS@TWNCTU01.BITNET

HSCHUREN@TWNAS886.BITNET

Abstract: In this paper we propose a framework of Information-based Case Grammar (ICG). This grammatical formalism entails that the lexical entry for each word contain both semantic and syntactic feature structures. In the feature structure of a phrasal head, we encode syntactic and semantic constraints on grammatical phrasal patterns in terms of thematic structures, and encode the precedence relations in terms of adjunct structures. Such feature structures denote partial information which defines the set of legal phrases. They also provide sufficient information to identify thematic roles. With this formalism, parsing and thematic analysis can be achieved simultaneously. Due to the simplicity and flexibility of Information-based Case Grammar, context dependent and discontinuous relations such as agreements, coordinations, long-distance dependencies, and control and binding, can be easily expressed. ICG is a kind of unification-based formalism. Therefore it inherits the advantages of unification-based formalisms and more.

1. Introduction

In this paper, we propose a lexicon-based grammatical formalism called Information-based Case Grammar (ICG). This formalism entails that the lexical entry for each word contain both semantic and syntactic information. It will be argued that our lexicon-based representation approach better focuses information for parsing and generation. In contrast, the phrase-structure rule approaches lack rule focusing capability. Even with the LR parsing strategy [Tomita 86], unnecessary branching and backtracking cannot be avoided when adopting these approaches. Therefore, modern linguistic theories share the tendency to be lexicon-based and to reduce PS rules. For instance, LFG and HPSG stipulate the argument structure as one of the attributes for each verb [Bresnan 82, Pollard 87] and Karttunen [Karttunen 86] proposes a radical lexicalism's approach to do without PS rules by representing syntactic information in terms of categorial grammar. Categorial grammar, however, offers no straightforward and intuitive mechanisms to handle context dependent or discontinuous relations such as control and long distance dependency [Uszkoreit 86]. Hence, we adopt an alternative approach equivalent to the ID/LP (immediate dominance and linear precedence) format of GPSG [Gazdar 87].

In the feature structure of a phrasal head, we encode syntactic and semantic constraints on grammatical phrasal patterns in terms of thematic structures, and encode the precedence relations in terms of adjunct structures. The feature structure of a potential phrasal head denotes partial information for defining the set of legal/grammatical phrases. It also provides enough information to identify the thematic roles for arguments and adjuncts [Chen 89]. In other words, with ICG, parsing and thematic analysis are achieved simultaneously without additional operation; and generation with thematic structure can be done with the identical formalism.

We take Mandarin Chinese as our representational target. Thus, the features were selected to account for Chinese only. However, the abstract design of this formalism is not limited to only the representation of Chinese. Since the Chinese lexicon is impoverished in inflection, it is necessary to fully stipulate both semantic and syntactic information for the purpose of both parsing and generation. Furthermore the precedence relationship of constituents is defined over thematic roles. This seems to be more appropriate for Chinese. By coincidence, Bresnen and Kanerva's [Bresnan 89] lexical mapping theory represents a shift towards the possibility of semantics major approaches.

2. Feature Structures for Mandarin Chinese

Chinese is a weakly marked language with no inflection. Nevertheless, the linear order of arguments and adjuncts are relatively free. Hence syntactic-only representations would cause tremendous ambiguities. For representational precision and for parsing adequacy, semantic information is indispensible. The most important semantic information includes 1. argument structures and their semantic restrictions, and 2. the semantic features for each word which are necessary to identify thematic roles. Hence the following feature structure (1) was selected so that each lexical entry can be uniformly represented by the same structure with lexically or syntactically defined value (including null).

```
(1)   Semantic: Meaning:
                Features:
                Arguments:
                Adjuncts:
      Syntactic: Class:
                 Constraints: Form:
                              Basic Patterns:
                              Adjunct Precedence:
```

A typical example (2) is given here before we introduce and summarize the values for each feature path below.

(2) *Chiuan* "persuade":

```
Semantic:  meaning:  "persuade"
           features:
           arguments: AGENT:  feature:  +Human
                      GOAL :    [1] feature: +Human
                      THEME:    feature:  +Event
                                  argument: AGENT: [1]
           adjuncts: time, location, manner,....

Syntactic:  class: V_NP,VP

           constraints: form: time [{NP, DM, PP, GP, ADV}, +time]
                              location [{PP, ADV}, +location]
                              manner [ADV, +manner]
                                                .
                                                .
                        BP: AGENT [NP] < * < GOAL [NP] < THEME [VP]
                        AP: 1. {time, location} < *
                            2. AGENT < manner < *
```

Meaning: an atomic value denoting the meaning of the word.

Features: a set of atomic values which are the semantic features of the word; e.g. ± animate, ± physical.

Arguments: a set of thematic arguments for the head if the word is a possible phrasal head; null otherwise. The value for each thematic argument is a feature structure of the same type as the value for the feature path 'semantic'.

E.g. 'a nice boy' is an agent, as in 'A nice boy drove the cattle.'

(3) AGENT:

```
meaning: 'boy'
features: +Human
arguments: none
adjuncts: quantifier: meaning: 'a'
                      features: -definite, +singular
          property: meaning: 'nice'
```

Argument structures for a verb are equivalent to case frames and case restrictions of this verb [Fillmore 68, Winograd 83]. Case restrictions indicate semantic preferences of thematic roles and function as a guide to identifying each case role [Chen 89]. The argument structures for other phrasal heads such as prepositions, post-positions and conjunctions serve similar purposes.

Adjuncts: a set of permissible adjuncts of the head word. The value of each adjunct is a feature structure of the same type as arguments.

E.g. 'yesterday'

(4) time:

```
meaning: 'yesterday'
features: +time
arguments: none
adjuncts: none
```

Arguments, adjuncts and head form the right-hand side of an immediate dominance rule.

Syntactic Class: atomic values denoting the syntactic class of the word.

The syntactic class of a word serves two purposes. The first is to denote the syntactic type. The second is as an index for inheriting common syntactic properties belonging to the mother node in the syntactic hierarchy.

Syntactic Form: a set of syntactic and semantic definitions for arguments and adjuncts.

The syntactic form for each thematic role in fact constrains the syntactic structures and semantic features of this role. We consider semantic restriction as part of the syntactic constraint. For instance, temporal expressions are instantiated by at least five different syntactic categories which are nouns phrase (NPs), compounds with determinatives and measures (DM, such as *San-dian Shi-fen* three-hours ten-minute, i.e. 'three-ten'), post-position phrases (GPs), preposition phrases (PPs), or adverbs (ADVs). They all share a common semantic feature +time regardless of their categories. Therefore the temporal expressions can be expressed as time [{NP, DM, PP, GP, ADV}, +time]. For complex expressions, we adopt the choice system used by Halliday in the systemic grammar [Winograd 83].

Basic Patterns: a set of linear precedence rules governing arguments and heads.

The basic patterns for verbs denote the possible sentential patterns, including optional argument omission. The familiar syntactic generalizations with regard to passivization, topicalization etc. can be captured by lexical rules [Gazdar 85, Pollard 87, Sells 85]. Adopting the lexical mapping theory, however, is a theoretical possibility to make basic patterns obsolete [Bresnan 89, Huang 89].

Adjunct Precedence: a set of linear precedence and cooccurrence constraints for adjuncts.

The following notations were adopted.

* : denotes a phrasal head.

< : e.g. "a < b" denotes a precedes b.

<< : e.g. "a << b" denotes a immediately precedes b.

>< : e.g. "a >< b" denotes a and b can not cooccur.

{ }: e.g. "a < {b,c}" denotes a precedes both b and c but there are no precedence constraints between b and c.

Although Basic Patterns and Adjunct Precedence both govern linear precedence relations, they differ in their nature. Basic Patterns encode phrasal heads and their arguments. Linear order is but an additional piece of information describing the pattern. It is conceivable, i.e. in non-configurational languages, that linear precedence relations play no role in Basic Patterns. Adjunct Precedence Rules, on the other hand, are constraints on linear precedence relations among possible adjuncts. In other words, linear order is central to Adjunct Precedence statements while the presence of each adjunct is not. Thus, a legal phrase can be viewed as a sequence of thematic roles arranged in a proper order defined by one of the basic patterns and satisfying all the constraints of Adjuncts Precedence when applicable. The division of Basic Patterns from Adjunct Precedence is similar to the ID/LP format with the additional information differentiating adjunts from arguments. The limited numbers of thematic cases, syntactic categories, and semantic features require only a finite number of notational symbols [Gazdar 87]. Therefore we claim that ICG falls into the class of context-free grammars.

The semantic features would be unified during parsing while the syntactic features are no more than constraints guiding appropriate unification. We have a scheme to identify the thematic roles by fully utilizing the semantic and syntactic information [Chen 89]. In [Chen 89], Chen et al. propose an information accumulation scheme (incremental description refinement in [Mellish 88]) for identifying thematic roles with the parametricalized information encoded with the ICG formalism. The four types of parametric information used in Chinese are:

a. the syntactic category and semantic features of the constituent,

b. the case frame and semantic restrictions of the verb,

c. the syntactic configuration and word order, and

d. oblique case assigner, including prepositions and postpositions.

Following is the parsing result of

(5) A nice boy persuaded John to go to school yesterday.

```
meaning:  "persuade"
features:  past
arguments: AGENT: meaning: "boy"
                  feature:  +Human
                  adjuncts: quantifier: meaning: "a"
                                        feature: -definite, +singular
                            property: meaning: "nice"

           GOAL : [1] meaning: "John"
                      feature:  +Human

           THEME:  meaning: "go to"
                   feature:  +Event
                   arguments: AGENT: [1]
                              GOAL:  meaning: "school"

adjuncts: time: meaning: "yesterday"
                features:  +time
```

3. *The Formal Definition for Information-based Case Grammar: The Lexicon and Principles*

The ICG is composed of two major components. One is the lexicon which is the set of feature structures as described in section 2. The other is the principles. Each feature structure can be viewed as a set of interpretable representations of syntactic and semantic information governed by a formal syntax. A parser or generator interpretes feature structures while parsing or generating sentences. The interpretation processes are guided by the principles of the grammar. The principles define well-formedness conditions and the rules for information management for sentences and phrases. The parser or generator takes lexical information and unify it in such a way that changes of lexical information would not affect the phrasing or generating process. Therefore we claim that ICG is declarative. The major principles of ICG are summarized below:

1) Head Driven Principle

The feature structure of a head contains the partial information that defines the permissible set of phrases with it as a phrasal head. The possible head types and respective phrase types for Mandarin Chinese are as follows.

(6)

Head	Phrase
V	S, VP
N	NP
Prep	PP
Post P	GP
Determinant	DM (determinative and measurecompound)
Conjunction	CP (conjunctive phrase)

The syntactic classes of the heads determine the syntactic types of their projections. VP is defined as an S without subject [Gazdar 87].

2) Well-formedness Conditions

Like LFG [Bresnan 82], we have the following well-formedness conditions:

a. Completeness and functional biuniqueness conditions,

b. Coherence conditions,

and c. Linear precedence and syntactic form constraints.

In fact, the completeness condition is enforced with respect to Basic Patterns (BP) only. The cases of argument omission are idiosyncratically determined by verb classes, and are not governed by the completeness condition. On the other hand, adjuncts are optional and constrained only by linear precedence rules AP and form restrictions. The functional uniqueness condition is also relaxed a bit to account for cases of multiple occurrences of some adjuncts such as modal at the sentential level, and property on the noun phrase level, by annotating Kleene's star on top of the adjunct modal and the adjunct property.

3) Feature Propagation Principles

a. Head feature convention [Gazdar 85],

b. Foot feature principle [Gazdar 85],

c. Explicit feature paths:

Explicitly denoting the daughter's feature in the mother node by a feature path such as the 'argument feature',

d. Conditional features:

Locally ambiguous semantic features can be expressed by conditional features. For instance, the preposition *bei* 'by' can mark an agent, a causer, or an instrument depending upon whether the arguments are animate, nonphysical, or physical inanimate respectively. Therefore the feature structure of *bei* 'by' would be (7).

(7) *Bei* 'by':

```
Sem   features: DUMMY feature,
                1. AGENT,
                2.CAUSER,
                3.INSTRUMENT
      argument: DUMMY:features: 1. +animate
                                2. -physical
                                3. -animate

Syn   form: DUMMY [NP]
      BP: 'bei' < < DUMMY
```

There are many possible types of semantic features allowed in ICG. They are differentiated by attribute value pairs. The three required types of semantic features for Chinese are 1. semantic classes e.g. ± animate 2. syntactic/semantic feature, e.g. ± NEG 3. thematic roles, e.g. AGENT. Different types of features can be identified simply by their attribute names e.g. we know AGENT in <ARGUMENT AGENT> is a thematic role.

The head feature principle is the same as the one in GPSG [Gazdar 85] which states that in any local subtree, the head features of the mother are identical to the head features of the head daughter. Similarly, our Foot feature principle also follows GPSG. Roughly speaking, foot features are passed up from any daughter in a tree, with the upper and lower limits of this propagation are determined by prior specification [Gazdar 85]. In Chinese, ± Question, ± Negation, ± Plural, ± Definite etc. are all considered foot features.

The semantic class of a thematic role is usually determined by its head daughter. However, for the marked cases with the syntactic categories of PP/GP, the semantic classes are determined by complement daughters. We can not define every semantic type as a foot feature. Therefore, we propose to explicitly state the daughter's feature in the mother node by a feature path such as 'DUMMY features' in (7).

4. What Makes ICG a Good Representational Language

We think that the simplicity and flexibility of the ICG formalism makes it a good representational language. It is simple since ICG is a type of context-free grammar and the attributes for feature structures are uniform for all different types of phrases.

Preparation of lexical feature structures are straightforward. Linguists can start with the categorial feature structure of each entry by discovering idividual idiosyncracies and then modifying the categorial feature structure accordingly. As for flexibility, ICG is much more flexible than the other context-free grammatical forms such as BNF, GPSG, etc., due to the wider scope of accessibility and the richness of information encoding on each thematic constituent. Context dependent and discontinuous relations such as agreements, coordinations, long-distance dependencies, control and binding, can be easily expressed in ICG.

4.1 Agreement and Coordination

Agreement and coordination pose similar problems in representation (but different problems in processing). Both have to express the relations between daughters. Coordination can be viewed as the agreement of syntactic or/and semantic classes between two daughter arguments of a conjunction. In contrast, other types of agreements are the relations between head daughter and complement daughters. Since ICG provides an explicit accessing capability to the daughters features, both types of agreements can be easily solved. The only difference is that coordination requires a variable ranging over a finite domain of syntactic classes and/or semantic classes to denote the common features of two arguments (8). For instance, the subject verb agreement problems are solved by stating agreement constraints on each subject role in every basic pattern. For example, the AGENT of the verb "persuades" is expressed as AGENT [NP, + singular, + third] in active voice.

(8)

```
meaning : "and"
feature : DUMMY feature
arguments: DUMMY 1: feature: class x
           DUMMY 2: feature: class x
```

4.2 Long-distance Dependency

Kaplan and Zaenen [Kaplan 88, Kaplan 89] proposed functional uncertainty as a new descriptive technique, assuming grammatical function in LFG, in order to account for long-distance dependencies. The schema of functional uncertainty avoid the problem of infinite specifications so that the attribute positions for topicalizable constituents could be realized as a regular expression such as $(\uparrow \mathrm{comp}^{*}\ \mathrm{subj}|\mathrm{obj}) = (\uparrow \mathrm{topic})$. This equation expresses the uncertainty about what the with-in clause functional roles of an extraposed topic might be. It offered a clearer and more accurate characterization of long-distance dependencies but still could not handle the case of context dependency in topicalization. For instance, there are many verbs in Chinese which do not allow the object to be topicalized. When such a verb is the head of an embedded sentence, it is clear that (comp obj) is not toplicalizable either. There is no way to predict the topicalizable constituent in depth by way of regular expressions. There are similar examples in English.

(9)

a. Who$_i$ did Mary think that Bill saw e$_i$?
?Who$_i$ did Mary quip that Bill saw e$_i$?

b. Who$_i$ did John tell you (that) Mary thought that Bill saw e$_i$?
?Who$_i$ did John tell you (that) Mary quipped that Bill saw e$_i$?

However we can solve such context dependent problems in ICG by recursively defining the topicalizable constituents in terms of topicalizable constituents of embedded sentences. The verb with sentential complements usually causes the problems. We may encode the topicalized sentence patterns as a part of BP or as derived by lexical rules. The topic of the embedded sentence is one of the topicalizable constituents of such verbs. For instance the topicalized sentence pattern of example (2) is:

TOPIC [TOPIC (THEME)]< AGENT[NP]< • <
GOAL[NP]< THEME/TOPIC

The topics of embedded sentences are defined recursively under the sentence patterns of the verbs of embedded sentences. Detailed discussion is given in [Chen 90, Huang 90].

4.3 Control and Binding

Functional control is the relation that exists between an antecedent and the missing subject in an XCOMP or XADJUNCT [Bresnan 82, Sells 85]. The coindex label adopted in the unification-based grammars is a simple solution to such problems. We use the same scheme in ICG, e.g. in (2). Anaphoric binding was solved in LFG by the concept of F-command [Bresnan 82, Sells 85]. The same concept is also applicable to ICG.

5. Conclusion

From the brief sketch above, it is clear that each lexical entry contains a large amount of redundant syntactic and semantic information shared by other entries belonging to the same category. Hence, a more elegant and economic strategy is to form a category hierarchy and to store shared information on higher level nodes. Each lexical entry contains only individual idiosyncracies. Thus the redundancy in representation can be removed and data consistency can also be easily maintained.

ICG is a kind of unification-based formalism. Therefore all the advantages of unification-based formalism were kept in ICG. Furthermore, additional advantages are incorporated via the following design features of ICG.

1. Declarative

   The lexical information defines legal sentences or phrase. And the changes in the above information does not affect processing procedures and results.

2. Algorithm-free

   Different control processes are allowed for parsing or generation. Regardless of whether it is sequential, parallel, or heuristic control, the result will be the same.

3. Suitable for parallel processing

   The processes are head driven. Each possible head can initiate a phrasal construction process. Thus, processes can be executed in parallel.

4. Allows a partial parse

   At any moment of the unification processes, the accumulated information shows the partial result up to that moment even if the input is ungrammatical.

5. Semantic structure is universal

   The semantic information contained in the feature structures of each lexical entry is universal. Different languages differ with regard to syntactic of information.

6. Incorporated with thematic role identification scheme

   The parametrical information for identifying thematic roles can be encoded in ICG such that syntactic parsing and semantic analysis can be done in parallel.

7. Parsing result is a thematic structure

   Recent theoretical linguistic studies are concerned with the relationship between thematic structures and argument structures (e.g. Brensan and Kanerva 1989). Our formalism dircetly and explicitly represents thematic structures.

Last but not least, we expect future studies of ICG to confirm the plausibility of the following advantages.

1. Efficient parsing

   ICG has the advantages of lexicon-based parsing which is better focused on the relavant syntactic and semantic information.

2. Cognitive reality

   The language capability of a man seems to be improved day after day by learning and polishing lexical information. ICG is able to reflect this phenomenon.

3. Generation power

ICG provides a way of generating surface sentences from thematic structures. The pragmatic consideration of the selection of the sentence patterns can be accomplished by incorporating the pragmatic features in the basic patterns and lexical rules.

5. Acknowledgement

Research of this paper was partially supported by NSC grant #78-0408-E001-001 and the Electronic Research and Service Organization, Industrial Technology Research Institute, Taiwan, R.O.C. under Contract No. X2-79007. Some of the examples are taken from the research results of the Chinese Knowledge Processing Group at the Computing Center of Academia Sinica. We alone, of course, are responsible for any possible errors contained in this paper.

6. References

Bresnan, J., 1982. The Mental Representation of Grammatical Relations, Cambridge: MIT Press.

Bresnan, J. and J. Kanerva, 1989. Locative Inversion in Chichewa: A Case Study of Factorization in Grammar. Linguistic Inquiry 20, pp. 1-50.

Chen, K. J. and C. S. Cha, 1988. The Design of a Conceptual Structure and Its Relation to the Parsing of Chinese Sentences. ICCPCOL'88, Toronto.

Chen, K.J., C.R. Huang, and L.P. Chang, 1989. The Identification of Thematic Roles in Parsing Mandarin Chinese, Proceedings of ROCLING II (1989), Taipei, Taiwan, pp. 121-146.

Chen, K.J., C.R. Huang, and W.P. Chen, 1990. Resolution of Long-Distance Dependencies with Recursive Information Embedding, In preparation.

Fillmore, C. 1968. The Case for Case. In E. Bach and R. Harms (Eds.), Universals in Linguistic Theory, New York: Holt, Rinehart, and Winston.

Gazdar, G. et al. 1987. Category Structures. CSLI report 102, Stanford: Center for the Study of Language and Information.

Gazdar, G., E. Klein, G.K. Pullum, and I.A. Sag, 1985. Generalized Phrase Structure Grammar. Cambridge: Blackwell, and Cambridge, Mass.: Harvard University Press.

Huang, C.R. 1989. Mandarin Chinese and the Lexical Mapping Theory. The 1989 International Conference on Sino-Tibetan Languages and Linguistics. Hawaii.

Huang, C.R., K.J. Chen, W.P. Chen and T.Y. Hu, 1990. Resolution of Long-Distance Dependencies in Mandarin Chinese - With an Algorithm Based on Functional Uncertainty. To appear in the Proceedings of the 1990 International Conference on Computer Processing of Chinese and Oriental Languages (ICCPCOL'90).

Kaplan, R.M. and J. Maxwell, 1988. An Algorithm for Functional Uncertainty. Proceedings of Coling '88, Budapest, 297-302.

Kaplan, R.M. and A. Zaenen, 1989. Long Distance Dependencies, Constituent Structure, and Functional Uncertainty. In M. Baltin & A. Kroch (Eds.), Alternative Conceptions of Phrase Structure. Chicago: Chicago University Press, 17-42.

Karttunen, L., 1986. Radical Lexicalism, CSLI Report No. CSLI-86-68, Stanford: Center for the Study of Language and Information.

Mellish, C.S., 1988. Implementing Systemic Classification by Unification, Computational Linguistics, Vol. 14 #1, 40-51.

Pollard, C. and I. Sag, 1987. Information-based Syntax and Semantics, Vol. I. Fundamentals, CSLI Lecture notes #13, Stanford: Center for the Study of Language and Information.

Sells, P., 1985. Lectures on Contemporary Syntactic Theories. CSLI Lecture Notes no. 3. Stanford: Center for the Study of Language and Information.

Shieber, S.M., 1986. Introduction to Unification-based Approaches to Grammar. Stanford: Center for the Study of Language and Information.

Tomita, M., 1986. Efficient Parsing for Natural Language, Boston: Kluwer Academic.

Uszkoreit, H., 1986. Categorial Unification Grammars. In Proceedings of Coling 1986. Bonn: University of Bonn. Also appeared as Report No. CSLI-86-66, Stanford: Center for the Study of Language and Information.

Winograd, T., 1983. Language as a Cognitive Processes, Vol. 1, Syntax, Addison-Wesley.