Computational Linguistics and Chinese Language Processing
vol.1, no.1, August 1996, pp. 183-204




# A Model for Robust Chinese Parser

**Keh-Jiann Chen***

**Abstract**

The Chinese language has many special characteristics which are substantially different from western languages, causing conventional methods of language processing to fail on Chinese. For example, Chinese sentences are composed of strings of characters without word boundaries that are marked by spaces. Therefore, word segmentation and unknown word identification techniques must be used in order to identify words in Chinese. In addition, Chinese has very few inflectional or grammatical markers, making purely syntactic approaches to parsing almost impossible. Hence, a unified approach which involves both syntactic and semantic information must be used. Therefore, a lexical feature-based grammar formalism, called Information-based Case Grammar, is adopted for the parsing model proposed here. This grammar formalism stipulates that a lexical entry for a word contains both semantic and syntactic feature structures. By relaxing the constraints on lexical feature structures, even ill-formed input can be accepted, broadening the coverage of the grammar. A model of a priority controlled chart parser is proposed which, in conjunction with a mechanism of dynamic grammar extension, addresses the problems of: (1) syntactic ambiguities, (2) under-specification and limited coverage of grammars, and (3) ill-formed sentences. The model does this without causing inefficient parsing of sentences that do not require relaxation of constraints or dynamic extension of the grammar.

**Keywords: Chinese Parser, Robust Parser, Information-based Case Grammar, Branch-and-Bound Algorithm**

## 1. Introduction

One of the steps in language understanding is producing the syntactic and semantic structure of a sentence. This step is called parsing. Other steps, such as inference, discourse analysis, and responding actions, can thus be performed accordingly. In general, sentence parsing is carried out by two different modules. The first module is a lexical analyzer which identifies the words in a sentence and provides their syntactic

* Institute of Information Science, Academia Sinica,Taipei, Taiwan.
E-mail: kchen@iis.sinica.edu.tw

and semantic information from a lexicon. From the output of the lexical analyzer, the second module produces the syntactic and semantic structure of the sentence. The syntactic structure and thematic role of each constituent are identified according to the grammar of the language. World knowledge is used in resolving structural ambiguities as well as in identifying thematic roles.

It is well known that context-free languages, such as programming languages, can be easily parsed. Except for some very exceptional context-sensitive constructions, natural languages are context-free languages. In addition, if the depth of center-embedded constructions is bounded, natural languages are regular languages. That is to say, there is theoretically an efficient and simple parser for natural languages.

Why then are there no natural language parsers that are practical and useful in general domain? Of course, there are parsers of limited success which can parse subsets of grammatical sentences in special domains. Stated simply, the difficulty is due to the complexity of the constraint relations between words, and not by the level of structural complexity. It is almost impossible to write a grammar that is precise and that fully describes the target natural language. Less constrained grammars may have better coverage, but they also tend to over-generate because of incorrect ambiguities in the grammar. Not only do these ambiguities cause a parser to incorrectly identify a parse as grammatical, they also frequently cause a parser to stop processing before finding a correct parse because its memory resources are exceeded. On the other hand, a more constrained grammar tends to have poor coverage (under-generates) because its syntactic and semantic constraints are too restrictive. Usually, a natural language grammar suffers from both over-generation and under-generation.

In this paper, a lexical feature-based grammar formalism is adopted with which complex word-relations can be expressed, and a robust parsing model is proposed which overcomes the previously mentioned difficulty. In the next section, a prototypical chart parser for Chinese is presented in order to illustrate the basic procedure for parsing Chinese sentences using a chart parser. The characteristics of Chinese that make parsing difficult are discussed in section 3. In section 4, a grammar representation and data management model are proposed. They are a central part of the robust parsing model. In section 5, a robust parsing model based on a priority controlled chart parser is proposed. This parser is proposed to be able to handle ill-formed sentences without sacrificing parsing efficiency on well formed sentences. Summary and future work are discussed in section 6.

## 2. A Chart Parser for Chinese Language

A model for a Chinese language parser is given in Figure 1.

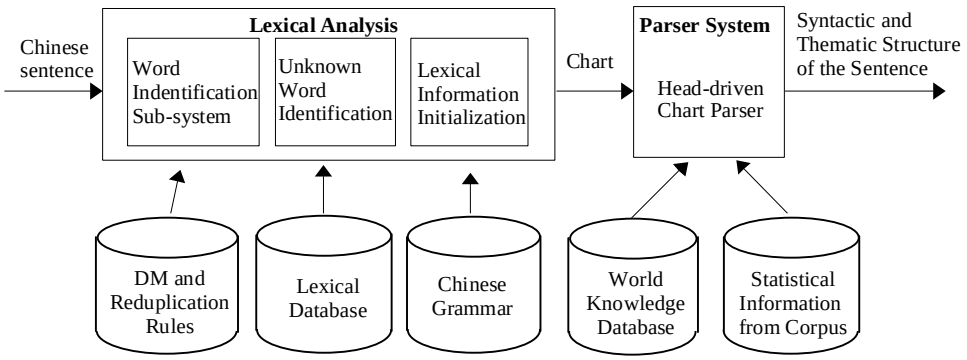
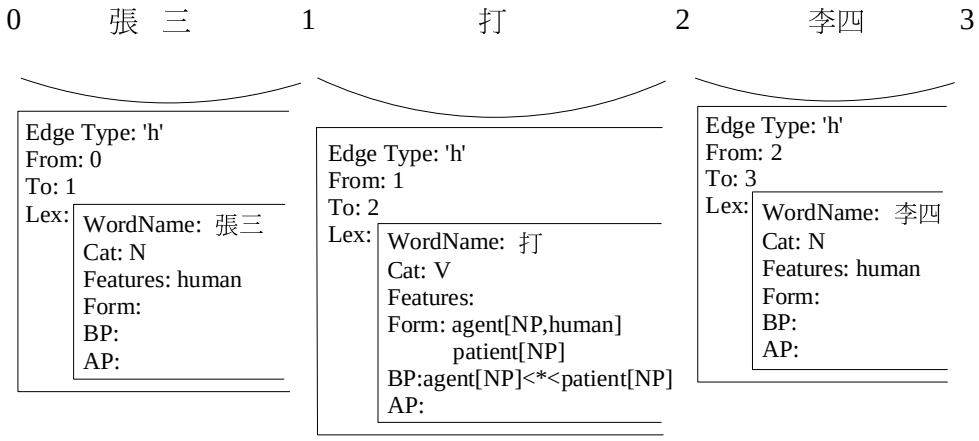



***Figure 1*** *A model of Chinese parser system*

The lexical analyzer takes a string of Chinese characters as its input and produces a feature embedded chart as its output. There are three parts to the lexical analyzer: word identification, unknown word identification, and lexical information initialization. Morphological rules and a lexical database are used to identify words. The database provides information on common words and the morphological rules assist in the identification of compound, determinative-measure, reduplicative, and derived words [Chen 92]. The unknown word indentification process makes guesses on words which cannot be identified using the database and morphological rules. The unknown word identification process is also required to handle proper names since they are not identifiable by regular morphological rules [Sun 96, Chen 96]. An input string may often have more than one possible segmentation. Heuristic methods, such as longest matching or statistically most plausible sequence, may be used to resolve these conflicts [Chen 92, Chiang 92].

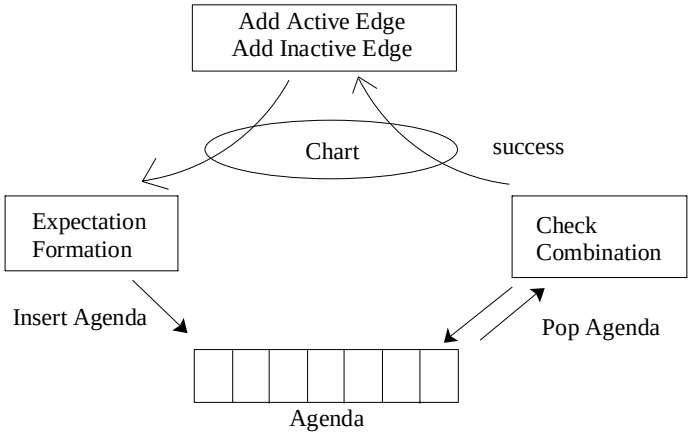
The lexical analyzer then initializes a chart data structure for the input sentence using the categories provided by the word identification process. A feature structure which represents the syntactic and semantic features of a category is associated with each edge in the chart. Diagram (1) shows an example of a simplified feature embedded chart.

(1) A simplified feature embedded chart




The chart parser takes the feature embedded chart and assigns syntactic and semantic structure to the input string. A world knowledge database and a statistical database provide semantic and statistical preference information which the parser uses to the amount of processing required to resolve syntactic ambiguities. The details of these two functions are discussed in a later section. A bottom-up head-driven parsing strategy is used in order to avoid unnecessarily building predicted or redundant structures. The parser uses an agenda-driven control because of the flexibility it provides in ordering the parsing steps. In a head-driven parser, the parsing process begins with the potential phrasal heads of the input and is guided by the phrasal patterns registered in the heads. It examines constituents to the left and right of the head to see if they match the syntactic and semantic restrictions imposed by the possible thematic roles (FORM) and the phrasal patterns (BP or AP) of the head. The actual parsing procedure is depicted in the following diagram (2). (3) shows the top level algorithm.

(2) First-come-first-serve Agenda Control




(3) A Head-driven Parsing Algorithm

---

Step 1: For each edge, do /*initialization*/

case 1: It is a potential phrasal head

/* potential phrasal heads are N, V, P.*/

Create an active edge for this word;

If it is a complete phrase, then create a new inactive edge for this phrase. /* such as N */

case 2: It is not not a phrasal head:

/* such as Adj, Adv, Aspect.*/

Create an inactive edge for this word.

Step 2: For each active edge, pair it with each neighboring constituents to form a series of (head, neighbor) agenda items and add them to the agenda.

Step3: While the agenda is not empty Loop

For the (head, neighbor) agenda item that is first in the agenda Do

1. Check the form constraints of the head to determine which thematic roles can be taken on by the neighbor constituent.
2. If a role also satisfies the constraints posted by BP or AP rules as well as principles of grammatical functions, such as functional completeness and uniqueness conditions and the coherence condition [Bresnan 82, Chen & Huang 90], then create a new active edge for the thematic role and its head.
3. For each newly created active edge, pair it with each neighboring constituent to form a new (head, neighbor) edge pair and add it to the agenda.

End While Loop.

---

The head-driven parsing algorithm will succeed in finding a parse for the input string if the input is covered by the grammar. However, it would fail on processing many sentences that are used in everyday situations. Why? Because grammars used by parsers do not fully and precisely describe the word-relations of the target natural language.

## 3. Difficulties of Parsing Chinese

Because of the complexity and flexibility of word usage in natural languages, there are many sentence structures of a target language which are not covered by any of the existing grammar representation for that language. Another inadequacy of current grammar representations is the over-simplification of grammar rules, causing multiple and spurious parses for a given input. For Chinese, there are additional difficulties at the stage of lexical analysis. Word identification is difficult in written Chinese since there are no blanks to mark word boundaries. In addition, when parsing real-life input, there are often many words that are not in the lexicon or that cannot be indentified by morphological rules.

In order to parse a sentence with unknown words, the lexical analysis system must first be able to detect the existence of unknown words in the input. The identification process would then need to find the boundary of the unknown words and to determine their syntactic categories and semantic types. Since a single Chinese characters can be a mono-syllabic word and many unknown words tend to be multi-syllabic, detecting the existence of an unknown word is not an easy task. An unknown word that is multi-syllabic will often be incorrectly segmented into a sequence of low frequency monosyllabic words. The parser will then simply fail to parse the input rather than detect the existence of unknown words.

Identifying an unknown word, after one has been detected, is an even more difficult task. The first difficulty is in determining the boundaries of the unknown word. Once the boundaries have been determined, a complex process of identifying its syntactic category and semantic type is required. Context dependent rules are adopted for proper names since proper names lack any morpho-syntactic regularity and so they cannot be identified by morphological rules, which can only identify regular compounds. Identifying proper names requires supporting evidence from context (e.g. proper names associated with titles, agreement between conjuncts, reoccurrence, etc. Miss-spelled words, abbreviations, newly coined words, and novel word usage make the problem even

worse.A detailed discussion of Chinese word identification is out of the scope of this paper. Interested readers should refer to [Sun 96, Chen 96].

Even if the lexical analyzer were to function perfectly, the parser may still fail on the following types of sentences:

a. The sentence structure is not covered by the grammar.
b. The sentence is syntactically or semantically ambiguous, and the parser fails to find the most contextually appropriate parse.
c. The sentence structure is very complicated, causing the parser to run out of memory time resources before a result is reached. This is often the case with long sentences.
d. The sentence is ill-formed.

The inherit complexity of natural languages makes it almost impossible to describe a language completely and accurately. The creation of new constructions, new word usages, idiomatic type substitutions, the occurance of unknown words, etc. make it almost impossible to describe the target language completely. The lack of a complete grammar is the reason why the parser fails to process the first type of sentence above. For the second and third types, the parser fails because of syntactic ambiguities. For instance, in Chinese, syntactic ambiguity is present in a grammar of Chinese for the following reasons.

a. Chinese is a weakly marked language with little inflection. Other than a few derived words with derivational suffixes, such as '- 化 ', '- 性 ', and '- 度 ', there are no morphological inflections in Chinese.
b. Words may serve different grammatical functions in different context. Since there are no morphological inflections, it is impossible to establish a one-to-one relation between the functional role of a word and its part-of-speech. For example, in (4), the word ' 攻擊 ' may function as a subject [Chen & Hong 95] as in (4a), and it can also function as an adjective as in (4b).
   (4) a 這次攻擊並不成功
   b 攻擊火力不足
c. A thematic role can be instantiated by many different types of syntactic categories. For example, a time adverbial can be instantiated by a noun, a determiner-measure compound, a postpositional phrase, a prepositional phrase, or an adverb. This is illustrated in (5), respectively.
   (5) 今天 , 三年 , 開會前 , 在秋天 , 常常
d. The linear order of grammatical function roles are relatively free. Chinese is a topic prominent language [Chao 68], and the topic can be either before or after subject, as shown in (6).

(6) a 我看過這本書。
b 這本書我看過。
c 我這本書看過。

e. Flexible and creative word use.

(7) a 張三很聰明。
b 張三很寶貝。
c 張三很豬。

The sentences in (7b) and (7c) show that nouns can be used as stative verbs.

f. An incomplete grammar.
Even after many years of education, a person will not have fully mastered all of the complex constructions and all of the uses of all of the words of their native language. Therefore, natural language grammars that are constructed by humans are usually not complete enough to describe the full range of the complex phenomena of a target target language. Therefore, any real grammar will suffer from under-coverage.

g. An over-simplified grammar.
Grammars are usually written to capture syntactic patterns of a language. Grammaticality constraints are stated in a simplified way to capture generalizations. However, in doing so, it is almost inevitable that any real grammar will suffer from over-generation since generalization often have many exceptions.

From the above discussion, it seems that the degree of success of a natural language parser is very dependent on the grammar used. The parsing algorithm and the grammar must work together in order for the parser to achieve a high degree of success. The characteristics of Chinese mentioned above strongly suggests that for parsing Chinese sentences, both semantic and syntactic information should be used in determining the thematic structure of an input sentence. In fact, in [Chen 89], we found that the five major features that help to determine the thematic role of a constituent are:

a. the syntactic category and the semantic features of the constituent,
b. the predicate argument structures and the semantic restrictions on the arguments of a head,
c. word order,
d. oblique case assigners, including prepositions and postpositions, and
e. world knowledge.

Therefore word-relations are expressed by the syntactic, semantic, and word order constraints. Constraint features are carried by each word, especially words that are phrasal heads. The grammar formalism used in the present research was especially designed for Chinese and is described in the next section.

## 4. Information-based Case Grammar

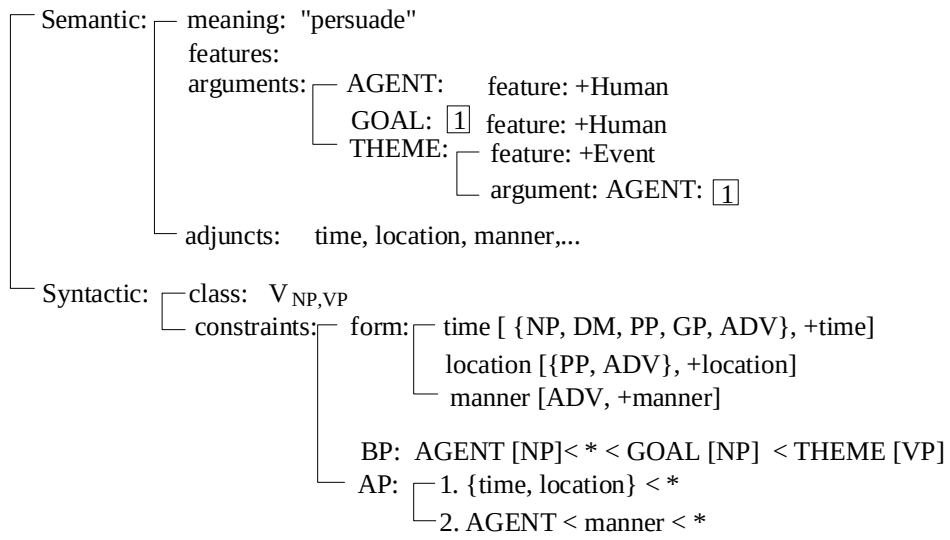
Many contemporary grammar formalisms, such as GPSG [Gazdar 87], LFG [Bresnan 82], HPSG [Pollard 94], CUG [Uszkoreit 86], etc., adopt feature unification as their major processing mechanism. However, the five major features that help in identifying the thematic role of a constituent in Chinese cannot be easily represented in any of these formalism. These grammar formalisms are primarily designed to represent syntactic constraints. For Chinese, it is better to represent grammaticality constraints primarily as thematic constraints. That is, it is better to describe grammaticality constraints in terms of the linear order of thematic roles and their syntactic and semantic restrictions. Therefore, the Information-based Case Grammar (ICG) [Chen & Huang 90] formalism was designed to allow the grammar writer to write constraints primarily in terms of thematic constraints. ICG is designed to provide better coverage and accuracy by adopting a generative lexicon approach [Pustejovsky 95] which stipulates that the lexical entry for each word contains both the semantic and syntactic features of the word and how it may be used. For a phrasal head, the syntactic and semantic constraints on grammatical phrasal patterns are expressed in terms of thematic roles and are encoded as ID/LP (immediate dominance and linear precedence) rules [Gazdar 87]. A uniform feature structure is used for each lexical entry and is shown in (8). The attribute values for an entry are defined using an inheritence hierarchy, with null as a possible value.

(8) Semantic: meaning:
features:
arguments:
adjuncts:
Syntactic: class:
constraints: form:
basic patterns(BP):
adjunct precedence(AP):

Many important features of conventional grammatical formalisms, such as ID/LP and feature based representation of GPSG[Gazdar 87], head driven principle of HPSG[Pollard 94], lexicalism approach of CUG[Uszkoreit 87], are retained in ICG. In addition, the representation is based on thematic roles with additional features to handle the complex structure of Chinese.

The example below illustrates how lexical information and constraints are encoded.

(9) 勸 Quan "persuade":




The entry of a word can be constantly refined without requiring any changes to the design of the parser. These refinements can be done to improve parser performance, without causing any side-effects. How the lexical data is managed is discussed below.

### 4.1 The Goals of Lexical Data Management

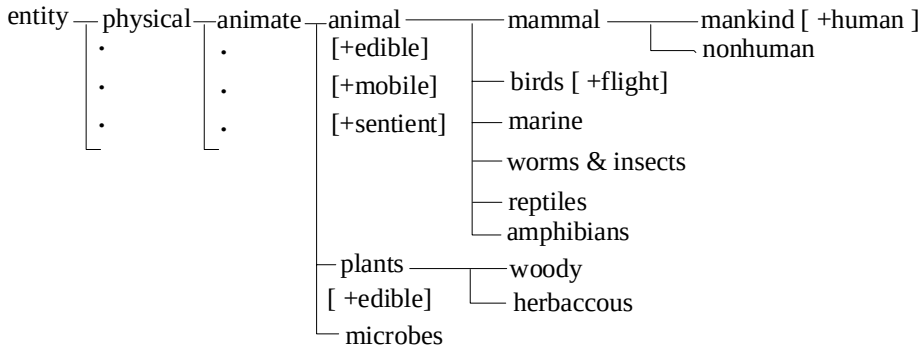
A formal ICG syntax is used to describe the lexical feature structures. Some of the important goals of lexical data management are:

(a) making the data more succinct and less redundant,
(b) keeping the data consistent and error free, and
(c) updating the data with or without global effect.

If all of the grammatical information of a word is encoded individually for each lexical entry, there would be a lot of redundant information in the lexical database since words of the same class and/or category would share much of the same information. If information was individually encoded, how would the data be kept consistent as changes are made? Furthermore, the syntactic phenomena of natural languages are very complicated and cannot be fully and precisely described by any linguistic theory in advance. Subsequently, the syntactic and semantic features and values need to be constantly refined. This requires frequently updating the existing lexical data, and hopefully it can be done without causing any side-effects. These are very important goals which have often been ignored and yet they need to be achieved since they are also crucial to the

success of any research effort. A hierarchical representation (i.e. taxonomy) for both syntactic categories and semantic classes is proposed to achieve these goals. A partial hierarchy of the semantic structure, the Animate Branch, is shown in (10). The taxonomy is organized under the is-a relation which defines a partial order [Chen 88].

(10) The Animate Branch of the Semantic Structure




The is-a relation maintains the inheritance property such that the lower level nodes inherit the properties of their ancestors. Therefore a partial but essential part of real world knowledge is encoded and distributed in the conceptual hierarchy. For instance, a swallow is a kind of bird. The features [+birds], [+animals], [+edible], [+mobile], [+sentient], [+animate], [+physical], and [+entity] are therefore automatically inherited. Similarly, the syntactic classes can also form a syntactic hierarchy such that properties of a syntactic category can be inherited by lower nodes in the hierarchy (i.e. by its subcategories).

The only the idiosyncrasies of a word need to be specified at the individual word level. All other properties are inherited from its higher level classes. For example, the feature structure of 勸 'persuade' in (9) is distributed along the path ' 勸 -Vnpvp-Va-V'.The value of the 'syntactic-form' feature of 勸 is common to all of the verbs and so its value is specified at the 'V' node. The value of the adjunct precedence (AP) feature of 勸 is common to all of the active verbs and so its value is specified at the 'Va' node. The value of the basic patterns (BP) feature is shared by all of the ditransitive verbs which have VP as one of the objects and so its value is stored under the 'Vnpvp' node. The feature structure of 勸 is the unified result of the feature structures belonging to all of the ancestors of 勸 . This type of data management scheme achieves the above three goals for managing lexical data. Keeping data consistent and less redundant is achieved by placing common information at a single node in the hierarchy. Furthermore,

the global or non-global effect of updating the lexical data is controlled for in the following way. If a property is shared by every word in a category then the feature structure of this category can be updated to include this property. This new property will automatically belong to all of the words in this category, achieving the desired global effect. On the other hand, the idiosyncrasies of an individual word can be kept at the leaf node. Updating a leaf node does not cause any undesirable side-effects because it only affects the feature structure of the individual word. As the grammatical information in the lexicon improves, the parser will become more accurate and achieve greater coverage. Furthermore, the use of a taxonomy hierarchy provides a similarity distance between two categories. Robust parsing requires the parser to be able to relax the constraints and similarity distances are used by constraint relaxation techniques. This point will be elaborated on in section 5.

### 4.2 Advantages of the ICG Formalism

Due to its richer and more accessible encoding of thematic information, the ICG formalism is much easier to manipulate than other grammar formalisms. The preparation of lexical feature structures is straightforward. Complex syntactic structures such as coordinations, long-distance dependencies, control and binding, can be easily expressed in ICG [Chen & Huang 90, 91]. The inclusion of semantic constraints along with thematic constraints allows a more accurate grammar to be written for Chinese. Better grammar coverage and accuracy is achieved by gradually improving the lexical feature structures. New words, new word uses, and discovery of new phenomena can be encoded into lexical feature structures at different representational levels of the syntactic and semantic hierarchies without causing over-generation or other side-effects. The declarative property of ICG allows the grammar to be revised at any time without requiring the parsing algorithm to be changed. The ambiguities of syntactic structures can be resolved by semantic preference checking. The semantic preferences between a head and a modifier, or a predicate and its arguments, are represented in ICG by semantic restriction or by statistical association strengths between words or semantic classes [Resnik 93]. The association strengths can be estimated by the statistical value of mutual information computed from a large corpus [Church & Hanks 90]. For instance, the semantic restriction on the agent of the predicate 'persuade' is +human which is denoted by Agent [NP, +human] in ICG. Therefore, by using semantic hierarchy distance measures and association strength values, the following semantic preference relation [teacher > tiger > tree > color] can be derived for the agent of 'persuade'.

The use of thematic roles, syntactic and semantic restrictions, and heirarchical data management allows a grammar to be written with better coverage and less spurious

structural ambiguity. Furthermore a lexically-based representation has better information focusing. If a priority control parsing strategy is also used, the likelihood of the parser running out of memory on long sentences is reduced. The proposed priority control strategy is discussed in the next section. The present research has not yet addressed the problems of ill-formed input or creative word usage. So far it seems that the proposed ICG grammar formalism and chart parser can handle normal sentences. In the next section, a more robust parsing strategy is proposed which can handle abnormal cases by making reasonable guesses and yet still handle normal sentences quickly and efficiently.

## 5. A Robust Parsing Scheme

Natural languages are not static systems, but are constantly changing because the human mind is very creative and intelligent. New words, new word uses, and new constructions are constantly appearing. Any enumerative type of grammar representation is unable to handle changes in a language. For example, in (7), the 很 'very' construction is a very common expression that expresses the characteristic of the subject and has the structural pattern of THEME[NP] < DEGREE [ADV[' 很 ']] < *[Vs] where Vs denotes a stative verb. As was shown in (7b) and (7c), the stative verb can substituted by what appears to be a noun ( 寶貝 'treasure' and 豬 'pig'). With an enumerative representation, both 寶貝 and 豬 would have to also be listed as stative verbs. This approach makes grammar writing tedious and endless because there do not appear to be any limitations on creative word uses [Pustejovsky 95]. In addition, representational ambiguities increase dramatically if no distinction is made between normal uses and creative uses. Creative word uses are strongly associated with particular constructions. For example, 豬 cannot function as a stative verb in most constructions. Therefore, it seems advantageous to separate normal from new and creative word uses. The same is true with new constructions. For example, argument omission, abnormal word ordering, partial structures, noisy word insertion etc. are very common, especially in spoken language. Trying to handle these phenomena as part of the normal grammar only leads to an unnecessarily complex grammar. Therefore, any robust natural language parsing system must have the ability to predict, infer, and extend the grammar and lexicon to handle new words, word uses, and constructions.

A robust natural language parser should not fail to process nor prematurely terminate on any input, even ill-formed input. It may not always succeed in producing the best interpretation, but it should at least make a reasonable prediction or provide partial structures for the input. In order to achieve robust performance, a model is proposed that uses a core grammar to cover the set of normal sentences and a method of grammar

extension to cover abnormal sentences. The core grammar, written in the ICG formalism, enumerates prototypical word usage and phrasal structures in the lexicon. It was argued in Section 4 that the core grammar can be expressed by the lexical feature structures of an ICG grammar. Grammar extension is accomplished by predicting and inferring, based upon the syntactic and semantic context, the required extension for a word. The specific method of dynamic grammar extension under the ICG framework will be discussed next.

### 5.1 The Evaluation Function and the Control of Grammar Extension

One of the most important functions of a robust parser is to evaluate the grammaticality and logical soundness of a phrase structure or a partial phrase structure for a string of words. Any grammar must allow structural ambiguities and so an evaluation scheme is needed to rate the possible structures at every step. A priority control strategy then uses these ratings to guide the parsing process. The degree to which the core grammar is extended depends on the degree to which the parser must construct structures with a worse rating in order to reach a parse for the input.

The evaluation function rates structures using integer values from 0 to N, where a value of 0 denotes a grammatically perfect structure (i.e. a syntactically and semantically proper structure). Therefore, the greater the evaluation score, the lower the preference. The set of phrase structures with preference scores of 0 can be defined as the language generated by the core grammar G0. Similarly, the extended grammar Gi denotes the grammar which generates the set of phrase structures with preference scores less than or equal to i. A robust parsing algorithm, based upon the priority controlled chart parsing algorithm, could first search for a parse of the input with the G0 grammar. If none is found, it could then search for a parse using the G1 grammar. Ill-formed sentences are thus parsed by gradually relaxing the grammaticality constraints from grammar G0 to G1, G2,..., until a parse is found using grammar Gi. The algorithm will always terminate if a grammar G* is defined to accept any input string. Therefore, normal sentences (i.e. sentences covered by the G0 grammar) are still parsed quickly since no grammar relaxation is required. There is only a performance cost when grammar relaxation is required to find a parse of the input.

However, if the parsing process were to start over from an initial state at each extension of the grammar, the parser would unnecessarily rebuild structure previously built. Therefore, a better control algorithm is needed. A priority control algorithm that avoids this redundant searching is discussed in the next section. A possible definition for the evaluation function and the grammar extensions based on the ICG formalism are described in section 5.4.

## 5.2 A Robust Chart Parsing Model

A robust parser requires:

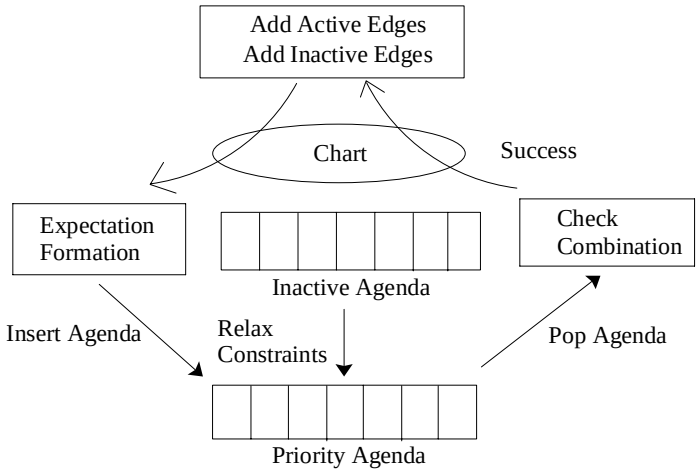
(i) an evaluation function which gives a preference score for each structure,

(ii) a sequence of grammars G0, G1, ..., Gn, G* where L(G0) < L(G1) < ... < L (Gn) < L(G*); G0 denotes the core language grammar; Grammar Gi+1 is an extension of of Gi, and G* is a universal grammar which generates any word sequences; L(Gi) denotes the language generated by the grammar Gi, and

(iii) a parser control mechanism which controls the parsing sequence such that there is no performance penalty when parsing normal sentences and best preference guesses are made on ill-formed input.

The above three requirements are closely related. Discrete values are needed for both structure preference scores and the related grammar extension (i.e. grammar Gi generates all and only the structures with scores less than or equal to i). For convenience, any non-discrete preference score can be normalized into finite integer values. The grammar G* guarantees that the robust parser system will terminate on every input string. The grammar G* may be a purely statistical preference grammar [Collin 96]. A bigram lexical dependence grammar is another possible choice [Collin 96]. If the parser fails to find a parse for an ill-formed sentence using grammar Gn, then the statistical grammar will guarantee that some result is produced for the input. The bottom-up chart parser introduced in Section 2 modified from a First-In-First-Out agenda control to a priority agenda control. The diagram illustrating a priority agenda control is shown in (11).

(11) Priority Agenda Control



The parser maintains a priority agenda and an inactive agenda. The parser always processes the edge pairs in the priority agenda with the lowest evaluation score first (i.e. the highest preference pairs first). The preference score of the edge pair is estimated by the sum of the preference scores of the two edges of the pair. If the edge pair successfully forms a new structure, a new edge is created with its own preference and a new expectation will be created as usual. A newly created edge pair is added to either the inactive agenda or the priority agenda depending the preference score of the edge pair. If the preference score is less than i of the current level of grammar extension Gi, then the new edge pair is added to the priority agenda; otherwise it is added to the inactive agenda. The parsing process relaxes the grammar constraint to Gi+1 when the priority agenda queue becomes empty. It does this by moving all edge pairs in the inactive queue to the priority agenda queue and starts the parsing process again with the current grammar extension set to Gi+1. Under such a control mechanism, normal sentences will be matched by grammar G0 without any unnecessary processing. If an ill-formed sentence does not match grammar Gn, then grammar G* makes the most probable guess on the structure of input word string.

In order to ensure that the above control algorithm works properly, the evaluation function must increase monotonically with the length of the processed structure (i.e. the preference score of any structure will not be better than the score of any of its substructures). This requirement is a reasonable, since a structure that contains a bad substructure should not be better than that substructure. In fact, the evaluation function proposed below defines the evaluation score of a structure to be the accumulated sum of the penalty scores of grammatical constraint violations for each substructure plus any additional penalty incurred in forming that structure.

### 5.3 A Step Toward the Evaluation Function and Grammar Extensions

It is difficult to define a perfect evaluation function. A possible evaluation function is proposed based upon the framework of ICG. The grammar G0 is assumed to be the core grammar of Chinese in terms of the ICG formalism described in section 4. The core grammar G0 describes the set of normal phrasal structures which are syntactically and semantically proper. Grammar extensions are obtained by relaxing the syntactic, semantic, and structural constraints of G0.

(i) The Relaxation of Semantic Constraints

In ICG, predicate-argument and head-modifier relations are expressed in terms of semantic features. For example, the agent of 吃 'eat' (eat) is an NP and has the semantic class of +animal represented as AGENT[NP, +animal] in ICG. Using common sense,

the suitability of teacher, tree, book, and color to be the agent of 吃 is teacher < tree < book < color. The semantic constraint +animal can be relaxed into +animate, +physical, +entity in order to accept 'tree', 'book', and 'color' as agent of 吃, respectively. Distance measures between two conceptual nodes in the semantic hierarchy can be used as the preference measure for the evaluation function. A natural way of measuring semantic distance between two compared nodes in a taxonomy is the shortest path to their common ancestoral-node [Resnik 95]. The order of the preference teacher < tree < book < color agrees with the distance measures between the +animate, +physical, +entity nodes and the +animal node.

(ii) Relaxation of Syntactic Constraints

Similarly, the distance measure between syntactic categories can be defined either in terms of their distance on the syntactic taxonomy hierarchy or by tabulating the distance values of each pair of syntactic categories. A syntactic constraint in ICG is represented hierarchically. The major syntactic constraint for a thematic role is a set of categorial structures. Each categorial structure is a category hierarchy with at most three levels. For instance, the thematic role AGENT [PP[P3[' 被 ']] has the syntactic constraint levels of PP, P3, and ' 被 ' .The first level constraint is the phrasal category. The second level constraint restricts the lexical categories of the head of the phrasal category PP. The third level constraint is the specific word of the phrasal head. The relaxation of syntactic constraints can be done at any level. The relaxation of a higher level constraint means a higher level grammar extension.

(iii) The Relaxation of Structural Constraints

Similar to LFG [Bresnan 82], there are three well-formedness conditions which define the grammaticality of the thematic structures of the phrases in ICG:

a. completeness and functional uniqueness conditions,
b. coherence conditions, and
c. linear precedence and syntactic form constraints.

In fact, the completeness and functional uniqueness condition are enforced with respect to the Basic Patterns (BP). The argument structure of a phrase should match at least one Basic Pattern defined by the phrasal head. Omitting a thematic role in the BP, called role deletion, is considered a violation of the completeness condition. Violation of the coherence condition occurs when there is a thematic role in the phrase that is not licensed by the head. This is called role insertion. Similarly, role insertion is also considered a violation of the functional uniqueness condition. Adjuncts of a phrase are optional and constrained only by the linear precedence rules (AP) and the form con-

straints. Therefore, there are three types of structural relaxations: role deletion, role insertion, and precedence reordering. The relaxation of more well-formedness conditions means a higher level of grammar relaxation.

Conceptually there exists a sequence of grammar extensions which correspond to the combination of different levels of constraint relaxation. In fact, grammar extensions are not actually carried out by creating a new sets of grammar rules. Rather, constraint relaxation of feature structures happens during parsing since the grammar rules relevant to the current parse are specified in the feature structure of each input word. The core grammar is a head-driven grammar (i.e. phrase structure patterns are specified in the feature structure of the phrasal head). The relaxation of semantic and syntactic features does not actually change the constraint features of the input words. Rather, by measuring the similarity distance between the candidate feature and the constraint feature, the preference score can be measured. Thus, the grammar level that the resulting structure belongs to can also be determined. However, if the head word of input is missing or type shifted, there will be insufficient information to perform grammar extension because the phrasal patterns (constraint features) are registered in the BP of head. In fact, in ICG, special constructions are not only registered in the phrasal head; they are also redundantly registered at key words of the constructions. For instance, the construction pattern of example (7), THEME<DEGREE[' 很 ']<*[Vs], is also registered in the BP of the word 很 'very'. There are no verbs in the sentences in (7b) and (7c). At some point during grammar extension, the structural pattern in 很 'very', will be accepted as the basic pattern of the extended grammar and the type of 寶貝 'baby' and 豬 'pig' will be shifted from N to Vs using the type coercion operation [Pustejovsky 95].

The domain of the evaluation function F are all structural hypotheses and each structural hypothesis s is a combination of two structural hypotheses s1 and s2 which correspond to the pair of edges in the priority agenda of the chart parser. The evaluation score F(s) is defined to be F(s1)+F(s2)+ p, where p is the level of relaxation with respect to G0 in order to accept the structure (assuming that the substructures s1 and s2 are permissible at the current level i). Therefore the evaluation function F is monotonically increasing, since F(s) >= F(s1)+F(s2). Whether or not s belongs to L(Gi) can be determined by G0 and the value of F(s) since the value of F(s) can be derived by the step by step composition of its substructures and each step is guided by the G0 grammar. If the constraint relaxation is carried up to a maximal level n, and there are still no solutions found, then grammar G* and statistical parsing takes over. With this type of a robust parsing model, processing normal input sentences will not suffer any performance loss since no structures other than ones allowed by the G0 grammar will be constructed, and it will produce a most preferred structure for any input string.

### 5.4 A Step by Step Example

The following is a step by step example which illustrates how priority control chart parsing and grammar extension work. Suppose that the input sentence is as shown in (7c). After lexical analysis, the feature embedded chart for (7c) is produced, as is shown in a simplified form in (12).

(12)

| 張三 | 很 | 豬 |
|---|---|---|
| Cat:Nb | Cat:D | Cat:Na |
| Sem:+human | Form:THEME[NP] | Sem:+animal |
| | DEGREE[D[' 很 ']] | Adjunct:PROPERTY,QUANTITY,... |
| | *[Vs] | Form:PROPERTY[N,Vs,...] |
| | BP:THEME<DEGREE<* | QUANTITY[DM] |
| | | AP:QUANTITY<PROPERTY<* |

The word string cannot be accepted as a complete G0 structure since there are no verbs in the word sequence. The potential structural hypothesis is NP, since 豬 'pig' and 張三 'Zhangsan' are potential heads of NP. Therefore, step 2 of parsing algorithm (3) creates the edge pairs ( 很 , 豬 ) and ( 張三 , 很 ) and adds them to the priority agenda. At step 3, both edge pairs cannot form a new structure since the edge 很 'very' cannot meet any of the syntactic constraints for the adjuncts of the NP. Therefore, both edge pairs are moved to the inactive agenda. The priority agenda is now empty and no complete structure has been formed for the input word string (i.e. the core grammar G0 does not accept (12)). The grammar extension process now takes effect and the edge pairs with the best preference score in the inactive agenda are moved to the priority agenda. The preference score of these edge pairs becomes the current level of the grammar extension. Therefore, the grammar will be extended to fulfill the 很 'very' construction since the categorial type of the word 豬 'pig' will be shifted from Na to Vs. On the other hand, the robust parsing algorithm might also consider 很 as a property role of the noun 豬 by relax the syntactic constraint of the thematic role PROPERTY to D and create a new NP edge 很豬 'very pig'. These two constructions, the 很 construction and the NP construction, are now in competition with each other. Hopefully, the preference function will correctly rate the 很 construction with a better preference score, causing the following result to be reached.

(13)

| 張三 | 很 | 豬 |
|---|---|---|
| THEME:NP | DEGREE:D | HEAD:Vs |

## 6. Summary and Future Work

Parsing is much like searching a tree. The priority-controlled robust parsing algorithm is

similar to branch-and-bound searching algorithm [Horowitz 78]. If the penalty value is incremental, then it is always finds the best solution.

The best solution is judged in terms of the value determined by the preference evaluation function. Therefore how good the best solution is depends heavily on the choice of the preference evaluation function. In section 5.3, a way of defining the evaluation function was proposed without giving a detailed definition. The evaluation score of a structure reflects the level of constraint relaxation or grammar extension required for accepting the structure. Therefore, one future task is to define in detail an evaluation function that is the best measure for semantic, syntactic, and structural distance. Resnik in [Resnik 95] proposed a similarity distance measure between two nodes in a conceptual taxonomy hierarchy as a possible way to define the relaxation levels on semantic and syntactic constraints. Conceptually there is a hierarchy of extended grammars G0,G1,...,Gn. In fact the grammar rules for each extension do not need to be actually generated. It can be simulated by measuring the distance between structural hypotheses and the grammar G0. The lexically-based representation of the ICG grammar formalism encodes grammatical information in the lexical feature structure of words. The redundant encoding of special constructions on both key words and heads avoids the loss of important grammar rules caused by type shifting or missing heads. The hierarchical semantic and syntactic feature representation of ICG makes the grammar constraint relaxation easy and systematic. Under such a representation each level of grammar extension can be dynamically emulated from the core grammar G0.

## References

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

Chao, Y.R., A Grammar of Spoken Chinese. Berkeley, CA: University of California Press, 1968.

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

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

Chen, K.J. and C.R Huang, "Information-based Case Grammar," *COLING'90*, Vol.2, 1990, pp.54-59.

Chen, K.J. and C.R Huang, "Resolution of Mandarin Chinese Unbounded Dependencies Based on Local Constraints," *Proceedings of ICCPCOL'91*, Taipei, 1991.

Chen, K.J. and S-H Liu, "Word Identification for Mandarin Chinese Sentences "*Proceedings of COLING'92*, 1992, pp.101-107.

Chen, K.J. and M.W. Hong, "Verb-Object Phrase and modifier-Head Compounds in Mandarin VN Constructions", *Proceedings of ROCLING VIII*, 1995.

Chen, K.J., L.P. Chang and Evan Bai, "Unknown Word Identification for Chinese Sentences", in preparation, 1996.

Chiang, T-H, "The Study of Different Statistical Word Segmentation Models," *Proceedings of ROCLING'92*, 1992.

Chien, L.F., K.J. Chen and L.S. Lee, "An Augmented Chart Data Structure with Efficient Word Lattice Parsing Scheme in Speech Recognition Applications". *Speech Communication. 10*, 1991, pp.129-144.

Church, K and P. Hank, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, Vol. 16,#1, 1990, pp.22-29.

Collion, M.J., "A New Statistical Parser Based on Bigram Lexical Dependencies," cmp-lg/9605012, 1996.

Gazdar, G., A. Franz, K. Osborne and R. Evans, "Natural Language Processing in the 1980s." CSLI, Stanford University, 1987.

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

Horowitz, E. and S. Sahni, Fundamentals of Computer Algorithm, chap.8, Computer Science Press, 1978.

Jiang., "Chinese Parsing:An Initial Exploration at LRC." *Computer Processing of Chinese and Oriental Language* 1985, 2(2):127-138.

Lee, L-S, *et al.*, "An Efficient Natural Language Processing System Specially Designed for the Chinese Language," *Computational Linguistics*, Vol.17, \#4, 1990, pp.347-374.

Li, C.N. and S.A Thompson, Mandarin Chinese. University of California Press, 1981.

Pollard, C. and I. Sag, Head-Driven Phrase Structure Grammar, Stanford: Center for the Study of Language and Information, Chicago Press, 1994.

Pustejovsky, J., The Generative Lexicon, MIT Press. Resnik,P.S.,(1993). "Selection and Information: A Class-based Approach to Lexical Relationships," Ph.D. dissertation, Department of Computer and Information Science, Univ. of Pennsylvania, 1995.

Resnik, P.S., "Using Information Content to Evaluate semantic Similarity in a Taxonomy," *Proceedings of IJCAI-95*, 1995.

Sproat, R. and C. Shih, "A Statistical Method for Finding Word Boundaries in Chinese Text," Computer Processing of Chinese and Oriental Languages, Vol.4, No.4, March 1990.

Sun, M. S., " Word Segmentation and Unknown Word Identifications", Lecture note of ICCC 96 Conference, Singapore, 1996.

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

Yang, Y., "Semantic Analysis in Chinese Sentence Analysis." In Proceedings of International Joint Conference on Artificial Intelligence (AAAI). Milano, Italy, 1987.