# VERB SEMANTICS AND LEXICAL SELECTION

ZHIBIAO WU

NATIONAL UNIVERSITY OF SINGAPORE

1994

# VERB SEMANTICS AND LEXICAL SELECTION

BY

ZHIBIAO WU

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

NATIONAL UNIVERSITY OF SINGAPORE

1994

# ACKNOWLEDGEMENTS

I wish to express my deep gratitude to my supervisor Professor Hsu Loke Soo, for his stimulating insights, constant encouragement, and for all the guidance he has given me throughout the entire process of my graduate studies.

Many thanks also go to Dr. Martha Palmer and Dr. Tan Chew Lim for being an inspiration throughout my years of graduate education. Dr. Martha Palmer has been a most supportive mentor. I am grateful to her for suggesting me the right research direction, providing me the chance to be a visiting student in University of Pennsylvania and reading my drafts and providing helpful comments. Her caring attitude and constant encouragement have been a great help through this difficult process. I am grateful to Dr. Tan Chew Lim for his constant help and valuable insights. I also grateful to other staffs and research scholars in the computer science department for their help and support.

I am especially grateful to my family for their loving and unfailing support that have provided much strength and motivation. Finally, I must record my gratitude to my wife Zhu Qin for her sacrificial love, her tender but firm support. Without her sacrificial love, this dissertation would not have been possible.

# SUMMARY

A Predicate with Selection Restrictions (PSR) is the main verb representation scheme in today's rule-based MT systems. They are the consequences of linguistic theories such as Fillmore's case theory, and Jackendoff's lexical conceptual structure. As Levin has addressed (Levin 1985), the decomposition of verbs is proposed for the purposes of accounting for systematic semantic-syntactic correspondences. This results in a series of problems for MT systems: inflexible verb sense definitions; difficulty in handling metaphor and new usages; imprecise lexical selection and insufficient system coverage. It seems one approach is to apply probability methods and statistical models for some of these problems. However, the question reminds: has PSR exhausted the potential of the knowledge-based approach? If not, are there any alternatives that can improve the handling of these problems?

We suggest an alternative that represents verb semantic knowledge and accounts for not only fine-tuned systematic semantic-syntactic correspondences, but also semantic-interpretation correspondences. A verb is not represented by a predicate or simple primitives, but by a set of semantic components that are sensitive to the syntactic alternations and the semantic interpretations. These semantic components are rooted in lexical conceptual lattices, so that the conceptual similarities among different verb meanings can be easily measured. In this way, the system will be able to handle verb senses outside the dictionary coverage and guess the best meaning of the verb.

We take lexical selection problem as the testbed of above verb representation methods. For PSR scheme, we choose a practical English-Chinese translation

system TranStar to see how PSR affects the lexical selection. Experiments show that the accuracy and coverage of the lexical selection are hard to be improved. Because PSR scheme does not care much about the relation among meanings and the way to guess the meaning with world knowledge, the only way to have an improvement under PSR is to exhaustively list every possible translation pairs. We then built a prototype system UNICON to demonstrate our suggested method. English and Chinese verb senses are defined on conceptual lattices. Experiments show that the accurate rate for lexical selection has been improved 13.8% after the system makes use of the extended selection process. This proves that once the verb senses are defined with the consideration about the relation among meanings and ways of guessing the new meaning, the system performance can be improved.

# CONTENTS

# LIST OF FIGURES

## LIST OF TABLES

# Chapter 1

# Introduction

Verb senses and verb interpretations in particular contexts are different concepts. When we talk about verb senses, it is easy for us to think about the verb entries in the dictionary. These entries divide a verb into several different senses and tell us the meaning of each sense. For example, the following is an entry that defines one sense of ‘break’ in the Longman dictionary.

> To (cause to) separate into parts suddenly or violently, but not by cutting or tearing: *to break a window/a leg. The rope broke when they were climbing. The window broke into pieces.*

This definition tells us that the meaning of ‘break’ in these examples is ‘to separate into parts’. By generalization, we can easily conclude that the ‘break’ in the sentence ‘The branch broke’ has the same meaning as ‘to separate into parts’.

The definition describes a particular set of events in which a concrete object is separated into parts suddenly or violently. However, in these events of ‘separation’, many things other than the manner of ‘whether the event happens suddenly or violently’ can be perceived by human. For example, the shapes of the concrete objects, the actions that make the separation happen, the result states of the objects, and so on. The sense of ‘break’ does not tell us about them. When ‘break’ is in a particular context, such as ‘The branch broke’, a human can easily perceive that the branch is separated into two or more shorter branches rather than separated into two thin branches with the same length as the original one. This is the same for the sentence ‘The rope broke when they were climbing’. In a human’s perception or interpretation process, more characteristics of the event are figured out by reference to world knowledge. So we can say the interpretation process assigns a new value to the verb ‘break’. This new perceived meaning can be called the interpretation for the verb 'break' in the particular context.

Verb senses and their interpretations are different. An interpretation is a combination of the verb sense, context information and world knowledge. A correct interpretation is important for computer natural language processing. This can be demonstrated in the following examples:

- Metaphor understanding needs correct interpretation. For example, in the metaphor processing, The 'drink' in the sentence 'A car drinks gasoline' means 'the car consumes gasoline for its energy', rather than the basic meaning of 'drink' in 'a man drinks water'.

- Lexical selection in machine translation needs correct interpretation. Because of the linguistic incompatibilities, different verbs in different languages highlight different characteristics of an event. For example, the English sentence 'The branch broke' is translated into Chinese:

  | | | |
  |---|---|---|
  | 树枝 | 断 | 了. |
  | shuzhi | duan | le. |
  | branch | separate-in-line-segment | Asp. |

  Here, 'duan' has the meaning that specifies that 'the branch is separated into the line-segment shape'. This meaning differs slightly from the sense of 'break' in the sentence 'The branch broke'. So only when a correct interpretation is made, can the system choose 'duan' as the correct target translation.

There is a gap between the verb sense representation and the verb interpretation in a particular context. To get a correct interpretation, an obvious solution is to encode each verb interpretation as one sense. In this way, the system can omit the process that derives the verb interpretation from the verb sense. For the above 'break' sense, we can further divide the 'to separate into parts ...' into a set of fine-tuned senses, such as 'to separate into two parts', 'to separate into the line-segment shape', 'to separate into very small pieces', etc. However, this kind of interpretation encoding faces two unsolvable problems. One is that there is an unlimited number of verb occurrences, and therefore there should be an unlimited number of interpretations. It is impossible to list all of them. The other problem is that the linguistic incompatibilities are pervasive. It is difficult to forecast these linguistic incompatibilities and pre-define them in the verb lexicon. Therefore, we still have to separate the verb sense and the verb interpretation, and provide ways for taking care of the correspondence between verb senses and the interpretation.

Some researchers have addressed a similar problem. Kilgarriff (Kilgarriff, 1992) put it in the following way: Each occurrence of the word in a particular context is called a usage of that word. Each usage has a particular meaning interpretation in that particular discourse. A usage type is formed by grouping those usages that have similar interpretations together. A word sense then can be viewed as such a fuzzy usage type. This same problem has also been addressed in other literature. Ruhl (Ruhl, 1989) addressed the interpretation as 'perceived meaning in particular context'. Cruse presented a classification of different types of 'interpretations' in (Cruse, 1993). All these works demonstrate the importance of the word interpretation problem.

However, none of them has addressed the problem of the verb semantic representation and its key influence on the verb interpretation. A correct verb interpretation depends on a good verb sense representation. It is obvious that we cannot list each interpretation in a machine dictionary. Only a limited number of usage types can be encoded in the dictionary. Then the question is: How can the system make correct interpretations for an unlimited number of usages based on the encoding of a limited number of usage types? After a careful investigation, we think we must get ride of the Predicates and Selection Restrictions (PSR), and find out a new representation method which can be adpative to various interpretations.

In today's rule-based MT systems, predicates with selection restrictions are the main method for verb sense representation. It is the consequences of linguistic theories such as Fillmore's case theory, and Jackendoff's lexical conceptual structures. As Levin has addressed (Levin 1985), the decomposition of verbs is proposed for the purposes of accounting for systematic semantic-syntactic correspondences. This results in a series of problems for MT systems: inflexible verb sense definitions; difficulty in handling metaphor and new usages; imprecise lexical selection and insufficient system coverage. Those probability methods and statistical models are applied for some of these problems. However, the question reminds: has PSR exhausted the potential of the knowledge-based approach? If not, are there any alternatives that can improve the handling of these problems?

We suggest an alternative that represents verb semantic knowledge and accounts for not only fine-tuned systematic semantic-syntactic correspondences, but also semantic-interpretation correspondences. Comparing to the old static verb semantic representation, this representation is dynamic, the identified concepts can be compared and shifted to its super-concepts, sub-concepts or even metaphor concepts. The constraints for sense disambigurration are relaxable and measurable. We propose that: A verb is not represented by a predicate or simple primitives, but by a set of semantic components that are sensitive to the syntactic alternations and the semantic interpretations. These semantic components are rooted in lexical conceptual lattices, so that the conceptual similarities among different verb meanings can be easily measured. In this way, the system will be able to handle verb senses outside the dictionary coverage and guess the best meaning of the verb.

Our preliminary experiments on the lexical selection problem of a practical machine translation system have demonstrated that the PSR representation lacks the strength in handling the verb senses outside of the dictionary listing. If a verb sense is out of the dictionary listing, there is no way for the system to guess the meaning and do coercion based on the world knowledge. In order to demonstrate the power of our representation method, a prototype system UNICON is built. We have done several experiments. One of the experiment is to follow the PSR scheme, the system don't guess the verb sense when the verb sense cannot be found in the dictionary listing. The other experiment is to guess the verb sense for unknown verb senses and attempt to give the best lexical selection. Testing results show that the correct rate of the lexical selection has been improved 13.8% in the second experiment.

## 1.1 Verb entries in MT systems

In this thesis, for the sake of discussion, I will define an abstract representation scheme – the PSR scheme. Most other representation methods can be viewed as an instance or a variant of the PSR scheme. The PSR scheme is defined as follows:

> The PSR scheme is used to denote those verb representation schemes that use only one symbol, i.e., predicate, to denote the event described by the verb, and a set of conditions imposed on another set of symbols. This set of symbols denotes the entities taking part in the event. The set of conditions is named as selection restrictions. We call this scheme the Predicate with Selection Restriction scheme. Therefore the PSR scheme.

As we will see, most of the current NLP systems essentially use the same scheme, although the representations are slightly different. The PSR scheme is a static verb semantic representation scheme, because its identified concepts are fixed and the constraints are fixed too.

### 1.1.1 Typical representations in a bilingual dictionary

In an English to Chinese machine translation system TranStar, the English verb 'break' is defined as:

| Chinese | Meaning | Selection restrictions |
|---|---|---|
| 打碎 | to break into pieces | Object is a brittle object |
| 决裂 | to break (the relation) | Object is a kind of connection |
| 打断 | to break the continuity | Object is a continuous event |
| ... | ... | ... |

The selection restrictions classify the events denoted by 'break' into several rigidly divided subcategorized events. The Chinese translations can be viewed as the predicate in this verb semantic representation.

### 1.1.2 Typical representations in interlingua dictionaries

In a KBMT (Nirenburg:1992) system, a verb 'eat' is represented as follows:

SEM:

```
(%ingest
   (AGENT (value ^$var1)
          (sem *animal))
   (THEME (value ^$var2)
          (sem *ingestible)
          (relaxable-to *physical-object))))
```

In this representation, a paraphrase %ingest is given as the predicate name, and two selectional restrictions are imposed on the thematic roles. Even if it is an interlingua MT system, the verb semantic representation is still a PSR style representation.

### 1.1.3 Verb sense primitive representation

The verb representation in UNITRAN is not a PSR representation. Based on Jackendoff's Lexical Conceptual Structure theory, Bonnie Dorr presented a deeper representation in UNITRAN (Dorr 1990). The decomposition of verbs into a set of primitives provides a way to relate different verb senses together. However, representing verb senses only with a small number of primitives seriously hampers the power of the representation. For example, the verb 'jog' is defined as follows:

```
(DEF-ROOT-WORDS (GO-LOC Y (FROM-LOC
                 (AT-LOC Y Z1))
                (TO-LOC (AT-LOC Y Z2)))
    :ROOTS ((JOG (Y (* Y))
                 (Z1 :OPTIONAL ((* FROM-LOC)
                (AT-LOC (Y) (Z1))))
                 (Z2 (UC (CASE ACC)) ((* TO-LOC)
                 (AT-LOC (Y) (Z2))))
                 (MODIFIER JOGGINGLY))
```

The verb 'jog' is decomposed into several primitives such as GO-LOC, FROM-LOC, AT-LOC, TO-LOC and a MODIFIER 'JOGGINGLY'. This representation scheme certainly captures some parts of the meaning of the verb 'jog'. However, the representation attempts to cover a large part of the meaning in the MODIFIER as 'JOGGINGLY'. When other similar verbs 'run', 'walk' and 'sneak' are defined, the modifiers will have to be 'runningly', 'walkingly', 'sneakingly', etc. It is difficult to fine-tune the differences among these similar verbs. On the other hand, the representation depends highly on the syntactic structure. It falls short of a powerful representation by merely taking care of the syntactic-semantic correspondence.

## 1.2 Problems that are difficult to handle by PSR

With the shallow PSR representation of verb semantic knowledge, a system using a PSR scheme can not handle many problems efficiently. The problems are listed below:

### 1.2.1 Linguistic incompatibilities

Between each pair of different languages, linguistic incompatibilities are a pervasive phenomena. Verbs in many languages tend to be sensitive to fine-grained conceptual distinctions that other languages do not recognize. For example, in the following sentence:

End Gene Raesz, who broke a hand in the Owl's game with LSU, was back working out with Rice Monday, and John Nichols, sophomore guard, moved back into action after a week's idleness with an ankle injury.

The English phrase 'broke a hand' has the meaning: ' To make the bone separated'. The Chinese translation is 弄断 longduan. The translation meaning is 'To do something and make the hand bone separated in the line-segment shape'. The meaning of the Chinese translation is slightly different from the meaning in the English sentence. This is because different languages have different categorizations of lexicalized concepts. With a shallow verb meaning representation, the linguistic incompatibility problem is difficult to resolve.

### 1.2.2 Lexical selection in MT

Experiments to be discussed in Chapter 3 have shown that when human subjects translate the same material, the scope of their lexical choices is larger than a computer translation system. For one particular verb 'break', the number of Chinese translations chosen by the human subjects is many more than the Chinese verbs listed in an English-Chinese translation system. This is because Chinese has serial verb compounds. English 'change-of-state' verbs must be translated into Chinese serial verb compounds. The translation depends a lot on the context information and the world knowledge. In the selection process, to choose the best target verb, the similarities among verb senses need to be measured. The PSR representation scheme does not provide such a similarity measure.

The experiments also demonstrate the following: The human subject grasps more knowledge about the verb's sensitivities to the particular characteristics of an event and the particular features of the entities involved in the events. Therefore, he is able to make a more accurate lexical choice in a wider range. This is exactly what a PSR shallow representation lacks.

### 1.2.3 Fine-tuned syntactic-semantic correspondence

Recent research (Levin, 1992) (Dowty, 1991) also shows that the verb syntactic behavior is sensitive to a more fine-tuned set of semantic components than the thematic role distinctions. As Levin has observed, English verbs that have the same set of syntactic alternations tend to possess the same semantic components. This can be shown in the following:

Four verb classes *break, cut, hit, touch* display correspondence between syntax and semantics.

| Properties | touch | hit | cut | break |
|---|---|---|---|---|
| Conative: | No | Yes | Yes | No |
| Motion and Contact: | No | Yes | Yes | No |
| Body-part Possessor Ascension: | Yes | Yes | Yes | No |
| Contact: | Yes | Yes | Yes | No |
| Middle: | No | No | Yes | Yes |
| Causing a change of state: | No | No | Yes | Yes |

This table shows that the conative alternation is sensitive to 'motion' and 'contact'. These four verbs' behavior with respect to conative alternation corresponds to their possession of 'motion' and 'contact' concepts. The same applies to the correspondence between Body-part Possessor Ascension alternation and 'contact' and the correspondence between middle alternation and 'cause a change of state' meaning (See Section 5.3 for more details). The meaning components such as 'motion', 'contact', and 'change of state' cannot be represented in a PSR scheme. Therefore, the fine-tuned sensitivities between syntax and semantics cannot be addressed in PSR.

### 1.2.4 The correspondence between semantics and interpretations

The other important problem is the correspondence between semantics and interpretations. Because PSR is a shallow representation, it provides no way to predict new usages outside the system coverage. The boundaries among different senses are set by the selectional restrictions imposed on the verb arguments. Therefore, the classification of verb senses is a rigid division that it is difficult to relax. For example, suppose that in the system, the 'break' sense as in 'the watch is broken' is encoded. The selection restriction on the patient is that the patient should be a mechanical device. The meaning of 'break' is 'to make the mechanical device lose its mechanical function'. When the system encounters the following sentence:

> No believer in the traditional devotion of royal servitors, the plump Pulley broke the language barrier and lured her to Cairo where she waited for nine months, vainly hoping to see Farouk.

The English phrase ' broke the language barrier' has the meaning 'To make the barrier unfunctional'. For a human being, it is easy to relax the constraint from 'mechanical device' to 'functional object', and therefore, the meaning can be relaxed to 'to make the functional object lost its function'. For a PSR representation, because the relation between mechanical device and functional object, and the relation between 'losing mechanical function' and 'losing function' are not encoded in the system. The PSR provides no way to relax this verb constraint.

Let us see another example:

> Vermont's sugar maples are scarlet from Sept. 25 to Oct. 15 , and often hit a height in early October.

If the core meaning of 'hit' has the semantic components involving 'motion', 'contact' and 'force', in the above sentence, the 'motion' and 'force' components disappear, leaving the 'contact' as the current interpretation. Chinese translation is 达到 'Da Dao' (reach). The extended meaning has a certain relation with the original meaning. We need to identify ways in which verb meanings extend. But a PSR scheme provides no way to extend the meaning unless the predicates can be related to each other.

## 1.3 Projection rather than decomposition

We have different ways of viewing the relation between verb sense and verb interpretation. A conventional view looking at this relation is that we know the verb sense first, then based on the knowledge of verb sense, we build our verb interpretation. But how can we know a verb sense? One may say that it is by learning many usages or by looking them up in a dictionary. In fact, the dictionary senses are the human generalizations from many usage interpretations. There goes a 'chicken and egg' problem. Which one comes first? It is hard to tell. All we know for sure is that different points of view influence our ways of handling the verb sense representation and its relation to the interpretation in the computer system. For the 'verb sense first' point of view, when people have a certain representation for the verb sense in the system, the interpretation is naturally constructed with the building blocks in the verb sense representation. People seldom provide processes that take care of the various possible perceived variations meaning of the verb in a particular context. Therefore, the interpretation is restricted to the domain that the composition function can map to. On the other hand, from the 'interpretation first' point of view, people define the verb sense on the usage type, i.e., generalizations from a set of similar usages. In the verb sense representation, different interpretations for the same verb sense are allowed. So from this point of view, the verb sense representation is defined for the convenience of the interpretation. The system must incorporate the interpretation process with the discourse processing.

Most of the current representation methods such as PSR and decomposition to primitives belong to the point of view of 'verb sense first'. In such systems, the interpretation of the sentence is simply a composition of pieces of the sense representation. There is no other interpretation process involved. In this way, the problem of constraint relaxation, meaning extensions, and metaphor are hard to handle.

For the PSR scheme, the shortcomings of the representation are:

- The representation makes a rigid division of the different verb senses.
- PSR representation is mostly designed for the "systematic semantic-syntactic correspondences" (Levin: 1985), not for the Semantic-Interpretation Correspondence.
- The predicate representation makes it hard to compare the conceptual similarities among different events.

Decomposing to primitives also has its own problems. The purpose of decomposition is to decompose the verb into primitive predicates with a set of arguments. There is no doubt that the primitives certainly capture some parts of the verb meaning. However, the particular characteristic of the event that a verb denotes is difficult to capture by primitives. The 'jog' series in section 1.1.3 is a good example. Those primitives might be thought of as certain concepts, but the composition of them will not be able to cover all possible interpretations.

Therefore, according to the second point of view as 'interpretation first', we propose a two-level structure. The bottom level is the interpretation level. The upper level is the verb semantic form level. For a particular verb, all the possible interpretations are supposed to be noted in the interpretation level. These interpretations will be categorized into different groups according to their similarities in meaning. The verb semantic form is then a meaning representation that generalizes on the usage groups. The interpretation of a particular occurrence is a projection from the verb semantic form level to the interpretation level. The context information and the world knowledge will decide what interpretation the verb semantic form should be projected onto. This idea can be shown in the figure below:

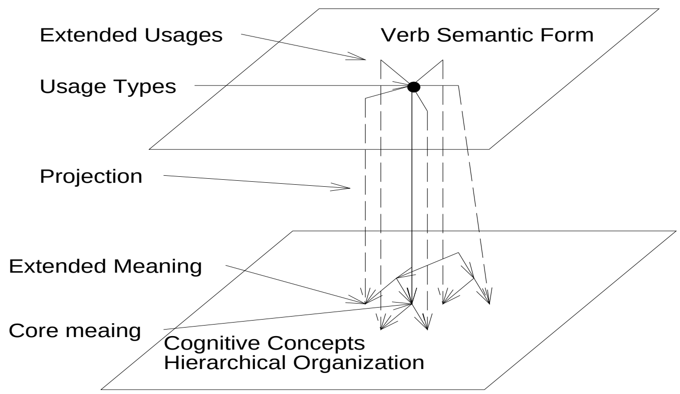



Figure 1-1: The two level structure

Actually the context information and world knowledge act as a coercive force to make sure that the system can get the right interpretation. To take care of the interpretation under this scheme, the relations among different concepts must be identified. We have found out that for some verbs describing a complex event, there are several different cognitive concepts involved. A verb semantic form can be projected onto several cognitive domains. Each projected domain contains a concept that is identified by that verb. In this way, the verb meaning is like a DOT in a multi-dimensional space. Each of the cognitive domains forms one of the dimensions. The concepts identified in each domain are just like the coordinates on those dimensions.

## 1.4 A tour of the thesis

The thesis is separated into four parts. The first part discusses the lexical selection and verb semantic representation problem. The second part is the literature review. The third part is to identify the problem and propose a theory. The fourth part is the application of the theory. A prototype system UNICON is presented.

### 1.4.1 Lexical selection in MT (Chapter 2, 3)

The lexical selection problem is the focus of our investigation of the problem of verb semantic representation. In Chapter 2, experiments show that there could be many different Chinese translations for the same English verb. The properties of Chinese serial verb compounds are discussed. Experiment results are presented to show that human subjects tend to have a large number of candidates for lexical selection and that within a certain degree, imprecise translation can be tolerated. Chapter 3 presents a review about approaches lexical selection problems in the history of MT systems, and attempts to show that the scope and accuracy of lexical selection is heavily influenced by the representation schemes used in these systems.

### 1.4.2 Literature review (Chapter 4, 5)

In Chapter 4, the organization of lexical semantic knowledge is discussed. Most of the hierarchical organization of lexical knowledge is used for syntactic information processing, similarly as Kilgarriff and Flickinger's work (Kilgarriff, 1992), (Flickinger and Nerbonne, 1992).

Verb semantic representation has a long history. Chapter 5 chronicles the historically important theories in this issue. From Fillmore, Jackendoff, Dowty and Levin, we attempt to draw a line of semantic theory development within the generative framework. What we attempt to point out is that the semantic representation is not enough if it only cares about the 'systemic syntactic semantic correspondence'.

### 1.4.3 Verb semantic representation (Chapter 6, 7)

In these two chapters, by pointing out the importance of the semantic interpretation correspondence and the fine-tuned syntactic semantic correspondence, a new theory about verb semantic representation is proposed. Those semantic components of the verb meaning that are sensitive to the syntax and the interpretation must be identified. And all these semantic components can be organized in a conceptual hierarchy so that the conceptual similarities can be measured. In this way, metaphor usage and new usages outside the system coverage can be easily handled.

### 1.4.4 The UNICON system (Chapter 8)

Chapter 8 discusses the prototype system UNICON. UNICON is a system that demonstrates a lexical selection package using the multi-dimensional semantic representation scheme. Each verb in English and Chinese is defined by several concepts in different domains. All of these concepts are organized in the hierarchical structure. Therefore, the similarities among different verbs and concepts can be measured. Four hundred sentences in the Brown corpus were used in the experiments. One hundred sentences were used as the training data. The other 300 sentences were used as the testing data. The result shows that because the verbs are represented in such a scheme, the accuracy of lexical selection is improved.

## 1.5 Summary

The thesis makes the following claims:

1. The PSR representation for verb semantics is not enough to capture the fine-tuned systemic syntactic semantic correspondence and the semantic interpretation correspondence.
2. To capture these correspondences, a set of semantic components, i.e., cognitive concepts that are sensitive to syntax and interpretation, must be identified.

3. All these cognitive concepts must be organized into a hierarchical structure in order to make a natural connection to world knowledge so that the system can handle metaphor usage and new usages outside the system coverage.

# Chapter 2

# The problem

Verbs are the keys in natural language sentences. Verb semantic representation is therefore crucial to natural language processing systems. In a machine translation system, the verb representation has a strong impact on the quality of the translation from a source verb to a target verb. The following experiments are designed to show that the lexical selection in machine translation is a tough problem. If the source verb is used in a new way that is out of the system coverage, it is very difficult for the system to choose the best or near best target verb as the target realization. In our first experiment, a bilingual transfer MT system TranStar (Dong, 1987) was used to translate a set of sentences in the Brown corpus. Because the system has small coverage on BREAK senses, it has translated most of the *break* verbs wrongly. The second experiment takes human beings as the subjects. Three human subjects translated the same sentences. The result of the translation shows that the lexical selection of human beings is more accurate and flexible. A natural question one may ask is: Why does human do a much better job than the computer system? Later in this chapter, I will try to point out that linguistic incompatibility is a determining factor in the lexical selection process. Humans have the knowledge and ability to handle these incompatibility phenomena. On the other hand, when facing a new usage, human has the knowledge and ability to predict the meaning and find out the best or a near best target word in the lexical selection process. These two types of knowledge and ability are exactly what the computer system lacks. Therefore, we conclude this chapter by proposing the problem of verb semantic representation: How to represent the verb meanings so that the computer system can easily handle the linguistic incompatibility and new usages in the input.

In the following sections, first, two experiments show the differences of human and machine translation of the same material. Second, the Chinese serial verb compound is introduced. Analysis shows the linguistic incompatibility between the English *change-of-state* verbs and Chinese serial verb compounds. Finally, by analyzing the differences of the human and machine lexical selection process, the key issue of verb semantic representation is proposed.

## 2.1 The Experiments

### 2.1.1 Results of a machine translation

The Brown corpus contains 246 English sentences that have 'break' words like 'break', 'breaking', 'broke' and 'broken'. These 'break' sentences cover a wide range of the 'break' usages. They can be viewed as a good sample of unrestricted text with 'break' words. TranStar is a commercial English to Chinese MT system that is currently recognized as one of the best MT systems in China. We have used TranStar to translate these 246 sentences, then the author translated the same material. Since our focus is the lexical knowledge representation and the lexical selection, the analysis will concentrate only on the translation result of the 'break' words. We have removed 89 occurrences from the data as instances of idiom or verb particle constructions such as 'break up', 'break through', etc. For 157 occurrences of break words, the translation result by TranStar can be classified into five groups.

**Group A:** 30 (19.1%) sentences were correctly translated. In the following sentence, the word 'broken' was correctly translated into the Chinese word 被打碎 'beidasui'.

> cc02:18:Another remained when an American Army car was recovered
> but with *broken* glass.

**Group B:** 42 (26.8%) sentences were translated into similar situations. There is a better word which could have matched the situation described by the

English sentence. In the following sentence, the latch was not necessarily broken into pieces, but the translation is 打碎 'dasui' that means 'broken into pieces'.

ca22:53:The younger Thomas ripped a screen door , *breaking* the latch , and after an argument struck his uncle with a rock , scratching his face.

**Group C:** 13 (8.3%) sentences were translated into related situations that are misleading. The 'breaking' in the following sentence was translated into 破坏 'pohuai' which means 'make damage to'.

cl01:33:I do not blame that girl for breaking her engagement with you .

**Group D:** 71 (45.2%) sentences were translated incorrectly. For example, in the following sentence, 'broke' was translated into 打碎 'dasui' which means 'to break into pieces'.

cn02:98:The two horses *broke* from the yard , from the circle of light cast by the lamp still burning in the house , into the darkness.

**Group E:** 1 (0.6%) sentences had no translation result due to a system error.

TranStar is a transfer-based approach. In TranStar's bilingual dictionary, there are only 13 Chinese words mapping to all English 'break' expressions. The lexical selection in TranStar simply makes use of the selectional restrictions to select target language lexical units through this bilingual dictionary. All the 'break' usages have to be translated into one of these 13 Chinese verbs. The accuracy rate of the translation is not high. Only 30 (19.1%) words were correctly translated. The Chinese verb 打碎 (dasui) acts more like a default translation when there is no other good choice. The 13 words are listed below according to their frequencies used for the translation in a decreasing order. The numbers next to the words are the frequencies that the 157 sentences translated into. Some of the zero frequencies are due to the fact that some Chinese verbs are mapped to English 'break' idiomatic usages.

| | | |
|---|---|---|
| 打碎 107 | 破坏 22 | 间歇 14 |
| dasui | pohuai | jianxie |
| to break into pieces | to make damage to | to have a break |
| 决裂 5 | 违反 2 | 爆发 0 |
| juelie | weifan | baofa |
| to break (a relation) | to against | to break out |
| 发生故障 0 | 闯入 0 | 打断 0 |
| fashenguzhang | chuanglu | daduan |
| to break down | to break into | to break a continuity |
| 突破 0 | 得失相当 0 | 违背 0 |
| tupo | deshixiangdang | weibei |
| to break through | to break even with | to break (a promise) |
| 完成绝大部分 0 | | |
| wanchenjuedabufen | | |
| to break with | | |

Table 2.1: The translation result of TranStar

### 2.1.2 Frequencies of human translation

The same 157 sentences are translated by the author into 68 Chinese verbs. The translation can be classified into four groups according the difficulty of the translation:

Group a: 128 (81.5%) sentences were easily translated.

Group b: 1 (0.6%) sentences were translated with a dictionary.

Group c: 15 (9.6%) sentences were translated with some uncertainties.

Group d: 13 (8.3%) sentences had no translations, because the human subject could not find a proper Chinese word for the translation. An example follows:

cg14:11:One looked down on a sea of leaves, a *breaking* wave of flower .

68 Chinese expressions were used to translate these 157 break verbs. They are shown in Table 2.2 according to the frequencies with which they were used for the translation in a horizontally decreasing order:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 打破 | 11 | 打碎 | 8 | 中断 | 7 | 破坏 | 7 |
| 破 | 7 | 坏 | 5 | 休息 | 4 | 违背 | 4 |
| 停顿 | 4 | 打断 | 4 | 违反 | 3 | 碎 | 3 |
| 垮 | 3 | 结束 | 3 | 分离 | 3 | 断 | 3 |
| 打开 | 3 | 闯入 | 3 | 突起 | 2 | 破碎 | 2 |
| 破裂 | 2 | 分开 | 2 | 发布 | 2 | 打破僵局 | 2 |
| 打坏 | 2 | 冲 | 2 | 变成 | 2 | 中止 | 1 |
| 越过 | 1 | 越 | 1 | 训练 | 1 | 完全无效 | 1 |
| 瓦解 | 1 | 挖开 | 1 | 突发 | 1 | 停止 | 1 |
| 逃出 | 1 | 衰亡 | 1 | 衰 | 1 | 切断 | 1 |
| 敲成 | 1 | 破土 | 1 | 破旧 | 1 | 弄断 | 1 |
| 流出 | 1 | 离开 | 1 | 垮掉 | 1 | 砍断 | 1 |
| 绝望 | 1 | 决裂 | 1 | 解除 | 1 | 击 | 1 |
| 分裂 | 1 | 分解 | 1 | 分成 | 1 | 断续 | 1 |
| 断裂 | 1 | 断开 | 1 | 断绝 | 1 | 断交 | 1 |
| 得失相当 | 1 | 打垮 | 1 | 打 | 1 | 冲进 | 1 |
| 崩溃 | 1 | 背弃 | 1 | 背离 | 1 | 爆发 | 1 |

Table 2.2: List of human translations

The distributions of the translated target words are shown in the Figure 2-1. It seems that the nature of the lexical selection in the translation obeys Zipf's law. That means for all possible verb usages, a large portion of verb usages is translated into a few target verbs, while a small portion of verb usages might be translated into many different target verbs.

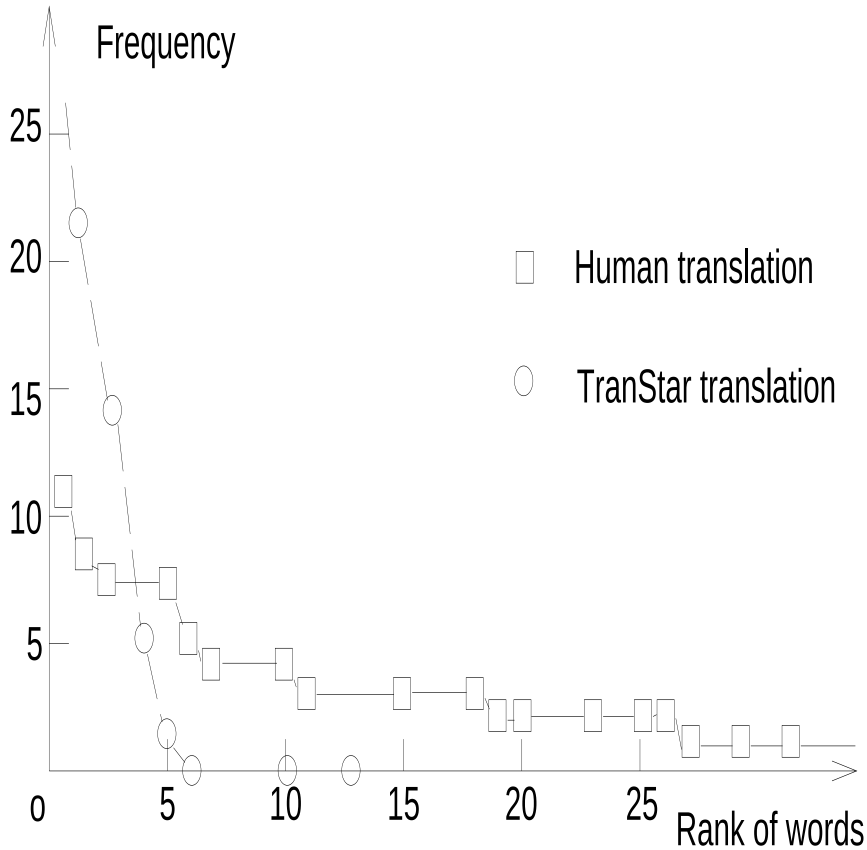



Figure 2-1: Distribution of the translation words

### 2.1.3 Results of human translation

In order to see how BREAK is translated by other independent human subjects, three Chinese university graduates from China were asked to translate the same set of BREAK sentences from the Brown corpus. For 246 English sentences including idioms and verb phrases, The following table shows the number of different target Chinese verbs chosen by the three different human subjects.

| | Subject 1 | Subject 2 | subject 3 |
|---|---|---|---|
| Total number of translations | 148 | 139 | 145 |
| Number of translations used by all three subjects | 33 | 33 | 33 |
| Number of translations used by one and another subject | 50 | 14 | 56 |
| Number of translations used by one alone | 65 | 92 | 56 |

Table 2.3: Results of experiment II

From the table, we can see that the three human subjects have used more than 246 different Chinese words for the translation. The following are some samples of the translations.

- **Three subjects translated the BREAK with the same Chinese words:**

i) cp28:105:Martin would have liked to break the man 's neck .

**Chinese translation**: 拧断 nin duan

**Meaning**: to twist and separate into line-segment shapes.

ii) " Well" , Heiser ventured, " why don't we hold an investigation with questioning and" – " That would be worse than useless", Alexander broke in.

**Chinese translation**: 打断 da duan

**Meaning**: to make into discontinued.

iii) No believer in the traditional devotion of royal servitors, the plump Pulley broke the language barrier and lured her to Cairo where she waited for nine months, vainly hoping to see Farouk.

**Chinese translation**: 打破 da po

**Meaning**: to do some action and make it separated.

- **Two subjects translated the BREAK in the following sentences with the same Chinese words:**

i) President Eisenhower held an 8:30 a.m. meeting with top military and foreign-policy advisers, decided to break off diplomatic relations immediately.

| | Subject 1 | Subject 2 | subject 3 |
|---|---|---|---|
| **Chinese translation** | 断绝 Duan Jue | 中断 zhong duan | 中断 zhong duan |
| **Meaning**: | Stop completely | Stop at middle | Stop at middle |

Table 2.4: Translation result 1

ii) It would be fine publicity for the man who was willing to walk to the mayor's throne over the broken reputation of a helpless girl!

| | Subject 1 | Subject 2 | subject 3 |
|---|---|---|---|
| **Chinese translation** | 败坏 Bai Huai | 损坏 sun huai | 败坏 bai huai |
| **Meaning**: | ruined | make demage to | ruined |

Table 2.5: Translation result 2

iii) cr04:126:Doors that wo- n't open , and doors that wo- n't close and shelves and broken – " " But those are the things I built the workshop for " , protested Mr. Crombie .

| | Subject 1 | Subject 2 | subject 3 |
|---|---|---|---|
| **Chinese translation** | 坏了 Huai La | 破了 po Le | 破了 po le |
| **Meaning**: | damaged | broken | broken |

Table 2.6: Translation result 3

- **Each subject translated the BREAK in following sentences with different Chinese words:**

i) Another remained when an American Army car was recovered but with a broken glass.

| | Subject 1 | Subject 2 | subject 3 |
|---|---|---|---|
| **Chinese translation** | 被打破 bei da po | 碎 sui | 被打碎 bei da sui |
| **Meaning**: | being hit and broken | being in pieces | being hit into pieces |

Table 2.7: Three different translations 1

ii) The will of its people, so crucial in time of peril, would be broken.

| | Subject 1 | Subject 2 | subject 3 |
|---|---|---|---|
| **Chinese translation** | 崩溃 ben kui | 破坏 po huai | 丧失 sang shi |
| **Meaning**: | corrupt | to make demage to | lost |

Table 2.8: Three different translations 2

iii) Miriam now ordered Pengally to break down the gate, but he said he really could- n't go that far.

| | Subject 1 | Subject 2 | subject 3 |
|---|---|---|---|
| **Chinese translation** | 砸破 za po | 捣碎 dao sui | 砸 za |
| **Meaning**: | to hit broken | to hit into pieces | to hit |

Table 2.9: Three different translations 3

This result implies another important fact about the translation. Normally, in statistical corpus based research, a bilingual corpus is used as the translation data, such as the Canadian Hansards. However, this might be only one aspect of data collection. From a different aspect, people might be able to come out with different translations for the same material. As another source of translation data, it is possible to have some statistical results from this new dimension.

**Five translation styles**

When classifying the translation result according to the Chinese verbs, five translation styles are derived:

1) Some sentences are translated into **Chinese collocations**. For example, the 'breaking marriage' is translated into 破裂 (polie) that is a collocation for marriage in Chinese.

> The marriage of John and Mary Black had clearly reached the breaking point after eight years.

TranStar: 打碎, da-sui, to hit it broken into pieces
Myself: 破裂, po-lie, broken and split
Subject1: 破裂, po-lie, broken and split
Subject2: 决裂, jue-lie, to separate a relation without hesitation.
Subject3: 破裂, po-lie, broken and split

2) Some sentences are translated according to **word meanings**. For example, the 'breaking of spectacles' can be translated into 打破 dapo or 打碎 dasui.

> cf37:48:When he had the mishap of breaking his spectacles , his ecumenical colleagues insisted on providing him with new ones .

TranStar: 破坏, po-huai, to make damage to
Myself: 打破, da-po, to hit it broken
Subject1: 打破, da-po, to hit it broken
Subject2: 打碎, da-sui, to hit it broken into pieces

Subject3: 打破, da-po, to hit it broken

3) Some sentences are translated from different **interpretations** by the translators. The following sentences are examples:

ca22:53:The younger Thomas ripped a screen door , breaking the latch , and after an argument struck his uncle with a rock , scratching his face .

TranStar: 打碎, Da Sui, To hit and separate into pieces

Myself: 打坏, Da Huai, To hit it unfunctional.

Subject1: 打断, Da Duan, To hit and separate somewhere along line-segment shape

Subject2: 打断, Da Duan, To hit and separate somewhere along line-segment shape

Subject3: 弄坏, Long Huai, To make it unfunctional.

Miriam now ordered Pengally to break down the gate, but he said he really couldn't go that far.

TranStar: 发生故障, Fa Shen Gu Zhan, to break down (functionality).

Myself: 打开, Da Kai, to open

Subject1: 砸破 Za Po, to hit and broken

Subject2: 捣碎 Dao Sui, Using big wood to hit it into pieces.

Subject3: 砸 Za, to hit

4) Some sentences are translated into **similar concepts**. For example,

End Gene Raesz, who broke a hand in the Owl's game with LSU, was back working out with Rice Monday, and John Nichols, sophomore guard, moved back into action after a week's idleness with an ankle injury.

TranStar: 破坏 Po Hui, to make damage to

Myself: 断 Duan, to separate along line-segment shape

Subject1: 骨折, Gu Zhe, bone separated

Subject2: 伤了, Shang Le, injured

Subject3: 骨折, Gu Zhe, bone separated

5) Some sentences are total **rephrasings of the interpretation**. For example:

> Within a few years the Scots, engaged in breaking the thick sod and stirring the rich soil of the valley, were joined by a group called Meurons.

TranStar: 破坏, Po Huai, to make damage to

Myself: 开垦, kai Ken, to cultivate

Subject1: 破开, Po Kai, to break open

Subject2: 开垦, kai Ken, to cultivate

Subject3: 开荒破土, Kai Huan Po Tu, to cultivate and break the ground.

## 2.2 Translations and the incompatibilities

The above section has shown that the human's lexical selection process is more accurate and flexible. Now, let us investigate the linguistic reason about why there are so many different translations. This section will show that because of the linguistic incompatibility between English and Chinese, English 'change-of-state' verbs must be translated into Chinese serial verb compounds, context information is involved in the target verb realization. Therefore, incompatibility plays an important role in lexical selection.

### 2.2.1 English 'change-of-state' verbs

We first examine two groups of verbs one in English and the other in Chinese. In English, there is a set of verbs which only identify change-of-state events. Levin

(Levin, 1992) has identified several groups of these verbs. They are:

- Group 45.1 break, chip, crack, crash, fracture, rip, shatter, smash, snap, splinter, split, tear
- Group 45.2 bend, crease, crinkle, crumple, fold, ...
- Group 45.3 bake, barbecue, blanch, boil, ...
- Group 45.4 abate, advance, ...
  Change of Color: blacken, brown, redden, ...
  -en Verbs: awaken, brighten, cheapen, darken, ...
  -ify Verbs: acetify, acidify, alkalify...
  -ize Verbs: americanize, caramelize, carbonize, ...
  -ate Verbs: accelerate, agglomerate, ameliorate, ...

According to Levin, these verbs are like 'break' verbs. They are pure verbs about changes of state.

> "the break verbs, unlike the cut verbs, are pure verbs of change of state, and their meaning, unlike that of the cut verbs, provides no information about how the change of state came about." (Levin p. 242)

These verbs have two syntactic alternations. The SVO alternation is the normal $Subject + Verb + Object$ construction. In this construction, the verb identifies concepts in the 'change-of-state' domain and the 'causation' domain. For example, the sentence ' John broke the vase. ' has the following meanings:

Causation: John

Change of state: vase

While the other syntactic OV alternation which is the passive construction only identifies a concept in the 'change-of-state' domain. For example, the sentence 'The vase broke.' has the following meaning:

Change of state: vase

Later on, we will see how these two alternations are translated into Chinese.

### 2.2.2 Chinese serial verb compounds

The other group of verbs are Chinese serial verb compounds. Most of the Chinese words are composed with several different Chinese characters. The meaning of the Chinese word has a firm connection to the meaning of each Chinese characters composing the word. Chinese serial verb compounds are one of these kinds of compounds in Chinese. For some two character verbs, each of them is composed with two single characters and each of the two characters is a verb too. We call this type of verb compound a VV compound. The following is an example of it.

| | | | |
|---|---|---|---|
| 狗 | 赶跑 | 了 | 猫 |
| Gou | gan-pao | le | mao. |
| dog | chase-run | Asp. | cat. |

The dog chased the cat and the cat has run away. (VV)

Here, 'gan-pao' is a VV verb compound composed of two single verbs 'gan' and 'pao'. The syntactic structure is quite special. The first verb 'gan' is the action of the subject while the second verb 'pao' is the resultant action of the object.

For other verb compounds, the first character is an action verb while the second character is a resultant state of the subject or the object. We call these VA verbs. The following are two examples of this construction.

| | | | |
|---|---|---|---|
| 张三 | 吃饱 | 了 | 饭 |
| Zhangsan | chi-bao | le | fan. |
| Zhangsan | eat-full | Asp. | meal. |

Zhangsan has eaten his meal and been full. (VA)

Here, 'chi-bao' is a VA verb compound composed of one verb 'chi' and one adjective 'bao'. chi is the action of the subject, while 'bao' is the resultant state of the subject.

| | | | |
|---|---|---|---|
| 约翰 | 打碎 | 了 | 花瓶 |
| Yuehan | da-sui | le | huapin. |
| John | hit-broken | Asp. | vase. |

John broke the vase. (VA)

Here, 'da-sui' is a VA verb compound composed of one verb 'da' and one adjective 'sui'. 'da' is the action of the subject, while 'sui' is the resultant state of the object. Daoping Wu also presents the theta-role analysis of the above three different syntactic structures (Wu, 1991). According to Wu, the verb compound has two theta-roles that need to be unified with the subject and the object. The following are the GB grammar analyses of the above three sentences.

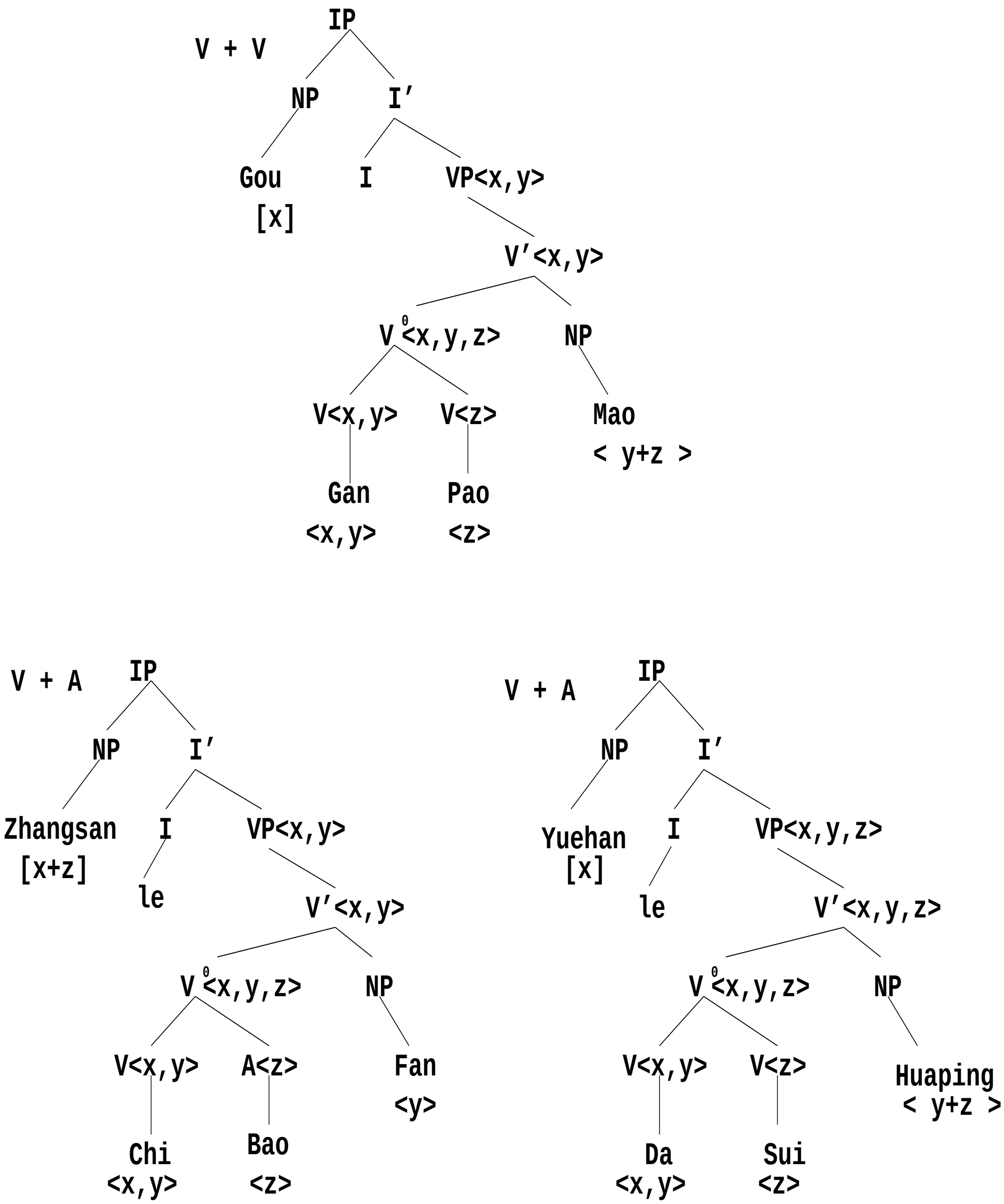



Figure 2-2: GB tree of the sentences

Our study will focus on the VA type of Chinese serial compounds. According to our observation, VA verbs are productive. Many single character verbs and single character adjectives can be combined to form VA verb compounds. The following is a set of examples.

| | | |
|---|---|---|
| 击碎 ji-sui | hit-being-in-pieces | |
| 击断 ji-duan | hit-being-in-line-shape | |
| 击破 ji-po | hit-being-in-irregular-shape | |
| 击开 ji-kai | hit-open | |
| 打断 da-duan | hit-being-in-line-shape | |
| 弄断 long-duan | do-something-being-in-line-shape | |
| 折断 zhe-duan | bend-being-in-line-shape | |
| 压断 ya-duan | press-being-in-line-shape | |

Since the verb compounds are so productive, it seems that we can separate the two characters and view them as two verbs. But this is quite difficult to do. Because the productive compounds must be listed in the lexicon, productivity should be semantically sound. Not any single character verb and single character adjective coming together can form a legal verb compound. For example, the following construction do not exist in the Chinese nature text.

| | | |
|---|---|---|
| * 赶红 | gan-hong | chase-red |
| * 折碎 | zhe-sui | bend-pieces |

The other reason is that some compounds become fixed expressions. For example, the verb 'song-zou' has become a fixed expression. The constructions like 'song-fei', 'song-kai' and 'song-pao' are anomalus.

| | | | |
|---|---|---|---|
| 张三 | 送走 | 了 | 客人 |
| Zhangsan | song-zou | le | keren. |
| Zhangsan | see-off-walk | Asp. | guest. |

Zhangsan saw the guest off and the guest went away.

| | | |
|---|---|---|
| 送走 | song-zou | see-off-walk |
| * 送飞 | song-fei | see-off-fly |
| * 送开 | song-kai | see-off-drive |
| * 送跑 | song-pao | see-off-run |

Since the verb compounds are productive, it seems that we can separate the two characters and view them as two verbs. But this is quite difficult to do.

However, the syntactic properties of the serial verb compounds are so special and the composed verb compounds must be semantically sound, these verb compounds must be listed in lexicon. Chinese serial verb compounds present a challenge to the computer processing system. Due to the active combination power of two Chinese characters, many legal verb compounds have not been listed in the dictionary. The worst thing is that new verb compounds can be continually generated in everyday life. A Computer system should have the knowledge and ability to handle these new verbs.

### 2.2.3 Translations from English verbs to Chinese

When we translate the English change-of-state verbs into Chinese, different alternations have different translations. For an OV alternation, the English verb is translated into a Chinese adjective that describes a resultant state with a Chinese character 'le' as the aspect denoting a past tense. For example, the sentence 'The clothes have dried.' is translated as follows:

衣服 干 了.

yifu gan le.

cloth dry Asp.

The following translation is redundant:

衣服 变 干 了.

yifu biao gan le.

cloth change dry Asp.

For a sentence that describes a resultant state, the Chinese verb shi must be presented. For example: The sentence 'The clothes are dry.' is translated as follows:

衣服 是 干 了.

yifu shi gan le.

cloth be dry Asp.

The following is another set of examples:

The vase broke.

| | | |
|---|---|---|
| 花瓶 | 碎 | 了. |
| huapin | sui | le. |
| vase | in-pieces | Asp. |

A redundant translation:

| | | | |
|---|---|---|---|
| 花瓶 | 变 | 碎 | 了. |
| huapin | bian | sui | le. |
| vase | change | in-pieces | Asp. |

If only the resultant state is described:

| | | | |
|---|---|---|---|
| 花瓶 | 是 | 碎 | 了. |
| huapin | shi | sui | le. |
| vase | be | in-pieces | Asp. |

An interesting phenomenon is the translation of the English SVO alternation into Chinese. The Chinese adjective can no longer be used for the translation. A VA type verb compound must be used for the translation. For example, the following sentence 'Bill dried the clothes.' cannot be translated as follows:

| | | | |
|---|---|---|---|
| * 比尔 | 干 | 了 | 衣服. |
| bier | gan | le | yifu. |
| Bill | dry | Asp. | cloth. |

A natural translation takes a verb compound long-gan as the translation. Here, 'long' means 'to do something'.

| | | | |
|---|---|---|---|
| 比尔 | 弄干 | 了 | 衣服. |
| bier | long-gan | le | yifu. |
| Bill | do-something-dry | Asp. | cloth. |

Another way is to use the default action to select the first action verb. The following is an example. The Chinese character 'hong' means 'bake'.

| | | | |
|---|---|---|---|
| 比尔 | 烘干 | 了 | 衣服. |
| bier | hong-gan | le | yifu. |
| Bill | bake-dry | Asp. | cloth. |

Someone may say that the sentence can be translated into a passive construction where the actor does not need to be specified.

| | | | | |
|---|---|---|---|---|
| 衣服 | 被 | 比尔 | 烘干 | 了. |
| yifu | bei | bier | honggan | le. |
| cloth | Pass. | Bill | bake-dry | Asp. |

The other way is to use 'shi', 'ran' verbs which are similar to 'to make' and 'to let' to translate the sentence as follows to avoid specifying the action.

| | | | | |
|---|---|---|---|---|
| 比尔 | 使 | 衣服 | 干 | 了. |
| bier | shi | yifu | gan | le. |
| Bill | let | cloth | dry | Asp. |

However, statistics shows that the last two types of Chinese sentences are rare in the Chinese corpus.

Following is another set of examples:

John broke the vase.

| | | | |
|---|---|---|---|
| * 约翰 | 碎 | 了 | 花瓶. |
| Yuehan | sui | le | huapin. |
| John | being-in-pieces | Asp. | vase. |

| | | | |
|---|---|---|---|
| 约翰 | 打碎 | 了 | 花瓶. |
| Yuehan | da-sui | le | huapin. |
| John | hit-being-in-pieces | Asp. | vase. |

| | | | |
|---|---|---|---|
| 约翰 | 摔碎 | 了 | 花瓶. |
| Yuehan | shuai-sui | le | huapin. |
| John | throw-being-in-pieces | Asp. | vase. |

| | | | | |
|---|---|---|---|---|
| 花瓶 | 被 | 约翰 | 摔碎 | 了. |
| huapin | bei | yuehan | shuai-sui | le. |
| vase | Pass. | John | throw-being-in-pieces | Asp. |

| | | | | |
|---|---|---|---|---|
| * 约翰 | 使 | 花瓶 | 碎 | 了. |
| Yuehan | shi | huapin | sui | le. |
| John | let | vase | being-in-pieces | Asp. |

### Statistical evidence

Our statistic data is based on the PH corpus. PH corpus (8M bytes) contains publications of the Xinhua News Agency of China from January 1990 to March 1991. We choose four resultant state adjectives and do statistics on sentences that contain one of the four adjectives and the related object in each of the sentences is a concrete object. The result follows:

| Chinese | 破 | 断 | 碎 | 烂 | Sum | Percentage |
|---|---|---|---|---|---|---|
| Meaning | irregular shape | in-line shape | in pieces | worn-out in-scraps | | |
| PinYin | Po | Duan | Sui | Lan | | |
| S + VA + O | 25 | 27 | 15 | 5 | 72 | 67.3% |
| O + VA | 2 | 8 | 3 | 0 | 13 | 12.1% |
| S + A + O | 2 | 0 | 0 | 0 | 2 | 1.9% |
| O + A | 7 | 9 | 0 | 2 | 18 | 16.8% |
| * Cause-type | 0 | 0 | 0 | 0 | 0 | 0% |
| Total | 36 | 46 | 18 | 7 | 107 | |
| | | | | | | |
| A + OBJ | 43 | 12 | 25 | 44 | 124 | |
| All occurrences | 971 | 1066 | 61 | 76 | 2174 | |

Figure 2-3: Statistical evidence

The result shows that single adjectives used in Chinese SVO alternations are very rare. The passive 'bei' construction and causation 'shi' or 'ran' construction are rare too. Chinese tends to use VA compound to describe a change-of-state event.

### A decision tree for lexical section

Therefore, we present a decision tree for translating English 'change-of-state' verbs in SVO alternation into Chinese:

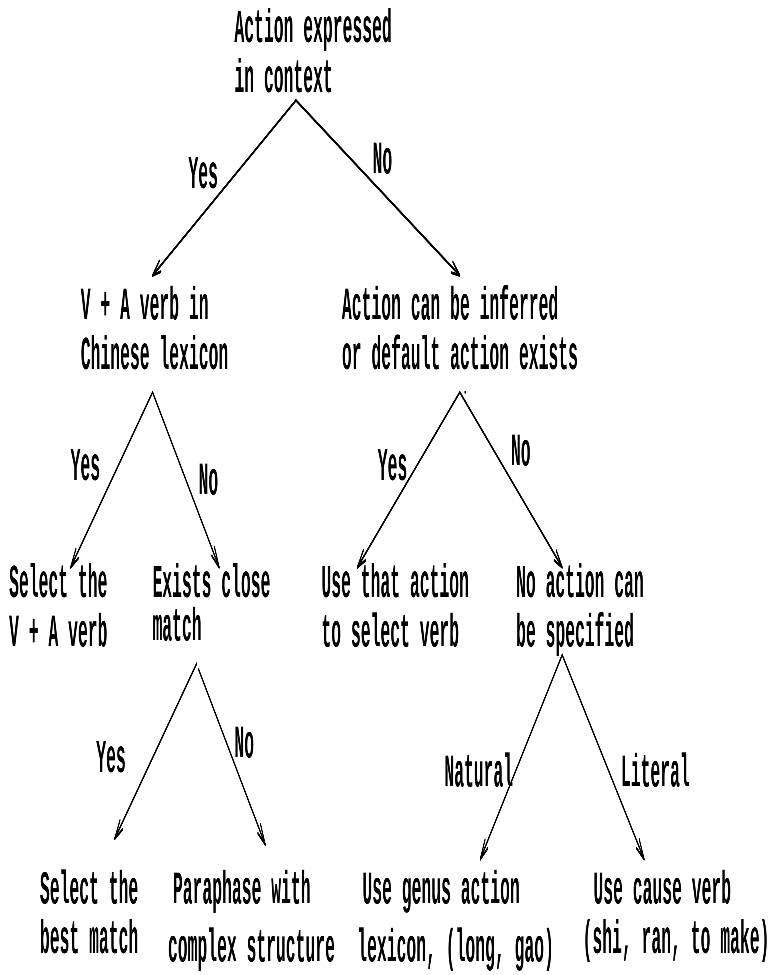



Figure 2-4: A decision tree for lexical selection

In summary, the VA type verb compound must be chosen as having a best match with the situation described by the context. The selection does not rely solely on the verb in the source text. The contexts and default situations play important roles in the selection.

## 2.3 The problem

As we have seen above, due to the linguistic incompatibility between different languages, an English verb such as *break* might have a wide range of target verb translations. To improve the selection accuracy, a natural solution is to enlarge the coverage of the dictionary. In the transfer-based MT system, it is to list more and more source-target verb pairs in the bilingual dictionary. However, this solution cannot handle the new usages outside the system coverage. It is also inefficient in representing the verb meaning. When we look into the human lexical selection process, we find out that rather than having a large storage of translation pairs in memory, human has the knowledge about the linguistic incompatibilities and the ability to predict or guess the new usages. The current computer system is lacking exactly these abilities.

### 2.3.1 Observations

From these experiments, we can have several observations.

First, 'break' is a very adaptable verb in English. It can be used to refer to many different situations which in Chinese are referred to by different expressions. For only 157 sentences, the subject translated them into 68 different Chinese expressions. When different people are asked to translate these sentences, or if the size of the corpus is increased continually, new Chinese words might be continually added into the set as the translation of 'break'. Figure 2-1 shows the frequency distribution of Chinese words used in the translation. It is a "Zipf-like" distribution which many other language facts also obey (Church, 1992). This implies that the set of Chinese words as the translation for 'break' can be very large, if not open ended. For a given English corpus, most of the words in the set are used in a very low frequency or not at all. This presents in the difficulties for the pure statistical approach. Because of the sparse data problem, a pure statistical approach cannot handle the linguistic incompatibilities well. Besides the processing of large corpora, the knowledge-based processing and generalization on facts are still useful to compensate the statistical approach.

Furthermore, human subjects can always select the best or near best word to fit the situation by approximate reasoning. The decisions are based on not only the single 'break' word, but also the syntactic structure and the semantic properties of the arguments. Some conceptual distinctions have played a very important role in lexical selection. For example, the shape of the 'branch' (a line segment) supports the choice of 打断 'daduan' for 'break the branch', and the brittle composition of the 'bowl' supports the choice of 打碎 'dasui' for 'break the bowl'. Here 碎 'sui' means 'pieces' and 断 'Duan' means 'separated into two'. Chinese tends to describe the 'break' situations more precisely than English. The linguistic incompatibility discussed above exists in many places among different languages. In English, 'break' situations for concrete objects are referred to by the same word 'break', but in Chinese, these situations are further subcategorized and referred to by several different Chinese words which are sensitive to certain conceptual distinctions. These conceptual distinctions are rooted in common sense knowledge and cognitive ability. They have probably gone beyond the coverage of the semantic features used in conventional MT systems.

Moreover, the differences between human and machine are that human has the knowledge about those linguistic incompatibilities and therefore can predict the meaning of the new usages. In the next chapter, we will see that most of the current verb semantic representation methods ignore the linguistic incompatibilities problem and provide no way to handle the new usages in the input.

### 2.3.2 Issues

Given the above observations, we can define some fundamental issues for representing verb semantics in MT systems.

First, the **lexical dictionary** must satisfy the following requirements:

a) Certain conceptual distinctions are necessary to resolve linguistic incompatibilities. The semantic features and selectional restrictions in conventional MT systems offer a partial solution; but they are insufficient.

b) We will never have complete knowledge stored in a computer system. The dictionary must be designed in such a way that the system can easily acquire new knowledge both manually and automatically.

c) The system needs to be fail-soft for incomplete information from the input or incomplete knowledge in the system. The system should be able to find out the best solution for all possible inputs, and has a quality measure for its own output based on the knowledge in the system.

Secondly, the **analysis** should not only reach a level deep enough to resolve the linguistic incompatibilities, but also be able to handle metaphorical usages by analogical reasoning.

Thirdly, due to the linguistic incompatibilities, there are rarely exact matches between different languages. The **generation** process must be able to perform approximate reasoning to select the best lexical item to match the analysis results.

The key to all of these issues is the representation and manipulation of the lexicon. In conventional transfer-based systems, source words and target words are organized as one-to-one pairs in the bilingual dictionary. When considering verbs that have a very large number of translations in the target language based on deep cognitive features, the bilingual dictionary seems to be an inefficient method for representing the necessary information. The structure of the dictionary also makes it difficult to measure the similarity among words. Therefore, the desired analogical reasoning in analysis and the desired approximate reasoning in generation are very difficult. On the other hand, as mentioned above, even in a conventional knowledge-based approach, such as UNITRAN, although an interlingua dictionary is used, the representation does not provide a similarity measure for different senses. The system still has the difficulty in performing inexact matching (Dorr 1990 p. 283). So our task becomes clear: to find ways to encode the verb meanings so that the lexical selection can be more accurate and flexible, and the system can easily handle the new usages.

# Chapter 3

# Lexical selection overview

Given the problems defined in Chapter 2, In order to improve system performance and coverage, we should handle the linguistic incompatibilities and new usages properly. In translating the English sentences to Chinese, the verb sense definitions in the MT dictionary are the key factor for the system performance. In a transfer-based bilingual dictionary such as TranStar, each source verb sense is linked to its target translation. If a verb sense is missing in the dictionary, the system provides no way to guess the meaning. On the other hand, due to linguistic incompatibilities, the conditions for choosing the correct target verb are context information and default common sense knowledge, as well as the selection restrictions on the verb arguments. So it is not enough to solely depend on selectional restrictions. This chapter will examine the lexical selection process in some of the MT systems. It will show that the scope and accuracy of lexical selection are heavily influenced by the representation schemes used in these systems.

## 3.1 The lexical selection problem

Lexical selection is to choose the best word in the target language to convey exactly the same meaning in the source text. For the verb selection problem, the accuracy of the selection is determined by two processes. One process is the source verb analysis process that disambiguates the verb sense. If the system cannot find out the correct sense and construct a correct meaning representation, the correct target verb will not be chosen. The other process is the target verb realization process. If the target verb has very limited definitions, the internal meaning may not be realized to a correct target verb.

In the source verb analysis, between the new input and the source verb sense representation, there are two different situations:

**Case A1: New usage** The input is a new usage out of the verb sense coverage. For example, if the system only encodes the 'hit' sense as in 'The man hit the table', the 'hit' in 'the price hit the top line' will be a completely new usage. Therefore, if the system does not have any ability to guess the meaning of the new usage, the input might be rejected or a wrong internal meaning representation might be produced.

**Case A2: Exact match** The verb sense of the input sentence is exactly the same as what has been encoded in the system. As in the above example, if the input is the sentence 'the man hit the table', the meaning representation of the system can be directly constructed with the verb sense representation of 'hit'.

The interesting case is the Case A1. If the new usage does not have any relations with the existing encoding, the system will not be able to do anything with the new usage. However, most of the time, the relations between the new usage and the encoded senses fall into the following two types:

**Type I: Metaphor** The new usage is a metaphor of the encoded sense. For instance, the English sentence 'The price hit the top line' is a metaphor of 'the book hit the table'. Both sentences imply some sort of 'contact'. The first one contains non-concrete objects. The 'contact' can be viewed as a kind of 'reach'. The sentence means that 'the price reached the top line'. While in the second sentence, the 'contact' is a physical contact. If the system knows that the 'price' and the 'top line' can have some sort of 'reach' relation, and that 'reach' and 'physical contact' can have a some sort of analogous relation, the system can predict the meaning of 'hit' in this new usage which might be a kind of 'reach'.

**Type II: Relaxation** As we know, sense defines a class of usages. A set of conditions draw a boundary for the usage class. If a new usage is near the boundary or just on the boundary, the system might rule out the new usage. However, if the conditions are relaxed a little bit, the new usage might be included into this class. The sense definition can be adjusted accordingly. For instance, Suppose the system has a sense definition for 'break' as in the sentence 'the man broke the watch' which means 'the object loses its mechanical function' and the condition is 'the object is a mechanical device'. If the new usage is 'the man broke the language barrier', the 'language barrier' can be viewed as a functional device. If the system can relax the constraint to 'the object is a functional device', the sense can be adjusted to 'the object loses its function'.

In the target verb realization process, there are three different situations between the internal meaning representation and the surface target word expression.

**Case R1: No Match** There is no target word that carries the meaning or similar meaning as the internal meaning representation. For instance, the Chinese idiom "WeiWeiJiuZhao" refers to an ancient event in the history:

> At the Warring States Age (475-221 B.C.), The Wei State attacked the Zhao State. Zhao then asked help from the Chu State. Instead of attacking the Wei army directly, the Chu state besieged Wei State so that Wei had to withdraw the army.

The meaning is that "to rescue the besieged by besieging the base of the besiegers". Due to the cultural differences, we cannot find an appropriate English expression to deliver the meaning. It seems that the paraphrase "to rescue the besieged by besieging the base of the besiegers" is not enough. The only way to clearly state the meaning behind this phrase is to tell the story.

Therefore the best translation will be to create a new word in English and let it have the same meaning as the Chinese phrase, while leaving a note behind to tell the story.

**Case R2: Exact match** The target language has the exact surface realization for the internal meaning representation. For instance, the English sentence "The man cut the bread" can be precisely translated into Chinese as 那人切面包 "na ren qie mianbao". Here, no matter what the internal meaning representation is, the Chinese verb 切 "qie" means precisely the same thing as the English verb "cut".

**Case R3: Inexact match** Most often, due to the following reasons, there is no exact match. From the point of view of evolution, culture, language and cognition are inter-related, and have firm connections among each other. The language is developed together with the history. The language expression represents the special categorization of the human cognitive process. For instance, in English, the sentence "The man broke the window" means "The man made the window separated", no action has been explicitly mentioned in the sentence. However, in Chinese, normally, people say 那人打破窗子"na ren dapo chuangzi", the verb 打破 "dapo" here means "Hit and break". To say that 那人破窗子"na ren po chuangzi" (that man broke the window) is an anomaly. This shows that the cognitive categorization incompatibility is the reason behind the linguistic incompatibility.

The interesting case here is Case R3. If both the source verb and the target verb are defined independently on an internal representation, the problem of linguistic incompatibility is unavoidable. If there is no relation among the internal concepts, the similarities between the source verb and the target verb are hard to measure.

In the above analysis, we can see that the abilities of handling new usages and the ability of selecting similar target verbs are very important for a correct translation. In the following sections, we will see whether different approaches have these abilities.

## 3.2 The direct replacement approach

As its name suggests, the direct replacement technique means that the internal representation can be directly replaced by the surface expressions. Early generation systems such as Winograd's SHRDLU system (Winograd, 1972) employed this technique. It allowed the researchers to use expressions directly from the internal representations and eliminated any need for a text planner or explicit representations of goals. Particular internal representations will activate particular procedures to generate particular surface expressions. The linguistic knowledge is encoded in these procedures. Taking SHRDLU as an example, for the following internal representation,

(#puton :B6 :table)

the generation procedure is a fragment of Lisp code:

(append (vbfix 'put) obj1 'on obj2)

The message #puton becomes an instruction in a program. It actually becomes a function call whose arguments are the objects :B6 and :table. The Lisp function *vbfix* is used to change the grammatical form of the verbs. This function in SHRDLU is defined very simply. For answering a 'how' or a 'when' question, the ending '+ing' is used. For answering a 'why' question, the infinitive form is used. This technique is only good for its data-driven inference. The internal representations and the realization procedures are connected so tightly that the surface expressions come out immediately, no extra inference is needed. Hence, no extra-inference mechanism is needed except those data-driven procedures. However, this grammatical information procedure encoding leads to a design that is unwieldy and awkward to extend. The lack of systematic treatment of the grammar greatly restricts the system coverage and performance. In the above example, the procedure is encoded such that the relation name *#puton* is definitely realized as *put*. There are no other conditions or choices for the system to realize the internal representation as another target word.

The direct replacement approach is seldom used in machine translation today. Because there are no source verb analysis steps and target realization steps, a source verb is simply linked to a target verb. The whole translation process is more like a dictionary look up or a table match. There is no way in this process to handle the new usages and select similar target verbs. The method can be shown in the following figure:

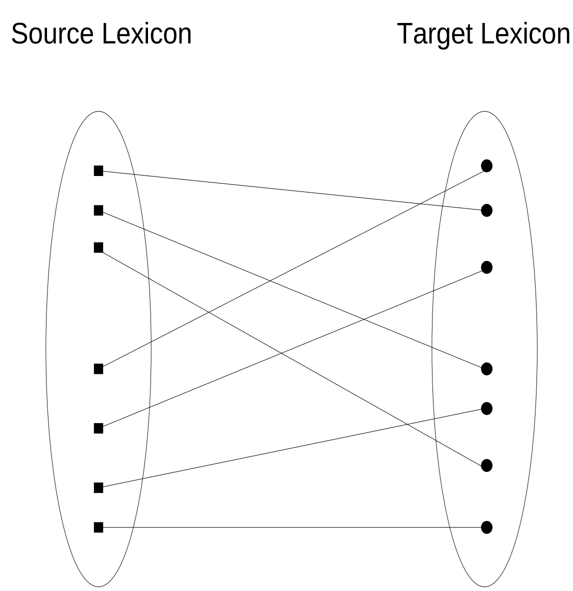



Figure 3-1: Direct replacement

## 3.3 Lexical selection in transfer approaches

In the transfer-based machine translation system, a source language expression is analyzed into an intermediate representation, then the intermediate representation is transferred into the target language representation. Based on the target intermediate representation, the target language surface expression is generated. The lexical selection in the transfer module is the lexical transfer, that is the replacement of a source item by a target lexical item. In the source verb analysis, different verb senses might have different intermediate representations. Before composing the intermediate representation, the verb sense must be selected. A normal solution to this problem is to use selection restrictions. The choice is made according to the restrictions on the verbs' arguments, i.e., the selectional restrictions of verbs. For example, English BREAK is translated into 打断 'daduan' when the object has a line-segment shape, and 打破 'dapo', if the object is a bowl. The knowledge of the mapping from the source lexicon to the target lexicon is in the bilingual dictionary. For each of the lexicons in the source language, a set of the target language lexicon is listed. A set of the conditions on the arguments of the verb is attached to each of the target lexicons as the selectional restrictions. This can be shown in Figure 3-2.

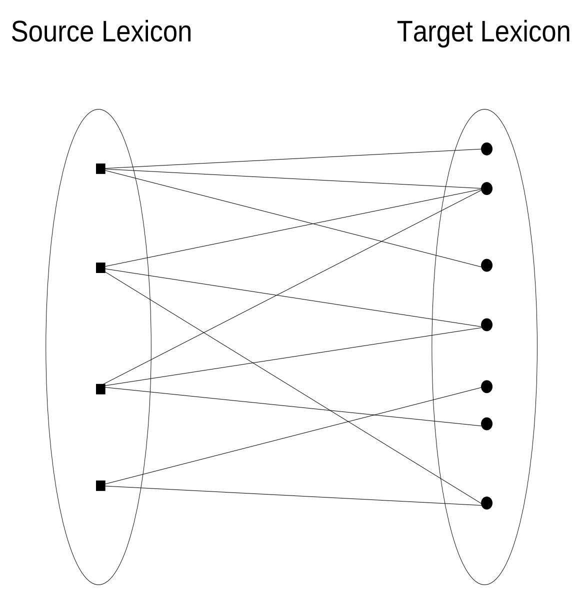



Figure 3-2: Selection with restrictions

Because the relations among different senses and the relations among selection restrictions have not been encoded in the dictionary, the relaxation of the constraints and the analogical reasoning cannot be performed. Therefore, the system has difficulty handling new usages and selecting similar target verbs.

## 3.4 Knowledge-based approaches

Knowledge-based approaches emphasize a deep understanding of the source text. This is the prerequisite for a successful translation. In KBMT system, a word has a very rich description. For instance, Nirenburg (Nirenburg et al., 1992) defines each word entry to have up to 10 zones. They are:

1. grammatical category (CAT slot)
2. user information (STUFF slot)
3. orthography (ORTH slot)
4. phonology (PHON slot)
5. morphology (MORPH slot)
6. syntactic features (SYN slot)
7. syntactic structure(SYN-STRUC slot)
8. semantics (SEM slot)
9. lexical relations (LEXICAL-RELATIONS slot)
10. pragmatics (PRAGM slot)

The particular interested zone is the SEM zone. Nirenburg defines one sense of **eat-v1** in SEM zone as follows:

SEM:

```
(%ingest
  (AGENT (value ^$var1)
         (sem *animal))
  (THEME (value ^$var2)
         (sem *ingestible)
         (relaxable-to *physical-object))))
```

One can see that the verb semantic definition in SEM zone is simply a paraphrase definition. There is no meaning decomposition in the representation. Therefore, the fine-grained relations among different verb senses cannot be encoded in the system. On the other hand, the semantic meaning of **eat** is reduced to a super concept **ingest**. After the meaning has been identified, the system reasoning is based on **ingest** rather than **eat**.

It has not been demonstrated that KBMT system can handle the inexact match between verb meanings. However, inexact match between nominals has been handled. For a nominal concept, the properties and property values that characterize the concept are used to define a meaning distance metric. Nirenburg (Nirenburg et al., 1992) presents an example. The concept that needs to be realized is:

a person who is male and among 13 to 15 years old.

This concept can be represented as:

```
(make-frame person-5
   (is-token-of (value *person))
   (sex (value male))     (age (value (<> 13 15))))
```

The candidate English words are listed below:

boy, kid, teenager, youth, child, young man, adolescent, man , human

Some of the words are represented below:

```
(make-frame boy
   (is-token-of (value *person))
   (sex (value male)
      (importance 10))
   (age (value (<> 2 15))
      (importance 4))
```

(make-frame kid
    (is-token-of (value *person))
    (sex (value male)
        (importance 2))
    (age (<> 5 15))
        (importance 6))

By taking 'sex', 'age' as the properties, the above words are defined with property values and importance values. The importance values serve as a rank for those properties with respect to the relevance of their values for the identity of a reference. By providing the meaning distance metric for different words, an inexact match can be performed and the 'boy' can be chosen as the surface realization.

Due to the inadequate representation of verb semantics, their approach does not include any measure for verb meaning distance or similarity. The verb selection is still done by selection restriction which is the same as in a transfer-based system.

## 3.5 Statistical approaches

Statistical machine translation is a self-organized type of approach. It was first proposed by Peter Brown and his colleges at IBM Thomas J. Watson Research Center in 1990 (Brown et al., 1990),(Brown et al., 1991). The approach is currently restricted to the sentence level. The high level involvement in the translation of information such as discourse, the author's style and background have not been considered. The basic idea employed here is that every sentence in one language is a possible translation of any sentence in the other in the degree of a probability measure. If we assign a probability to each pair of sentences as $P(T \mid S)$, then the problem of translation can be viewed as: Given a sentence T in the target language, we seek the sentence S from which the translator produces T and maximizes $P(S \mid$

$T$). We have

$$P(S \mid T) = \frac{P(S)P(T \mid S)}{P(T)} \tag{3.1}$$

According to the above formula, the statistical machine translation system has the following parts:

1. A method to compute the language model probabilities P(T).
2. A method to compute the translation probabilities $P(T \mid S)$.
3. A method to find out the possible source sentences S and the probability P(S).

Lexical selection is done by calculating the target sentence probability. The raw modeling data comes from the alignment of the bilingual corpus. An alignment as defined in (Brown et al., 1990) indicates the origin in the English sentence of each of the French words in the French sentence. One English word may produce zero, one or more French words. The number of French words that an English word produces in a given alignment is referred to as its *fertility* in that alignment. An example of this alignment is shown in Figure 3-3. The probability of the

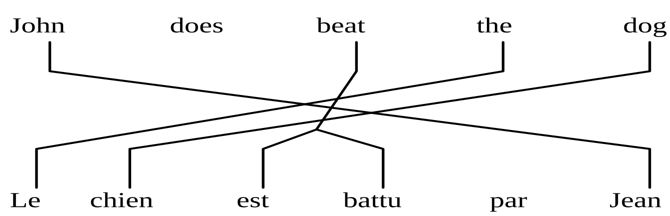



Figure 3-3: An alignment example

translation model is defined as the product of the word's alignment probability and its fertility probability. For the alignment *T=(Le chien est battu par jean)* | *S= (john(6) does(0) beat(3,4) the (1) dog(2))* is as follows:

$$\begin{aligned} P(S \mid T) &= P(fertility = 1 \mid John) \times P(Jean \mid john) \\ &\times P(fertility = 0 \mid Does) \\ &\times P(fertility = 2 \mid beat) \times P(est \mid beat) \times P(battu \mid beat) \end{aligned}$$

$$\times \quad P(fertility = 1 \mid the) \times P(Le \mid the)$$
$$\times \quad P(fertility = 1 \mid dog) \times P(chien \mid dog)$$
$$\times \quad P(fertility = 1 \mid < null >) \times P(par \mid < null >) \qquad (3.2)$$

Here, the parameters of the translation model are a set of fertility probabilities $P(n \mid e)$ for each English word e and for each fertility n from 0 to a moderate limited number; a set of translation probabilities $P(f \mid e)$, one for each element f of the French word and each number e of the English word; and a set of probabilities $P(i \mid j, l)$ for each target position i, source position j and target length l.

This approach gets ride of the verb sense definition and representation problem. The translation accuracy depends on the capacity of the statistical model. The performance of the system is influenced by the size of the training corpus, the availability of the cooked corpus, the quality of the statistical model. So far, no system using this approach has demonstrated good results in machine translation.

## 3.6 Summary

From the development of the lexical selection methods, we can see that the representation of the verb semantics has progressively become fine tuned. At the stage of direct transfer, the verb semantic representation is only a paraphrase point to a sequence of actions. There is no context information involved in the representation. At the stage of transfer approaches, the verb semantic representation has taken the context information into consideration. The arguments of the verbs are used to set the boundaries of the verb senses. However, most of the schemes are emphasized on the linking between syntax and semantics. Verb semantics is represented by an abstract symbolic notion together with a set of thematic roles that are linked to syntactic functions. Later at the stage of KBMT, more information has been encoded in the verb entry, such as orthography, phonology, morphology and pragmatics. However, as it is demonstrated in Section 3.4, the verb semantic knowledge encoding is still at the same level as the transfer approach, i.e., a symbolic notion together with a set of thematic roles. We have named this a PSR scheme in Chapter 1.

Yet in another stream of development, with the hope of building a good model for making the right choice from a source word to a target word, the self-organized statistical approach uses an n-gram language model to do lexical selection. This approach is designed without any knowledge representation but only with the training data. However, it is hard to design a statistical model good enough to handle all of the possible linguistic incompatibilities.

In either case, none of the streams has called for a deeper verb semantic representation than merely thematic roles. As will be discussed in Chapter 7, the verb semantics must go as deep as the cognitive level and must be rooted in the knowledge base hierarchy. In the following chapters, I will first review the problem of verb sense definition, representation and organization, and then show that fine-tuned relations among verb senses must be represented and how this representation handles new usages properly.

# Chapter 4

# Literature Review I: Lexical Knowledge Hierarchical Organization

Before discussing the verb representation, let us do some review of the lexical knowledge organization and the theories about verb semantic representations. In this chapter, we will discuss the question as to how to organize the lexical knowledge in the system and how to let the same amount of knowledge achieve its maximum performance and efficiency. This has long been viewed as only an implementation issue instead of a theoretical one. For computational linguists, the motivation for an inheritance organization is limited to the consideration of computer storage efficiency. With the bias towards emphasizing the representation of grammatically relevant parts of the verb lexicon, the recent approaches in this area focus on the issue of how to efficiently represent the verb's syntactic information. Inheritance networks are used for efficient memory manipulation. In the following, some recent works are discussed.

## 4.1 Why inheritance?

First, let us see why inheritance is needed in a computer system. Natural language interfaces to computers must deal with a wide variation in real world input or unrestricted texts. The real world input often has missing words and variant syntax. And for the unrestricted text, a system often suffers overgenerate or undergenerate problems. [1] To understand a natural language utterance, the knowledge about the lexicon, syntax, semantics and the world is needed. A human can easily understand real world input because he grasps more knowledge and has a smart

[1] Overgenerate means that the grammar might predict that a sentence is grammatical, when it is not. Undergenerate means that the grammar might predict that a sentence is not grammatical, when it is.

reasoning mechanism to help him to make decisions. Why do NLP systems have difficulty in doing that? This is because they are lacking sufficient knowledge coverage. Thus for a practical NLP system, the key problems are: 1) Decide which knowledge is needed for the task. 2) Decide the knowledge representation strategy. 3) Decide the knowledge reasoning mechanism.

Generally speaking, a NLP system is the same as a human being. It cannot possess all the knowledge in the world. A system can only solve a particular natural language task. Therefore all NLP systems are task-limited systems. The system domain can be defined based on a special task. In a conventional NLP system, the knowledge can be classified into different types: declared knowledge; implicit knowledge; system knowledge; domain knowledge; and world knowledge. Let me use an example to elaborate this point.

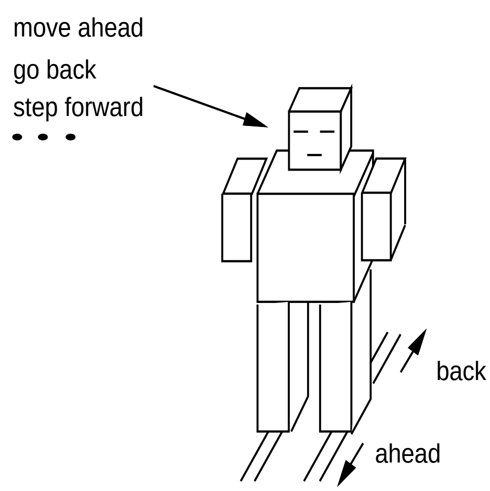



Figure 4-1: A robot system

Let us consider a NLP system that is built into a robot to accept human natural language instructions. Suppose the robot can only take two actions: 'move ahead' and 'move back'. The input consists of real world instructions made by a human. We can further suppose that the following knowledge has been declared in the system.

- **[Facts:]** The following phrases are 'move ahead' instructions:
  move ahead, go ahead, walk ahead
  move forward, go forward, walk forward

- **[Facts:]** The following phrases are 'move back' instructions:
  move back, go back, walk back

- **[Rules:]** Suppose the system has the generalized rules:
  'If the phrase is composed of two words, the first is a verb which is a space action and the second is a preposition which means a forward direction, then the robot moves ahead.'
  'If the phrase is composed of two words, the first is a verb which is a space action and the second is a preposition which means a back direction, then the robot moves back.'

With the example above, we define the five types of knowledge as follows:

**Definition 4.1 (Declared knowledge)** *The declared knowledge includes facts, rules which are explicitly represented by the NLP system.*

In the above example, all the facts and rules are declared knowledge.

**Definition 4.2 (Implicit knowledge)** *All the knowledge that has not been explicitly declared in the system, but can be inferred based on the declared knowledge with the inference engines.*

**Definition 4.3 (System knowledge)** *System knowledge includes the knowledge explicitly declared and the knowledge implicitly implied by the system.*

**Definition 4.4 (Domain knowledge)** *Domain knowledge is all the knowledge needed by the system to solve the particular task.*

**Definition 4.5 (World knowledge)** *World knowledge is all the knowledge in the world.*

The system knowledge of the robot system does not cover all of the domain knowledge. If we say: "forward one step", the robot system cannot correctly interpret this sentence. The domain knowledge for the robot system includes all the knowledge concerning space, motion, direction, and the related linguistic knowledge. However, the domain knowledge is not necessarily all world knowledge. For example, the robot system need not know the fact that "NLP is a difficult problem". The relations between these diferent forms of knowledge can be shown in Figure 4-2

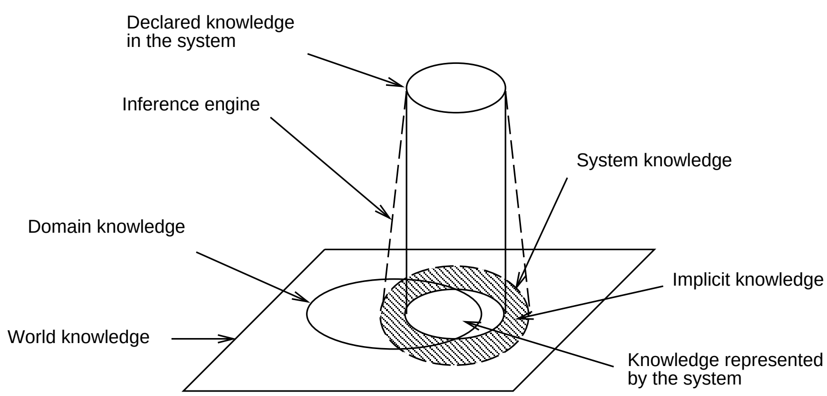



Figure 4-2: Relations among different types of knowledge

From the figure we can see that as long as the system knowledge covers more of the domain knowledge and less of the knowledge outside the domain knowledge, the system will become more efficient and can handle the real world input better. However, because of the huge amount of world knowledge and a vague boundary of the task domain, a full and efficient coverage is difficult to achieve.

Today's NLP systems suffer the same problem as in the example above, the difficulty of implementing an NLP system efficiently for processing real world input or unrestricted texts. There are many different NLP approaches that deal with this problem. When these approaches are applied to an idealized situation, they can lead to a good solution. However, the problem is that ideal situations are seldom encountered.

Inheritance can be an efficient way to pursue the goal. By using a structured lexicon, the system will be more efficient in achieving its performance than when using the non-structured lexicon. A conceptual hierarchy will certainly improve the efficiency of representation for domain knowledge, as well as for lexical knowledge.

## 4.2 Evans and Gazdar: Inference in DATR

DATR (Evans and Gazdar, 1989a)(Evans and Gazdar, 1989b) is a declarative network representation language with two principal mechanisms: orthogonal multiple inheritance and nonmonotonic definition by default. The primary unit of a DATR network description is called a NODE and consists of a set of PATH/DEFINITION pairs where PATH is an ordered sequence of arbitrary atoms enclosed in angle brackets, and DEFINITION is either a value, an inheritance specification or a list of definitions enclosed in round brackets. The primary operation on a DATR description is the evaluation of a QUERY, namely the determination of a value associated with a given PATH at a given NODE. Such a value is either defined directly for PATH at NODE or obtained through an inheritance specification for PATH at NODE or determined from the definition for the longest subpath of PATH defined at NODE, when PATH itself is not defined at NODE.

There are several different inheritance specifications in DATR. DATR provides a new node, a new path or both to seek a value form. The basic form of inheritance is called LOCAL inheritance, which changes the node and/or path specification in the current context. For example:

```
Node1: Path 1 == Node2
```

```
(inherit value from Path1 at Node2)
 Node1: Path 1 == Path2
(inherit value from Path2 at Node1)
 Node1: Path 1 == Node2:Path2
(inherit value from Path2 at Node2)
```

The second form of inheritance is called GLOBAL inheritance. It changes the node and/or path specification in the saved global context initially set to the node/path pair of the query and inherits from the new global context:

```
Node1: Path 1 == ``Node2''
       (set global node to Node2 and inherit
       value from global node/path)
Node1: Path 1 == ``Path2''
       (set global node to Path2 and inherit
       value from global node/path)
Node1: Path 1 == ``Node2:Path2''
       (set global node and path and inherit)
```

When a requested path is not defined at a node, the longest subpath is used to provide a definition, with all the paths in the definition specification extended by the extra requested atoms. Thus if paths <a b c> and <a b c d> are defined at Node1, a definition such as:

```
Node1:<a b> == Node2:<x>.
```

implicitly defines both the following:

```
Node1:<a b c> == Node2:<x c>.
Node1:<a b c d> == Node2:<x c d>.
```

This 'definition by default' (in the absence of any more specific path definition) gives DATR its nonmonotonic character: add a definition to a node and some of those previously valid definitions, but derived by this default mechanism, may cease to hold.

DATR turns out to be a good inheritant lexical representation language. However, since the major focus of DATR is inheritance and default reasoning, it only covers lexical relations partially. By knowing the syntactic information of one word, there is no way for DATR to retrieve those words that have similar syntactic behavior. Although the inheritance properties make it easy to get the supernodes of a word according to some features, there is no way to get the subnodes of a word. When word semantics is considered, this style of structured lexicon that maintains only the inheritance relation will not be enough for defining a good lexicon.

## 4.3 Flickinger and Nerbonne: Easy adjectives

Flickinger and Nerbonne (Flickinger and Nerbonne, 1992) illustrated that structured lexicon has its practical advantages. By analyzing a concrete example – the linguistic properties of the *easy* adjective, they show that the structured lexicon is easy to be maintained, modified and extended. The following example sentences are used to illustrate the properties of *easy*, such as obligatory and optional subcategorization, control, long-distance dependency, optional modification, and specification.

(1) a. Bill is easy to talk to.

b. It is easy to talk to Bill.

c. Bill is easy for Mary to talk to.

d. It is easy for Mary to talk to Bill.

HPSG (Head-driven Phrase Structure Grammar) is used as the demonstration vehicle for the lexical demands of grammatical analysis. They also listed a set of other adjectives that have the same properties. We refer to them as belonging to the *easy* class. This class is further divided into two subclasses. One is the **IT-EASY** class that contains lexical entries for the words having the properties displayed in sentences (b) and (d). The other is the **SLASH-EASY** class which contains lexical entries for the words having the properties displayed in sentences (a) and (c). According to Flickinger and Nerbonne's analysis, these two classes are located at the middle of a word hierarchy. On the top, there is the **CONTROL** word class that introduces a verbal complement subcategorization. Between the top and these two classes, there are two other classes **IT-SUBJ** and **SLASH-COMP**. **IT-SUBJ** is a superclass of **IT-EASY** that includes the set of adjectives like *possible*. The **SLASH-COMP** is a superclass of **SLASH-EASY** that includes the set of adjectives like *pretty*. So the whole hierarchy comes out as follows. It is directly involved in the definition of complex adjective lexical entries. The hierarchy shows that the same information for different word classes need not be represented redundantly (Flickinger and Nerbonne, 1992) p. 280.

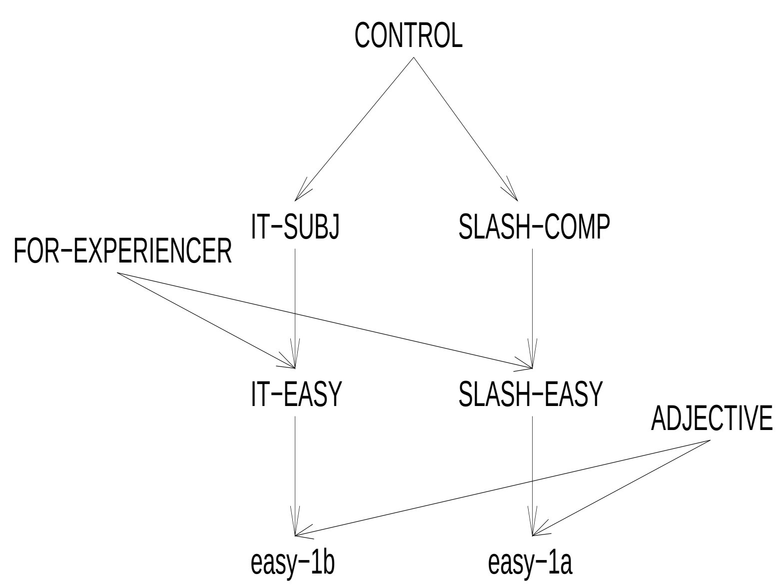


Figure 4-3: The structure of word classes

Flickinger and Nerbonne also used adjective classes like *pleasure* and *too* to show that the hierarchy can be easily extended with very little modification, merely addition. This is because significant classes are identified in a detailed grammatical description. The modified hierarchy is shown below (Flickinger and Nerbonne, 1992)p. 290.

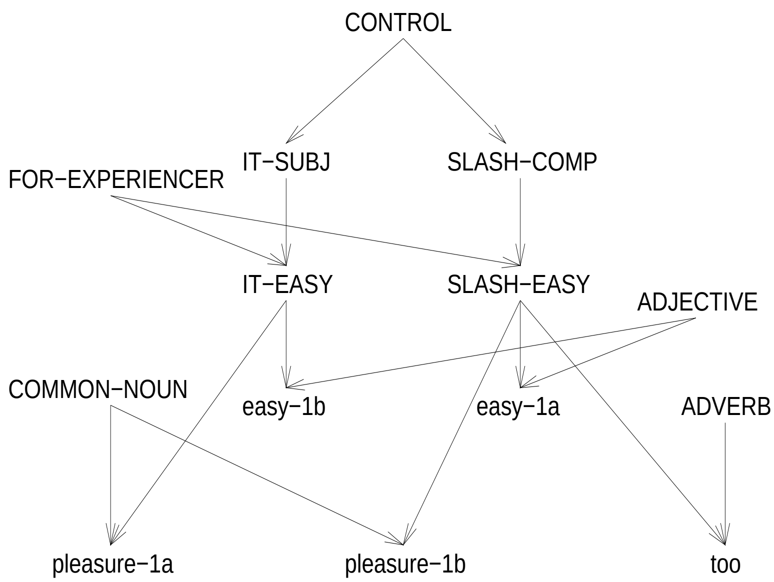



Figure 4-4: The extended hierarchy

## 4.4 Kilgarriff: Inheriting verb alternations

Kilgarriff (Kilgarriff, 1992) built a hierarchy according to (Levin and Rappoport, 1991) by using DATR. Levin (Levin, 1992) has listed a set of English syntactic alternations and verb classes having the same syntactic behaviors. Kilgarriff shows how the verbal lexicon can be formalized in a way that captures and exploits generalizations about the alternation behavior of verb classes. The HPSG is again used as the grammar formalism to represent linguistically related lexical knowledge. For example, "bake" in the sentence "John baked the cake" is represented by the Attribute-Value Matrix (AVM) notation:

$$\begin{bmatrix} \textit{WORD} & \textit{bake} \\ \textit{SYN} & \begin{bmatrix} \textit{MAJ} & \textit{V} \\ \textit{SUBJECT} & \textit{(NP[NOM] SEM @1,} \\ & \textit{NP[ACC] SEM @2)} \end{bmatrix} \\ \textit{SEM} & \begin{bmatrix} \textit{RELN} & \textit{BAKE} \\ \textit{BAKER} & \textit{@1} \\ \textit{BAKED} & \textit{@2} \end{bmatrix} \end{bmatrix}$$

Figure 4-5: AVM for *transitive bake*

For *ergative bake* in the sentence "The cake baked", the same BAKE relation holds in the base form, but the unspecified role filler is not bound to a complement. It is existentially quantified. So the AVM is:

$$\begin{bmatrix} \textit{WORD} & \textit{bake} \\ \textit{SYN} & \begin{bmatrix} \textit{MAJ} & \textit{V} \\ \textit{SUBCAT} & \textit{(NP[NOM] SEM @1)} \end{bmatrix} \\ \textit{SEM} & \begin{bmatrix} \textit{RELN} & \textit{BAKE} \\ \textit{BAKER} & \textit{EX-Q} \\ \textit{BAKED} & \textit{@1} \end{bmatrix} \end{bmatrix}$$

Figure 4-6: AVM for *ergative bake*

For the *unspecified-object bake* in the sentence "John baked", the BAKED role has not been filled. So the AVM is:

$$\begin{bmatrix} \textit{WORD} & \textit{bake} \\ \textit{SYN} & \begin{bmatrix} \textit{MAJ} & \textit{V} \\ \textit{SUBCAT} & \textit{(NP[NOM] SEM @1)} \end{bmatrix} \\ \textit{SEM} & \begin{bmatrix} \textit{RELN} & \textit{BAKE} \\ \textit{BAKER} & \textit{@1} \\ \textit{BAKED} & \textit{EX-Q} \end{bmatrix} \end{bmatrix}$$

Figure 4-7: AVM for *unspecified-object bake*

By looking into the syntactic behaviors of some other verbs such as *melt, clear, remove, wipe* and *wash*, Kilgarriff found out that all the alternation-related information can be organized into a hierarchy. This is shown below (Kilgarriff, 1992) p. 116.

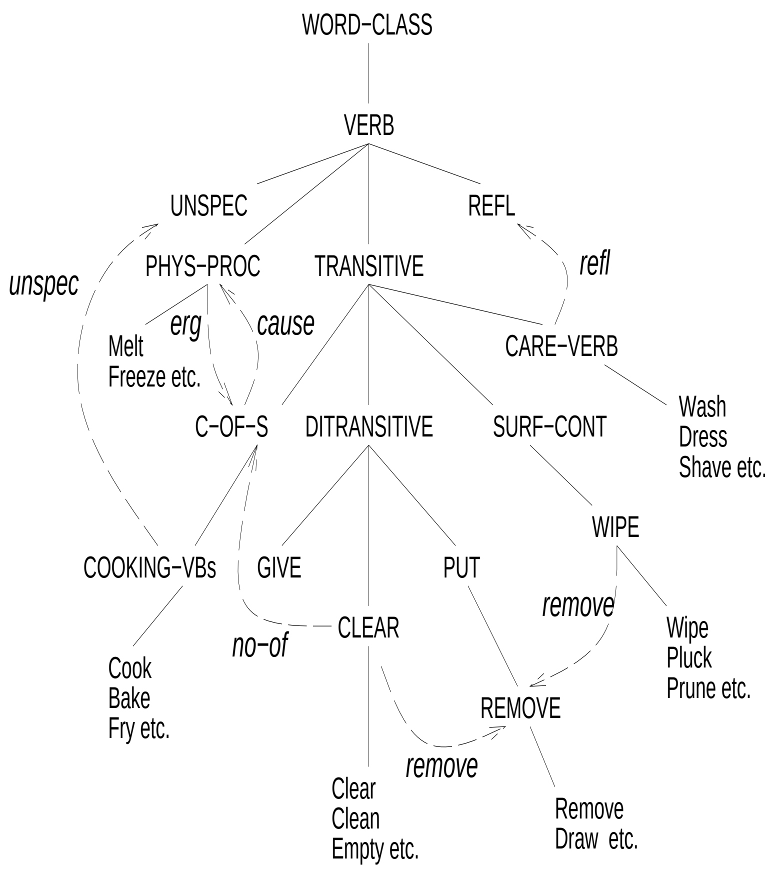



Figure 4-8: Kilgarriff's verb taxonomy

The figure shows that *bake* takes TRANSITIVE as its base form. The cooking verb class and the change-of-state verb class are the superclasses of *bake*. The alternation *unspecified-object* of *bake* can be inherited from the verb class UNSPEC. The "*ergative*" of word *bake* can be inherited from PHYS-PROC verb class by the route through COOKING-VB and C-OF-S verb classes.

Kilgarriff's work serves a good example of applying DATR to lexical knowledge hierarchical organization. On the other hand, he proposed a concrete example to show how Levin's English alternations and classes can be organized into a hierarchy. However, Kilgarriff did not go further to use the hierarchy to solve the semantic representation problem and handle the extension of a word meaning as his thesis implied. This problem will be dealt in the following chapters.

## 4.5 Summary

As we have seen above, the recent calls for a structured lexicon are limited to the issue of efficient storage of lexical syntactic information. The hierarchical organization of lexicon is for the purpose of inheritance only. Daelemans et al. have presented a good review on inheritance in natural language processing (Daelemans et al., 1992). They summarized the motivation as follows:

- Inheritance lexicons can be made one or two orders of magnitude smaller than their full-entry counterparts.
- Changes or corrections will typically only need to be made in one or two nodes, not in thousands of individual entries.
- Several levels of linguistic description can be encoded in the same way and be made subject to the same rules of inference.
- Multiple inheritance allows different taxonomies to apply for different levels of description.
- The relation that a lexical property at one level of description depends on a lexical property at another level of description can be stated.

However, as we will see later, the lexicon relations in the structured organization need not be restricted to the inheritance relation. And the structured lexicon organization can be used to define word meanings and their possible extensions. The structured lexicon will be able to handle all the usages of a word based on its current knowledge base.

# Chapter 5

## Literature Review II: Verb Semantic Representation

To achieve the goal of large coverage of the NLP system, we have to take care of the word meaning in each usage type. Besides the task of disambiguating word senses, a system must be able to handle unknown usages. If an unknown usage is totally new and without any relation to the existing knowledge in the system, naturally the system will not be able to know the word meaning. However, the system should be able to guess the word meaning at its best. Here the original word meaning representation plays a very important role in the inference. Unfortunately, studies in computational semantics have long been influenced by generative linguists who only worry about the "systematic semantic-syntactic correspondences" (Levin, 1987).

> Numerous arguments have advanced against the use of predicate decomposition, as in Fodor et al.'s paper "Against Definitions" (1980). Many of their arguments are inapplicable to the discussion of decomposition here. They assume that the decompositions are put to use other than that assumed here. In the works discussed, the decomposition of verbs is proposed for the purposes of accounting for systematic semantic-syntactic correspondences. ... Instead, Fodor et al.'s concern is whether the decomposition or definition replaces a lexical item whenever it is used. They are not interested in the independent question of whether a decomposition analysis as a lexical semantic representation enters into the statement of linguistic generalizations. (Levin, 1987) P. 39.

In generative linguistics, verb representations are built for linking between syntactic and semantic levels. But word meanings are more than that. Even only considering the semantic-syntactic correspondences, recent research has shown that word meanings are complicated and a deep analysis is needed. In the following, I will lead the reader on a tour of the history of semantic theories in the generative linguistic school. It is the most influential linguistic theory in computational natural language processing.

## 5.1 Fillmore: Case theory

Fillmore's case theory (Fillmore, 1980) has a pervasive influence on linguists and computer scientists. Fillmore's cases attempt to define the interrelations between syntax and semantics. The original idea was that case functions were expressed in the deep structure of a generative transformational grammar. For instance, the sentence "John gave the books to my brother" is analyzed below:

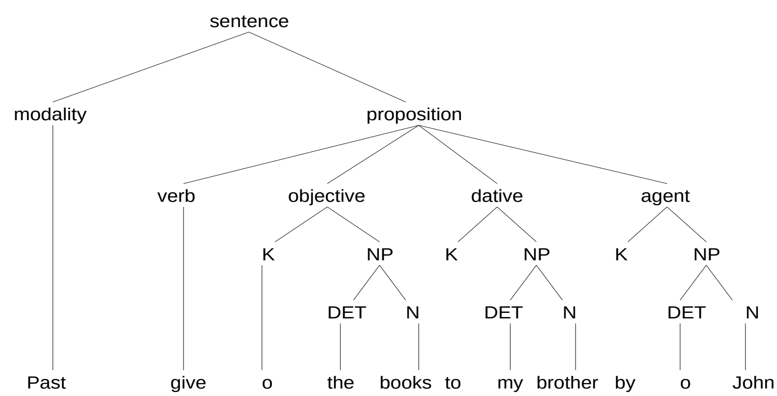



Figure 5-1: Case analysis for "John gave the books to my brother"

The cases used here are objective, dative, and agentive. The motivation behind this treatment is partially due to the evidence that different semantic roles can be indicated by the same syntactic constituent. For example, in the sentence "John gave the books to my brother", the "book" is the grammatical OBJECT, however, in the sentence "The book is given to my brother by John", "book" is the grammatical SUBJECT. That is why Fillmore rejected SUBJECT and OBJECT as the underlying universal syntactic-semantic relations and proposed a more abstract representation containing the case functions. Those cases were supposed to be universal to all languages and to provide common terminology for representing the deep structure of all verbs. Fillmore even listed a set of general rules for the syntactic linking of his cases in English. For instance, he proposed the following:

> "If there is an agentive, it becomes the SUBJECT; Otherwise, if there is an Instrumental, it becomes the SUBJECT; Otherwise, the SUBJECT is the Objective."

Other useful linking conditions were proposed by him as well, such as:

"No case can appear twice in the same clause.

Only noun phrases of the same case can be conjoined.

Each syntactic constituent can fill only one case."

The case theory with its generative tradition is naturally developed into the verb predicate decomposition theory (Jackendoff, 1972). By augmenting case frames with complex predicates, the deep structure of a sentence is no longer a case frame but rather a decomposition of the verb predicate into several embedded predicates. For example, a possible case frame for *kill*:

*kill(agent, patient)*

is changed into the semantic primitive representation:

*cause (<agent>, become(dead (<patient>)))*

Although the intuition that syntactic choices are largely a reflection of underlying semantic relationships is very appealing, the problem is that such an approach is difficult to define precisely and to implement satisfactorily. Computer scientists like Simmons (Simmons, 1973) who tried to adhere closely to a linguistic analysis of case get bogged down in modeling the necessary semantic representation and in accounting for all "exceptions" to the generalizations. The "exceptions" seem to form a larger set than the verbs that conform, with the result that most verbs seem to require individual sets of prepositions for indicating cases. Even only considering the syntactic-semantic correspondence, the following sections will show that cases are insufficient to capture the essence of linking between syntax and semantics.

## 5.2 Jackendoff: Thematic roles and Lexical Conceptual Structure

By recognizing the problem that case theory does not allow a noun phrase to be assigned more than one case, Jackendoff (Jackendoff, 1972) extended the case theory to thematic relations:

> The thematic relations can now be defined in terms of [these] semantic subfunctions. Agent is the argument of CAUSE that is an individual; Theme is the argument of CHANGE that is an individual; Source and Goal are the initial and final state arguments of CHANGE. Location will be defined in terms of a further semantic function BE that takes an individual (the Theme) and a state (the Location). (Jackendoff, 1972) p. 39

So in Jackendoff's representation for "Esau traded his birthright (to Jacob) for a mess of pottage", "Esau" fills the traditional AGENT position as well as the SOURCE position.

Jackendoff further extends the notion of "thematic relations" as the index to the argument positions in his lexical conceptual structure.

> The fundamental point, from which all else proceeds, is that *thematic roles are part of the level of conceptual structure, not part of syntax.* Recall Gruber's (1965) intuitive definition of Theme: the object in motion or being located. This can be structurally defined as the first argument of the functions in (lc,d). Source, "the object from which motion proceeds," appears structurally as the argument of the Path-function FROM. Note

> that Source is not a direct argument of the Event-function but is embedded within a Path-constituent. Similarly, Goal, "the object to which motion proceeds," is the argument of the Path-function TO. Agent is the first argument of Event-function CAUSE. Experiencer presumably is an argument of a yet unexplored State-function having to do with mental states. In other words, thematic roles are nothing but particular structural configurations in conceptual structure; the names for them are just convenient mnemonics for particularly prominent configurations. (Jackendoff, 1990) p. 46-47

In trying to identify the *I-concepts*[1], Jackendoff proposed his lexical conceptual structure(LCS) for verb representation. Verbs are decomposed into an embedding predicate. A small number of primitives such as PLACE, PATH, EVENT, STATE, CAUSE, GO etc. are used as the construction blocks for the composition. Although Jackendoff notices that only limited aspects of conceptual structure interact with syntax (Jackendoff, 1990) p. 49, the focus of the LCS is still on the level of syntactic-semantic correspondence, because the choice of those primitives and structure of the predicates have a generative tradition. Verb syntactic behaviors are the major evidences for justifying the approach. LCS deals more with how to derive an abstract representation for a sentence than on how to provide a good bridge between lexical meaning and the inference to the world knowledge. When it is used for NLP, Palmer's verb representation, Dorr's LCS representation need to be argumented. Later on, we will discuss some serious problems when we look into the UNITRAN system.

[1] I-concepts is Chmosky's notion about the concepts related to the internal language, the language as a body of internally encoded information

### 5.3 Levin: English syntactic alternation

Levin (Levin, 1977) pointed out many years ago that the case system adopts arbitrary categorizations because they cannot represent the complexity of events, and case names encode semantic concepts without explicitly defining their properties and interaction. In her recent work (Levin, 1992), she presents a detailed description of English verb classes and alternations. The main point of her work is to show that verbs in English and other languages fall into classes on the basis of shared components of meaning. The lexical knowledge of a speaker of a language must include knowledge of the meaning of individual verbs, the meaning components that determine the syntactic behavior of verbs, and the general principles that determine behavior from verb meaning. Her work shows that the syntactic behavior of the verb and the verb meaning are the inter-related two sides of the one coin. Once the relation is settled, by looking into the semantic components in a verb meaning, the verb's syntactic behavior can be predicted. On the other hand, by listing the syntactic behaviors of the verb, the exact meaning of the verb can be identified.

Levin used the following examples to show that various aspects of the syntactic behavior of verbs are tied to their meaning. Four verb classes *break, cut, hit, touch* is investigated. All four verbs are transitive:

(2) a. Margaret cut the bread.
    b. Janet broke the vase.
    c. Terry touched the cat.
    d. Carla hit the door.

However, only *cut* and *break* are found having the middle alternation.

(3) a. The bread cuts easily.
b. Crystal vases break easily.
c. * Cats touch easily.
d. * Door frames hit easily.

*cut* and *hit* appear in the conative construction, but *break* and *touch* do not.

(4) a. Margaret cut at the bread.
b. * Janet broke at the vase.
c. * Terry touched at the cat.
d. Carla hit at the door.

Another diathesis alternation – the Body-part Possessor Ascension alternation distinguishes *cut, hit, touch* from *break*.

(5) a. Margaret cut Bill on the arm.
b. * Janet broke Bill on the finger.
c. Terry touched Bill on the shoulder.
d. Carla hit Bill on the back.

It has been found that each verb here represents a class of verbs sharing the same syntactic patterns listed above. They are:

(6) a. Break Verbs: break, crack, rip, shatter, snap. ...
b. Cut Verbs: cut, hack, saw, scratch, slash, ...
c. Touch Verbs: pat, stroke, tickle, touch, ...
d. Hit Verbs: bash, hit, kick, pound, tap, whack, ...

Studies also show that the members of the verb class share certain aspects of meaning. Thus, the members of the class have common syntactic and semantic properties. As we know, the four verbs differ as follows: *touch* is a pure verb of contact, *hit* is a verb of contact by motion, *cut* is a verb of causing a change of state by moving something into contact with the entity that change state, and *break* is a pure verb of change of state. At the same time as we can see, the members in the classes also possess the same semantic components. The inter-related property of these verb classes is shown in the following table.

| Properties | *touch* | *hit* | *cut* | *break* |
|---|---|---|---|---|
| Conative: | No | Yes | Yes | No |
| Motion and Contact: | No | Yes | Yes | No |
| Body-part Possessor Ascension: | Yes | Yes | Yes | No |
| Contact: | Yes | Yes | Yes | No |
| Middle: | No | No | Yes | Yes |
| Causing a change of state: | No | No | Yes | Yes |

These inter-related properties are very important. It shows that when considering the linking between syntax and semantics, those highly abstract notions of cases or thematic roles are not fine-tuned enough to sort the syntactic behaviors clearly. A deeper semantic analysis is needed. The notions of motion, contact, change of state, and causation have firm connections with the verb behavior which must be taken into account in selecting a lexical representation of verb meaning. Although the inter-related properties between syntax and semantics have been known by linguists for ages, Levin's main contribution is that she is the first person to really look into the class of English verbs as a whole and attempt to present a coherent picture of these properties.

## 5.4 Pinker: Language acquisition of children

Child language acquisition studies have contributed many insights to the problem of verb semantics and representation. One of the major concerns in this area is how children acquire argument structure. This problem poses a learnability paradox (Pinker, 1991). The so called "Baker's paradox" states that: without benefit of negative evidence, children learn a grammar in which lexical rules allow productive generalizations of many verbs to new argument structures, while excluding other verbs that are otherwise syntactically indistinguishable. Recent studies have impact on verb representation from different perspectives.

### 5.4.1 Pinker: Grammatically Relevant Subsystem Hypothesis

Pinker proposes the following hypotheses to solve Baker's paradox (Pinker, 1991).

**Criteria-governed productivity** Children and adults both can use lexical rules productively. By noticing that a verb has a similar meaning component to a verb class such as the "break" class, they can predict that the verb has the same syntactic alternations as the "break" class. However, the semantic and morphological criteria, i.e., some fine-tuned semantic distinctions, will prevent the production.

**Thematic cores** Argument structure is the projection of the semantic structure. When a verb changes from one syntactic alternation to another, the semantic structure is changed. And this change will be more suitable for some verbs than others. Thus the production will be more natural for some verbs than for others.

**Narrow conflation classes** Verbs fall into numerous narrow classes, they have similar meanings and forms. This is consistent with Levin's work (Levin, 1992).

**Grammatically relevant subsystem** Verb meaning contains two parts according to whether or not they are relevant to the grammar. The grammatically relevant part can be represented by a very small number of primitives. We will focus on this later.

**Color-blind conservatism** The narrow class formation and narrow range rules are learned by setting a variable for some idiosyncratic pieces of information in verb meaning. The variable is free to vary across the members of the class as long as it is specified in some way.

**Minimalist hypothesis and childhood malapropisms** Children's lexical rules at all stages are formally similar to those of adults. The main developmental mechanism that makes them sound more like an adult as they grow up is the acquisition of more and more accurate meanings for more and more verbs.

The grammatically relevant subsystem hypothesis is close to Jackendoff's lexical conceptual structure (Jackendoff, 1990). What Pinker has claimed is:

> Perhaps there is a set of semantic elements and relations that is much smaller than the set of cognitively available and culturally salient distinctions, and verb meanings are organized around them. Linguistic processes, including the productive lexical rules that extend verbs to new argument structures, would be sensitive only to parts of semantic representations whose elements are members of this set. The set would consist of symbols that have cognitive content, such as "causation" and "location". but not all cognitively meaningful concepts are members of this privileged semantic machinery. Thus a verb like *to butter* would specify information about butter and information about causation, but only the causation part could trigger or block the application of lexical rules or other linguistic processes. (Pinker, 1991) P. 166

The claim has clearly specified the goal of the semantic representation in (Pinker, 1991). The representation is designed for the purpose of syntactic-semantic correspondence. It pays no attention to the *grammatically irrelevant information* in the lexicon. That is why they are not concerned about whether *cause to die* is no longer the same as *kill*, as long as the representation can sort the syntactic-semantic correspondence clearly. Similar to Levin, Pinker implies that by exhaustively investigating verb syntactic behaviors, the grammatically relevant portion of elements for building verb entries can be identified. However, the problem how to represent the grammatically irrelevant portion of the verb lexicon still remains unsolved.

### 5.4.2 Gropen: Affectedness and direct objects

Again, from the point of view of syntactic-semantic correspondence, Gropen et al. (Gropen et al., 1991) designed experiments for children and adults to test their willingness to encode the moving items or the surface as the verb's object. Evidence shows that speakers are not confined to labeling moving entities as *themes* or *patients* and linking them to the grammatical object. When a stationary entity undergoes a state change as the result of a motion, it can be represented as the main affected argument and thereby linked to a grammatical object instead.

Gropen et al. explicitly denounce the traditional theories about linking. The early theories posit that syntactically relevant information about semantic arguments consists of a list of thematic roles like *agent*, *theme*, and *goal*, which are linked onto a hierarchy of grammatical positions like subject, object and oblique object. For verbs involving motion, the entity caused to move is defined as the *theme* or *patient* and linked to the object. But these terms are too abstract to cover all possible linkings. This prompts the recent theories about more deep representation, i.e., verbs' meanings are multidimensional structures in which motions, changes, and other events can be represented in separate but connected substructures; linking rules are sensitive to the position of an argument in a particular configuration.

The main argument they have about a verb's direct object is that the verb's object would not be linked to the moving entity but to the argument specified as "affected" or "caused to change" as the main event in the verb's meaning. The change can either be one of location, resulting from motion in a particular manner, or of state, resulting from accommodating or reacting to a substance. So they conclude:

> Within the semantic structure theory, these generalizations are enabled not by a single lexical rule specifying the syntactic linking of the new form (as in the list-of-primitives theory), but by a combination of a specific lexical rule and general linking rules. The lexical rule is reduced to a simple manipulation of a verb's semantic structure, effecting a gestalt shift: the rule takes a semantic representation as "cause X to go into/onto Y" and generates a new, related representation like "cause Y to change by means of cause X to be in/on Y" (or vice versa). The linking rules would create the corresponding syntactic structures automatically. (Gropen et al., 1991) P. 182.

Three experiments about children's using language have been presented in (Gropen et al., 1991). Interested readers can refer to it for more detail.

## 5.5 Recent theories about deep representation

By facing problems raised by the list-of-primitives approach, researchers feel the need for a deep representation of verb meaning. Since the motivation comes from the consideration of syntactic-semantic correspondence, except Siskind (Siskind, 1992), the grammatically irrelevant parts of the verb meaning have not been considered.

### 5.5.1 Dowty: Proto-roles

Vexed by the questions of the theoretical status of thematic roles and the inventory of possible roles, Dowty (Dowty, 1991) has proposed a new theory about thematic roles and addressed the problem of argument selection. His major claims are:

> It is concluded that the best theory for describing this domain is not a traditional system of discrete roles (Agent, patient, Source, etc.) but a theory in which the only roles are two cluster-concepts calls Proto-Agent and Proto-Patient. Each characterized by a set of verbal entailments, an argument of a verb may bear either of the two proto-roles (or both to varying degrees, according to the number of entailments of each kind the verb gives it. Both fine-grained and coarse-grained classes of verbal arguments (corresponding to traditional thematic roles and other classes as well) follow automatically, as do desired 'role hierarchies'. (Dowty, 1991) P. 1.

He further proposed the contributing properties for the agent proto-role and the patient proto-role:

(27) Contributing properties for the Agent proto-Role:

a. volitional involvement in the event or state

b. sentence (and/or perception)

c. causing an event or change of state in another participant

d. movement (relative to the position of another participant)

(e. exist independently of the event named by the verb)

(28) Contributing properties for the patient proto-Role:

a. undergoes change of state

b. incremental theme

c. causally affected by another participant

d. stationary relative to movement of another participant

(e. does not exist independently of the event. or not at all)

Although Dowty's listed properties for proto-roles are not the same as the studies in children's language acquisition approaches (Gropen et al., 1991) (Choi and Bowerman, 1991), the basic motivation behind his proto-Roles theory is the same as (Pinker, 1991) (Levin, 1992). They all want to use more fine-tuned notions or finer categorization of rules to achieve certain distinctions in the linking.

### 5.5.2 Jackendoff: Parts and boundaries

By addressing a wide range of phenomena at the level of syntax and morphology, Jackendoff (Jackendoff, 1991) proposed some primitives that differ from those STATE, EVENT, and PATH that he proposed earier in his lexical conceptual structure. These primitives includes the features $\pm b$ and $\pm i$, the six extracting and including functions (PL, ELT, COMP, GR, PART, and CONT), the dimensionality feature (including the epsilon dimensionality), the directionality feature, and the two boundary function BD and BDBY. These primitives are used to handle phenomena like the plural, collective nouns, N-of-NP constructions and N-N compounds, boundary nouns like *end* and *crust* and prepositions like *to* and *form*, progressive aspect, and the "imperfective paradox".

However, Jackendoff has not solved the problems raised by Fodor (Fodor, 1970) (Fodor et al., 1980). First, in the way of reducing *kill* to *cause to die*, something is missing. It may be enough to use *cause to die* to represent *kill*, if only syntactic-semantic correspondence is considered. But when considering how to draw conclusions based on the knowledge of *kill*, it is obviously not enough. Second, there is no principal reason to stop the decomposition at any given level of detail. Third, there is no reason to prefer one decomposition at a given level of detail over another. Again, if only syntactic-semantic correspondence is considered, verbs' syntactic behaviors can be the justification, but for reasons based on the representation, there is no such justification.

### 5.5.3 Pustejovsky: Event structure

Pustejovsky's work (Pustejovsky, 1991b) is another attempt to use the syntactic-semantic correspondence to justify a deep semantic representation theory. The motivation for his event structure theory is that the event structure can provide a distinct and useful level of representation for linguistic analysis involving the aspectual properties of verbs, adverbial scope, the role of argument structure, and the mapping from the lexicon to syntax.

The major difference of the theory from Jackendoff's LCS is Pustejovsky's generative lexicon idea. He states that:

> We would like to suggest that lexical (and conceptual) decomposition is possible if it is performed *generatively*. Rather than assuming a fixed set of *primitives*, let us assume a fixed number of *generative devices* that can be seen as constructing semantic expressions. Just as a formal language is described in terms of the productions in the grammar rather than its accompanying vocabulary, a semantic language should be defined by the rules generating the structures for expressions rather than the vocabulary of primitives itself. (Pustejovsky, 1991b) P. 54.

However, this is consistent with Jackendoff's view of conceptual semantics:

> The conceptual structure is autonomous from language and that there is no intervening level of "purely linguistic semantics" intervening between it and syntax. (Jackendoff, 1991). p. 43.

The generative properties of the verb meaning have serious limitations. We will discuss this problem later in chapter 6.

### 5.5.4 Siskind: Naive physics

We finally come to the work of Siskind that is a call for deep verb semantic representation with motivation from the visual event perception domain. Siskind (Siskind, 1992) proposed a new definition for *throw*.

```
(define throw (x y)
 (exists (i j)
  (and (during i (move (hand x)))
       (during i (move y))
       (during i (contact (hand x) y))
       (during i (attached (hand x) y))
       (during j (not (contact (hand x) y)))
       (during j (not (attached (hand x) y)))
       (during j (move y))
       (during j (not (supported y)))
       (= (end i) (beginning j)))))
```

Siskind explains the definition as follows:

> Informally, this states a throwing event comprises two consecutive time intervals i and j, where during i, both x's hand and y are moving, and x's hand is in contact with and attached to y, while during j, x's hand is no longer in contact with and attached to y, and y is in unsupported motion. Note that this definition incorporates the grasping and releasing action of the agent followed by the supported motion of the patient, aspects of throwing not captured by the definitions advanced by Miller, Schank, Jackendoff, and Pinker. (Siskind, 1992) P. 108.

Siskind's approach attempts to formalize the *grammatically irrelevant parts* of verbs' meanings in the domain of visual perception. Although he pointed out the need for deep and detailed representation for verb events, he has not proposed a clear picture in this direction. The verb representations proposed in the thesis are only for the purpose of "using these definitions as part of an implemented computer program to accurately differentiate occurrences from non-occurrences of the events they describe in animated movies". (Siskind, 1992) p. 108. This makes his approach to lexical semantics lack generalities as a serious verb semantic theory.

## 5.6 Summary

We have seen the verb semantics theories developed so far in generative frameworks. These form the theoretical basis for most of the current natural language processing systems. Most of the theories focus on the syntactic-semantic correspondence. The whole history of development is shown as follows:

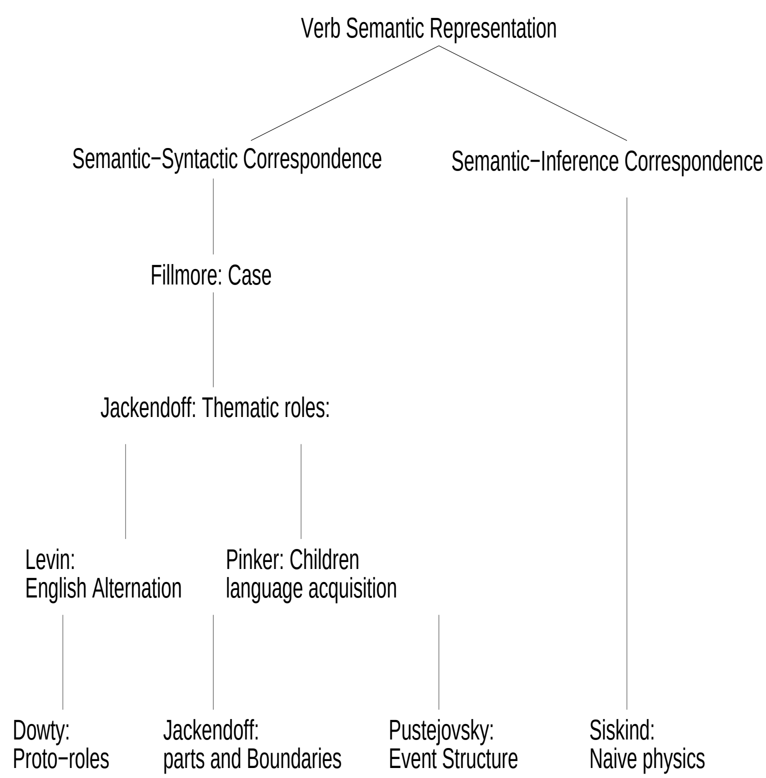



Figure 5-2: History of verb semantic theories in generative frameworks

Even only considering the linking between syntax and semantics, recent work shows that a deep detailed representation is needed. We call the early theories **list-of-primitives theory** and recent theories **semantic structure theory**. Gropen et al. (Gropen et al., 1991) have given a comprehensive comparison of these two main approaches that is summarized below. (Gropen et al., 1991) p. 155-163.

The problems of list-of-primitives theories are:

1. The early theories predict that all verbs denoting a kind of event with a given set of participant types should display the same linking pattern, and that is not true.

2. Verbs that violate the standard linking pattern would be non-canonical or "marked" and presumably would be rarer in the language and harder to learn. Not only does this reduce the predictive power of the theory, but its predictions do not seem to be true.

3. The list-of-primitives theory does not naturally explain systematic semantic differences between two forms of an alternating verb that involve the same kinds of thematic roles but different linking patterns.

4. The verb's semantic information relevant to linking should be exhaustively captured in its list of thematic roles, but the patterns of alternation (i.e., alternative linking patterns for one verb) are among verbs with identical lists of thematic roles.

The advantages of semantic structure theories are:

1. By going deep to represent the verb meaning, the actual roles are more finely differentiated and the verb's interaction with syntax can be sensitive to such distinctions. Thematic roles are not primitive types but are argument positions in the multidimensional structures.

2. The semantic structure theory in its strongest form holds that the linking pattern of a verb is fully predictable from its meaning.

3. Semantic structure theories predict which syntactic forms are related in an alternation, and how the verb's interpretation changes when it is linked to one form or another.

4. The new linking theory can be applied to a variety of constructions in a variety of languages.

5. The theory helps explain which verbs undergo alternations.

For computer text processing, the list-of-primitives theory has had great influence on the system design. With a few thematic roles, a small number of primitives and a neat predicate decomposition, this verb representation scheme strongly attracts computer scientists, because it meets the needs of scientific computing (i.e., neat theory with little tedious work) quite well. However, the computational linguists' dreams were quickly shattered, when these systems were required to handle unrestricted texts. The brittleness and small coverage of these systems all call for a re-evaluation of the representation theory on which the systems were based. Actually, this should be no surprise: the starting point of case theory has already restricted the representation within the area of syntactic-semantic correspondence. When the system is required for large coverage, with the lack of a deep and detailed representation of the verb meaning, the system goal of processing unrestricted text is doomed to fail. It is time for the NLP researchers to adopt the recent theories about semantic structures to deal with the syntactic-semantic correspondence. The other important aspect is to investigate the representation method for the "grammatically irrelevant" parts of verb meaning. Only when the verb representation is completed, can the goal of large coverage be reached.

# Chapter 6

## Theory I: The correspondence between sense and interpretation

The blind men of Hindustan attempt to comprehend the elephant:

*The first approached the elephant,*
*And happening to fall*
*Against his broad and study side,*
*At once began to bawl,*
*"Bless me, it seems the elephant*
*Is very like a wall."*

*The second, feeling of his tusk,*
*Cried, "Ho! What have we here*
*So very round and smooth and sharp?*
*To me 'tis mighty clear*
*This wonder of an elephant*
*Is very like a spear. "*

*The third approached the animal,*
*And happening to take*
*The squirming truck within his hands,*
*Then boldly up and spake;*
*"I see," quoth he, "the elephant*
*Is very like a snake."*

*The fourth stretched out his eager hand*
*And felt about the knee,*

*"What most this mighty beast is like*
*Is mighty plain," quoth he;*
*"'Tis clear enough the elephant*
*Is very like a tree."*

*The fifth who chanced to touch the ear*
*Said, " Even the blindest man*
*Can tell what this resembles most;*
*Deny the fact who can,*
*This marvel of an elephant*
*Is very like a fan.*

*The sixth no sooner had begun*
*About the beast to grope*
*Than, seizing on the swinging tail*
*That fell within his scope,*
*"I see," cried he, "the elephant*
*Is very like a rope."*

– *John Godfrey Saxe*

The problem of these blind men is that they only perceived one facet of the elephant. A verb meaning is just like an elephant. Besides the meaning components related to syntax, verb meaning has other components related to the interpretation. We should not ignore these components. As reviewed in Chapter 5, the verb semantic representation theory has long been influenced by syntactic theories. Word meaning is only decomposed for the purpose of semantic-syntactic correspondence. In this chapter, we will propose that word meaning, on the other hand, must be decomposed for the purpose of semantic-interpretation correspondence. As we have seen in Chapter 3, the translation accuracy depends on the verb semantic representation. The ability to handle new usages in the source verb analysis and the ability to measure the meaning similarity in the target verb realization are important for correct lexical selection. So it is necessary to investigate the definition and representation of verb senses for the purpose of lexical selection. In this chapter, we will discuss the following problems: Then what is a verb sense? What is important for handling new usages? And what is the correspondence between semantics and interpretation? We will then discuss verb semantic representations in several NLP systems. We will also try to show that the representation schemes in these systems are unable to handle the problem of linguistic incompatibilities and new usages well. This is because the verb semantic representations in these systems do not concern themselves with the correspondence between the semantics and interpretation.

## 6.1 Verb senses

### 6.1.1 Verb senses as usage types

First let us introduce a term *verb usage*. Kilgarriff has given a good description about the relation of usage of a word and the word senses (Kilgarriff, 1992).

> A 'usage' will be used to mean a particular occurrence of a word in a particular context. If we have two different sentences containing a word, or even the same sentence on two different occasions, we have two usages. Whereas usages are tokens, 'usage-types' or 'uses' are types. Thus wherever we might wish to say, in two usages, a word is being used in the same way, we say both usages exemplify a particular usage-type. Any set of usages of a word where all the members have some aspect of meaning in common, so there is some motivation for saying they all mean the same thing, is a usage type. 'Senses' or 'word senses' are a subset of usage-types. 'Senses' are those usage-types that are or ought

to be listed in a dictionary.

Normally, every particular usage carries particular information. So if we take the verb sense as the information it intends to carry, to the ultimate extreme, a verb in every usage has a particular verb sense. However, it is impossible to define the verb sense for each of the verb usages. We have to generalize on different usages and define the verb senses based on a group of similar usages. For example, the verb 'break' in a concrete object breaking event might have different degrees and styles of breaking for different concrete objects. However, we do not have to give a definition to each of these broken concrete objects. We can define the verb break as 'causing a concrete object to separate into two or more parts' that covers a class of break usages with concrete objects. So the verb sense definition problem becomes a usage classification problem. The number of verb senses is the number of classes of usages. We can also call the class of usages usage types. Verb senses are the conceptual representation of these usage types.

### 6.1.2 Problems in usage type classification

The verb sense is a meaning representation of a group of different usages, in the verb sense definition and representation, there are three problems.

#### Problem 1: How to classify the different verb usages into classes?

Taking all possible usages as one usage type and taking each different usage as a different usage type are the two extremes of the classification. The first one makes sense when all the usages have the same meaning. But when there exists usages that have very different meanings, the class has to be split. The second classification is neither possible nor necessary. Then what are the criteron for classifying the usages? From the system implementation point of view, different usages will trigger different sets of actions. It is natural to group those usages that trigger the same action into one class so that they may have the same meaning representation.

**Problem 2: How to define a usage type?**

Once a classification is done, given a new input, the system should be able to judge which class the new input belongs to. Normally, a set of conditions is given to test the new input. These conditions draw boundary lines between different usage classes. If these conditions are clear-cut conditions, the devision of all the usages will be rigid. For a new usage that is out of the boundary, it is difficult to measure the similarity between the meaning of the new usage and the existing sense. If the conditions are not clear-cut conditions, a similarity degree can be derived from some testing. So it is better to provide some numerical measurement on those conditions.

**Problem 3: How to represent the meaning of a usage type?**

The other important problem is to represent the meaning of a usage type. Different usages carry different meanings. If they are in the same class, their meaning must be similar. A natural way is to generalize on all the usages in the class, and use a generalized concept to represent all the usage meanings. For instance, the usages 'the man broke the branch' and 'the man broke the window' can be classified into the same class, where a generalized concept 'the physical object is separated' can be used as the representation of this class. However, when no extra information is given, in the interpretation of these two sentences, common sense knowledge tells us that 'the man broke the branch' might mean 'the branch is separated somewhere along the line segment shape', 'the man broke the window' might mean 'the window is separated into pieces'. Because of the generalization, such slight conceptual differences have been ignored. There are gaps between the semantic representation and the interpretation. In the next section, we will look into this problem in detail. Now let us look at some different methods of sense classification.

### 6.1.3 Ways of different classifications

#### Dictionary sense classification

There are different methods of usage classifications. One method is the classification by lexicographers, i.e., the dictionary senses. We can view the dictionary senses as the classification of the verb usages by human understanding. Because different people have different understanding and classification of verb senses, different dictionaries have different sense definitions. The definitions are clear-cut divisions of the usage types. The meaning representation is a paraphrase explanation of the verb sense. Since a human has common sense knowledge and the ability of analogical reasoning, there is no difficulty for a human being to understand the dictionary entries.

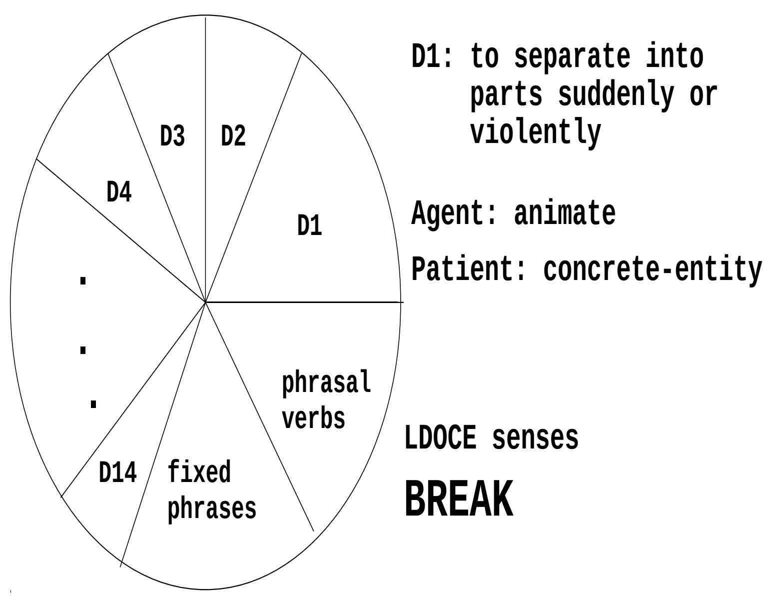



Figure 6-1: Dictionary sense definitions

### Translation sense classification

The other method of classification is to classify a verb sense according to its translation to a target language. The verbs in those usages have the same meaning when they are translated into the same target verb. This type of definition was also used in the transfer-based approaches. The translation pairs are exhaustively listed in the bilingual dictionary. The selection restrictions on the verb arguments are used to select different target verbs.

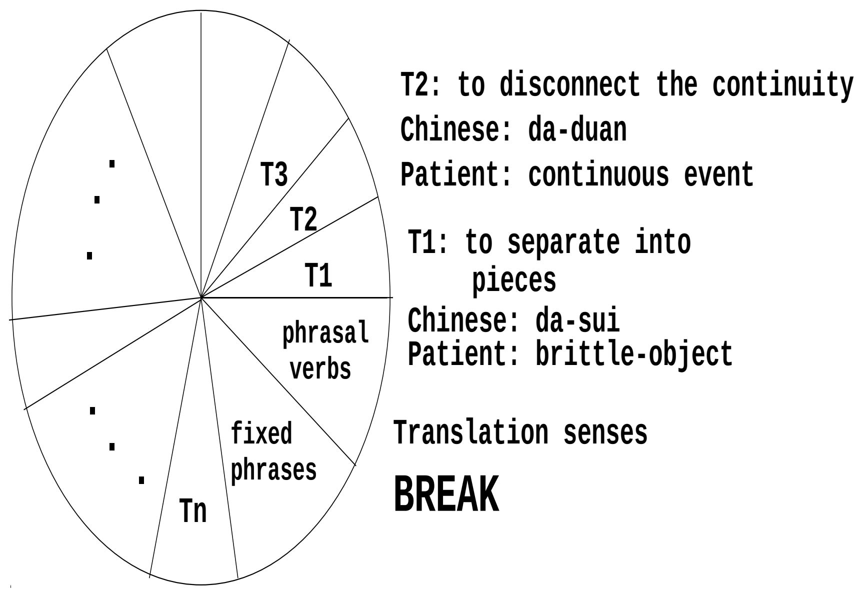



Figure 6-2: Translation sense definitions

### Fuzzy usage classification

The third method is to have a fuzzy classification of the usages. It is obvious that an exhaustive listing is infeasible. So it is necessary to define a verb sense according to fuzzy categories and provide ways to construct the interpretation of the particular usages in a usage type, and predict the new usage meanings in the machine systems. Therefore, we have the following figure:

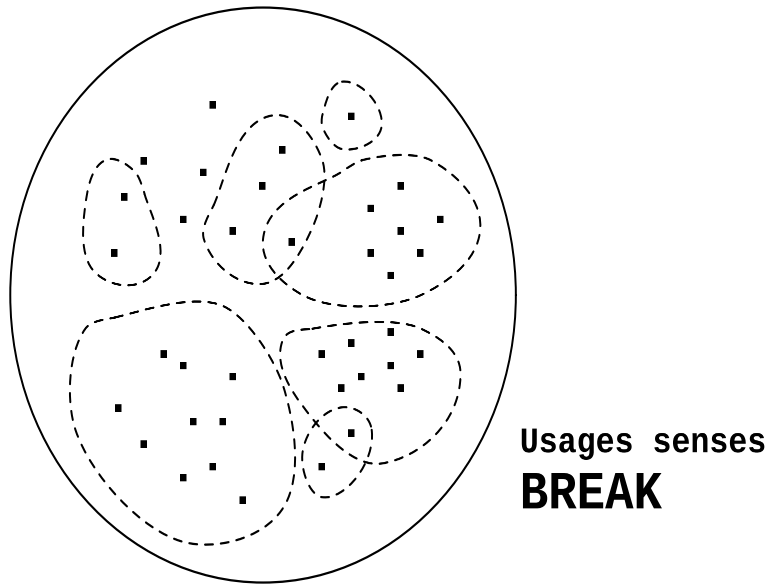



Figure 6-3: Usage sense definitions

## 6.2 Verb sense and interpretation

In the examples of 'the man broke the branch' and 'the man broke the window', we know that they have different interpretations. This is because common sense knowledge takes part in the interpretation process. This becomes a gap between the verb sense and the interpretation. There are different points of views for this problem. One is the *encyclopedic* view of linguistic semantics that proposes the storing of all related information in the lexicon entry. The other is the *self-contained* linguistic semantics view that supports separation of semantic knowledge and knowledge needed for interpretation.

From the *encyclopedic* view of linguistic semantics, each usage's context will not have exactly the same meaning. The word entry should contain all the knowledge that makes the usages different. An NLP system could wholly depend on the knowledge in the word entry. If the *encyclopedic* conception of linguistic semantics prevails, the word entry will be very large or even open ended. This is certainly not desired by computer scientists. It is better to hold the view of *self-*

*contained* linguistic semantics. This means that a boundary line must be drawn between the semantic knowledge and the encyclopedia knowledge. The system inferencing is based on the lexical entry and the domain knowledge together. The lexical entry is a general one that is designed to be highly abstract and easily connected with the domain knowledge. This kind of organization should not imply a lower system performance than the encyclopedia one. It should achieve the same system performance while the system becomes more efficient and portable. In order to achieve this goal, the following conditions must be met:

- Only the core senses are stored in the verb lexical entry.
- The representation of each verb sense should be designed in a way that each usage's meaning can be easily predicted.
- The relations between the core senses of the verb and the meanings in the usage context must be identified. We can call the rules for deriving the particular meaning from the core sense as the **linking rules between semantics and interpretation**.

If we hold the view of *self-contained* linguistic semantics, the correspondence between sense and interpretation must be considered. The representation scheme should be easily extended to construct interpretation representations.

The *semantic-interpretation correspondence* has long been ignored by the NLP community. As it is described in Chapter 3, the verb semantic meanings are normally represented by a paraphrase. They are not decomposed into semantic components. This is obviously not enough for the system to get a correct result for each usage of the verbs. Being similar to the decomposition of verb meaning for the purpose of the semantic-syntactic correspondence, the verb meaning here must be decomposed for the purpose of semantic-interpretation correspondence. *Systematic semantic-interpretation correspondence* can serve as a criterion for us to decide how to decompose the verb into semantic components, how to figure out the linking rules between semantic core meaning and the usage meaning. It is time for us to consider the **systemic semantic-interpretation correspondence**.

## 6.3 Problems in some practical systems

In this section, we will see that because the semantic-interpretation correspondence is ignored, practical systems face difficulties in achieving the goal of large system coverage.

### 6.3.1 UNITRAN

UNITRAN is a Principle Based Machine Translation (PBMT) prototype developed at MIT by Bonnie Dorr (Dorr, 1990). The system can translate single sentences among English, Spanish and German. UNITRAN elaborates the idea of PBMT to its full strength. At each level of morphological, syntactic and semantic processing, the system is designed based on a small set of principles. Particular languages realized in the system by a set of parameter setting files in the system. This approach brings in some merits. First, it is easy to extend the system's ability to handle a new language. By specifying the parameter files, the major parts of the system remain unchanged. Secondly, different languages have divergences in syntax, semantics and pragmatics levels. Since the divergences of the languages can be represented by different parameters for the same principle, language divergences can be easily resolved.

#### UNITRAN verb representation overview

UNITRAN employed LCS (Lexical Conceptual Structure) as its lexical semantic representation. LCS is a compositive representation method for lexical semantics. The building blocks can be classified into several types as: EVENT, STATE, PO-

SITION, PATH, THING, PROPERTY, LOCATION, TIME and MANNER. For each type, there is a set of semantic primitives. For example, we have HERE, THERE, LEFT, RIGHT, UP, DOWN ... in the type LOCATION. Some of the primitives are predicates that take arguments. According to Jackendoff's theory (Jackendoff, 1990), every sentence meaning or word meaning can be represented by primitives or the composition of primitives. The composition is done by observing the $\theta$ theory. The following is an example of Chinese "划伤 (HuaShang, Stab)" event.

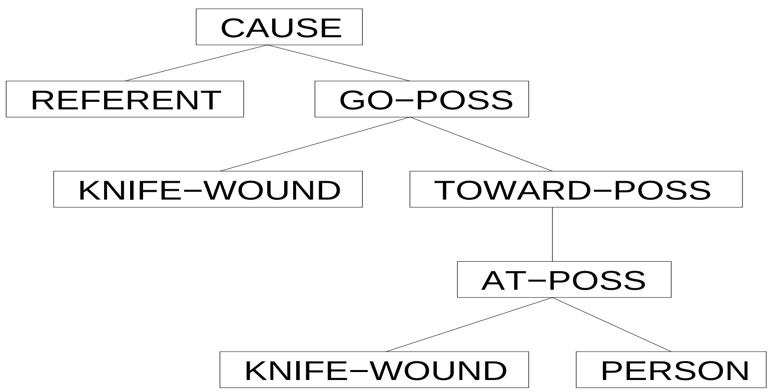



Figure 6-4: The underlying LCS for Chinese verb HuaShang

This is the interlingua that is used as the pivot language from source to target language. The underlying form conveys the meaning that a *referent* causes a *person* to have a *knife-wound*. For the Chinese sentence "小明 跑步 上 学 (XiaoMing PaoBu Shang Xue, XiaoMing jogged to the school)", the logical form of the whole sentence is derived from the verb semantic representation of "跑步 (PaoBu, Jog)". The LCS representation i.e. the argument structure of "跑步 (PaoBu, Jog)" is:

(DEF-ROOT-WORDS (GO-LOC Y (FROM-LOC (AT-LOC Y Z1)) (TO-LOC (AT-LOC Y Z2)))

:ROOTS ((跑步 (Y (* Y))

(Z1 :OPTIONAL ((* FROM-LOC) (AT-LOC (Y) (Z1))))

(Z2 (UC (CASE ACC)) ((* TO-LOC) (AT-LOC (Y) (Z2))))
(MODIFIER JOGGINGLY))

This representation defines "跑步 (PaoBu)" as falling into the class of GO-LOC. GO-LOC is a three-place predicate that represents "motion with manner". The definition of "跑步 (PaoBu)" can be read as "Y is in a motion from location Z1 to a location Z2 with a 'JOGGINGLY' manner". Without arguing about the semantic representation schemes, let us see how the surface structure "小明 跑步 上 学 (XiaoMing PaoBu Shang Xue)" can be analyzed to form a Logic LCS representation. The composed LCS of the sentence is shown in Figure 6-5.

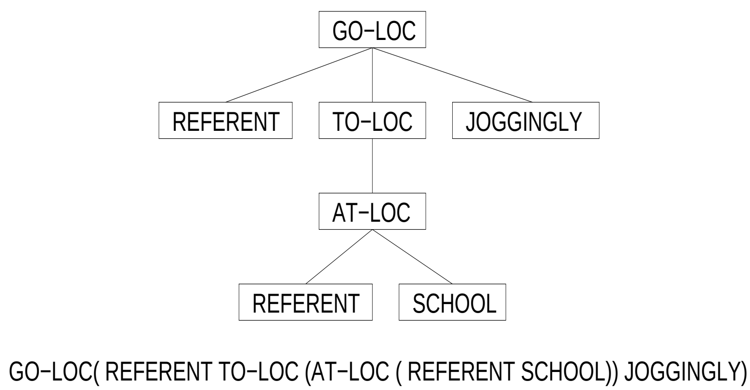



Figure 6-5: Logic form i.e. LCS representation

### Limitations of the verb representation

Taking LCS as the underlying verb representation scheme has resulted in UNITRAN some serious limitations (Dorr, 1990).

1. LCS does not provide a good bridge from the verb semantics to discourse and the domain knowledge. This makes it difficult to process "external pronominal reference", "paraphrasing", "story telling" and "interactive question-answering".
2. Because LCS primitives are too abstract, many open-ended classes of words are not distinguishable by their LCS's. This makes the lexical selection pro-

cedure unable to cut down the surface word possibilities during the mapping from the composed LCS to lexical items.

3. The similarities among words are difficult to measure. The system requires an exact match of the target-language root words to the underlying conceptual structure. The preference-assigning scheme is difficult to encode into the system. This may have been enhanced in the later approach.
4. The notion of *aspect* and *tense* is not represented in the LCS structures. But in some languages, these notions might be lexicalized. Lexical selections might depend on this information. For instance, there is no way in LCS to distinguish whether an event is prolonged, repeated or instantaneous. In Spanish, the surface realization requires on this missing information. This information should be at somewhere, either in the lexical entry or in the knowledge base. The translation of the repetitive version of *stab* is the surface form *dar puñaladas* (the plural form of knife-wound), whereas the translation of the non-repetitive version is the surface form *dar una puñalada* (the singular form of knife-wound).

#### Reasons behind these limitations

As we can see, all these UNITRAN limitations are due to the underlying LCS representation. In other words, the unsolvable problems behind the LCS style of decomposition are the key reason for limiting the system performance. It seems that the effort of breaking the limitation is doomed to fail if the LCS is used constantly as the underlying verb representation scheme. This is due to the fact that LCS is not fine enough to capture the conceptual distinctions. Now, let us see what those problems are.

- The primitives are chosen arbitrarily. They are separate from human cognitive ability. There are no criteria to justify which one should be a primitive and which one should not.

- The primitives are not fine-tuned enough. This makes the composed LCS representation unable to represent the meaning of the whole sentence.
- There are no criteria to justify in which way the verb is to be decomposed and whether the primitives should be decomposed more deeply, and when to stop the decomposition.
- LCS does not provide a good bridge between the primitives and the domain knowledge.
- LCS does not have a clear distinction between the grammatically relevant verb meaning components and the rest of the meaning. This makes the justification of the decomposition more vague. Sometimes, the syntactic behaviors are used to justify the decomposition. Sometimes, other facts are used.

In a word, the fundamental problem of LCS are the decompositions. There is no systematic way to justify the decompositions.

### 6.3.2 PUNDIT

PUNDIT is one of the state of art systems in the late eighties. The semantic processing in PUNDIT is Palmer's inference-driven semantic analysis (Palmer, 1990b). In PUNDIT, the verb is decomposed in a PROLOG format. The following is a decomposition of the verb *replace*.

```
replace :-
    cause (agent (A),
        use (instrument (I),
            exchange(object1(O1), object2(O2))))
```

This representation together with **Mapping Rules**, **Semantic class constraints** and **Common sense inference** forms the lexical semantic representation for PUNDIT. Although the representation is successful for very specified domains, such as "maintenance reports", "CASREPS", "TRIDENT" and "OPREPS", the customization of those representations from one domain to another gives rise to a big problem (Palmer, 1990a).

**Problem of representation in PUNDIT**

Martha Palmer is one of the pioneers handling the problem between the verb representation and inference. With the experience of PUNDIT, she found out the verb definitions change across domains in at least the following aspects:

- Change in thematic roles
- Change in selectional restrictions
- Change in semantic predicates
- Change in allowable common sense inferences

Palmer presented examples for the definition of verb *fail.* In the first domain, *fail* occurs in sentences such as *SAC failed.* The decomposition can be:

fail1 :- inoperative(patient(p))

But in the TRIDENT domain: *Magnetic tape unit failed to stop.* The new usage can be captured by saying that *fail* takes a single propositional argument; The representation can be:

fail2 :- not(proposition(actor(X), P))

Where P can be the proposition:

not(operate(sac))

In order to capture the part of meaning that might be shared between these two, Palmer proposed another more abstract notion to replace the first one.

fail3 :- not(perform-properly(actor(X)))

If we put all these three definitions into the conceptual hierarchy, the relations among them can be shown in Figure 6-6.

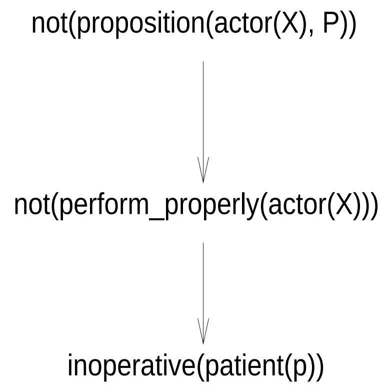



Figure 6-6: The relation of conceptual representations for *fail*

The figure shows that the concept *not-proposition* is a supernode of *not-perform-properly* in a conceptual type hierarchy. The concept *not-perform-property* is a supernode of *inoperative*.

### Possible solutions

Palmer proposed a possible solution to the problem of semantic-interpretation correspondence. Palmer has noticed that sometimes the interpretation is not based on the verb predicate representation only. For instance, the verb semantic representation of *break* has one core sense that is defined below:

**Meaning definition:** (Suddenly) to create a break (some form of separation) which causes alteration but does not necessarily violate the wholeness of the entity.

**Example sentences:** *break the skin, break the soil, break the branch off the tree*

**Representation:** cause(agent,

use(instrument,

separate-S(theme, source)))

Sometimes, the system is more concerned with the status of the *theme* or the *source* in the event. Palmer introduced a **change-of-state inference** to compensate for the insufficiency of the above predicate verb decomposition. So the following statement is attached to the verb definition as well(Palmer, 1990a) P. 25:

change-of-state inference [update(theme),
update(source)]

In the example of *breaks a branch off the tree*, before the breaking event, the branch is a part of the tree (source). After the event, the branch is a completely separate entity and can be involved in activities that have no connection with the original tree. So we must update both the source and the theme – tree and branch.

Palmer's representation is similar to the event structure representation later proposed by Pustejovsky. The sentence "The door closed" is represented as follows (Pustejovsky, 1991b) p.58:

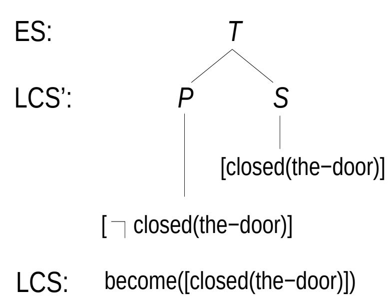



Figure 6-7: Event Structure for "The door closed"

However, Palmer is more concerned with how the verbs extend their meanings and the relation between verb meaning representation and inference. She also pointed out a serious problem for the extension of the coverage of the verb representation: "Any given sense can be extended almost indefinitely by applying it to items with new selectional properties", "It is not clear where to draw the boundary between senses." And "The crucial point for discussion is not exactly where to draw the lines between different verb senses, but how to develop processes whereby one sense can easily and naturally be extended into another one.". Her *change-of-state inference* can be one way of extending the verb meaning. However, this is not enough. A more systematic investigation on decomposing verb meaning for the semantic-interpretation correspondence is necessary.

## 6.4 Summary

NLP systems must have a broad coverage in a problem domain. As we all know, to let the system perform the task well, the system must have some mechanism for evaluating a language expression. It must find out whether the language expression is syntactically or semantically plausible. In the end, a syntactically and semantically best interpretation will be generated as the result of the analysis.

Chomsky has proposed the goal of the generative grammar as:

> The fundamental aim in the linguistic analysis of a language L is to separate the *grammatical* sequences which are the sentences of L from the *ungrammatical* sequences which are not the sentences of L and to study the structure of the grammatical sequences. The grammar of L will thus be a device that generates all of the grammatical sequences of L and none of the ungrammatical ones. (Chomsky, 1956) p. 13

The goal of an NLP system is much more than the goal above. The analysis of the language L is to find out all possible interpretations for the language expressions of L and evaluate every possible interpretation based on the system domain. Based on the best interpretation, the correct inference can be made. So the goal of the NLP system is to investigate the relations between the word semantics and the inference and make the correct prediction based on the current system domain.

As it is reviewed in Chapter 5, recent theories about lexical semantics focused on the semantic-syntactic correspondence. Researchers shun the *encyclopedia* view of word meaning and seldom touch the word meaning other than the scope of semantic-syntactic correspondence. Generally, the syntactic-irrelevant part of semantic representation falls into two extremes. One extreme is the *encyclopedia* view, which has no distinction between word semantic meaning and encyclopedic knowledge related to that word. This means all related knowledge can be in the contents of the verb entry: the verb entry might be open ended. The other extreme is the *paraphrase* view or *meaning postulates*, where the representation of the verb meaning is only a paraphrase of the verb. Most of the time, this paraphrase is a concept symbol in the knowledge base. The inference performed by the system will be limited by that concept in the knowledge base and has no connection with the verb any more.

Both of these views are not totally wrong. However, the word meaning is like the elephant in the story of "the blind men and the elephant". The particular connections between the encyclopedia knowledge and the paraphrase have not been perceived. In holding either of these views, it is difficult for computer systems to have a large coverage and efficient processing. Ignoring the relations between the encyclopedic knowledge and the paraphrase is the reason the computer systems are unable to to extend the verb meaning, and make correct interpretation.

Similar to Pinker's *grammatically-relevant subsystem hypothesis*, we propose the *inference-relevant subsystem hypothesis* as follows:

- Besides the grammatically-relevant subsystem, there is an Inference-relevant subsystem in the verb entry.
- Instead of having a set of primitives as the word meaning in the subsystem, there is a finite number of ways to decompose the verb into semantic components.
- Based on the decomposed semantic components, there is a finite number of ways to extend the verb meaning, and to make the right interpretation based on the domain knowledge.

If the *Interpretation-relevant subsystem hypothesis* is true, our task becomes clear. We must identify the ways in which a verb is decomposed for the purpose of semantic-interpretation correspondence, and find out the ways in which a verb is used in the system to make inferences based on the knowledge base to cover every usage. This problem will be addressed in the following chapter.

# Chapter 7

## Theory II: Verb Cognitive Domain Projection

> *"When I use a word," Humpty Dumpty said, in rather a scornful tone, "it means just what I choose it to mean, neither more nor less."*
>
> *"The question is" said Alice, "whether you can make a word mean so many different things".*
>
> *"The question is" said Humpty Dumpty, "Which is to be master – That's all".*
>
> – *Lewis Carroll: Alice's Adventures in Wonderland*

As it is claimed by Humpty Dumpty, people have great freedom in assigning meanings to words. The word meaning is **just what people choose it to mean**. In human communication, there are two ways to assign meaning to a word. One way is that there is no prior agreement on understanding of word meaning between the speaker and the hearer. The speaker must explain every unknown meaning first before he or she can use the word in the communication. We define this the way of Total Freedom With Explanation (TFWE). The other way is that both the speaker and the hearer have some consensus on the word meaning. The speaker can assign an unknown meaning that can be predicted by the hearer. We define this the way of Restricted Freedom With Prediction (RFWP). If a word is assigned a meaning that is out of the context and far away from the consensus, the hearer will not be able to understand what the speaker says. The communication is successful only when the hearer can correctly predict the word meaning. We are only interested in the second way. This kind of meaning extension has a close relation with the predictability of the unknown meaning. So we propose our **Meaning Extension**

**Restriction (MER):** The extension of the verb meaning should not go beyond the hearer's ability to predict the unknown meaning in the context.

As it is discussed in Chapter 4, for a computer system, the system coverage must be large enough to cover the whole domain. And this coverage is decided by the **Declared Knowledge** and the **Implicit Knowledge**. This is the same for the word meaning interpretation in the computer system. It is impossible to list each interpretation for all the word usages in the computer system. We have to list a small number of core verb meanings in the system and let other meanings be predicted by the system together with the context and the background knowledge. So we are facing the following problems:

- What information must be declared clearly in the verb entry?
- How should this information be represented?
- How should the unknown usage be interpreted within the system capability.

These questions are tightly connected and must be answered by all computer system designers who attempt to deliver a general lexicon processor. Traditionally, the analysis of the verb meaning is from the semantic-syntactic correspondence perspective, but in the following we will investigate the verb meaning from a new perspective: semantic-inference correspondence.

## 7.1 Cognitive relations among different usage types

The MER says that the only restriction for extending the word meaning is predictability of the unknown meaning. To investigate the predictability of the word meaning, we must understand the relations between the unknown meaning and the core meaning first. One good source for this study is dictionary verb entries. Entries of "break" in **Longman Dictionary of Contemporary English** (Longman, 1978) are chosen for the analysis. We are not saying that the definitions in this dictionary are the standard ones or that the analysis is the best. The entries in the dictionary serve as an example to show how the lexicographers categorize all usages into different prototypical usage types. The analysis aims at connecting all these usage types by locating them in the right positions in multi-dimensional cognitive domains. Before the analyses, we introduce two useful concepts from (Langacker, 1988):

**BASE** The BASE consists of those facets of pertinent domains that are directly relevant to the expression's characterization, hence necessarily accessed when the expression is used.

**PROFILE** The PROFILE for an expression is a substructure that the expression designates, making it maximally prominent within the BASE.

### 7.1.1 BREAK

The first entry in a dictionary can always be viewed as the basic or the core meaning of the word. Other entries can be viewed as the extension of the core meaning. So we take the first entry of BREAK for an analysis on meaning extension. Then all the other entries can be compared with the first meaning to see in which way the meaning is extended.

#### Core meaning of BREAK

In the dictionary, the first usage type is listed as follows:

> 1. To (cause to) separate into parts suddenly or violently, but not by cutting or tearing: *to break a window/a leg. The rope broke when they were climbing. The window broke into pieces.*

Clearly the definition is a usage type abstracting from many usages. The BREAK usage type in this definition lexically categorizes a set of cognitive events into one group. This group of cognitive events involves the following characteristics:

- Time domain. The event happens in a certain time and the duration is short. There should be the beginning of the event and an ending of the event.
- Space domain. The physical object that undergoes a separation exists in a location in the Space.
- Change of state domain: The physical object changes in its integrity and becomes separated into several parts. This concept serves as the PROFILE for these usage types.
- Causation domain: There is a cause for the change.
- Force domain (Optional): The force to cause the change of integrity is violently.
- Action domain: The actions involved in this event do not include the cutting or tearing action.

These cognitive domains are composed to form the whole picture of the core meaning of BREAK. This can be shown in the following Figure 7-1:

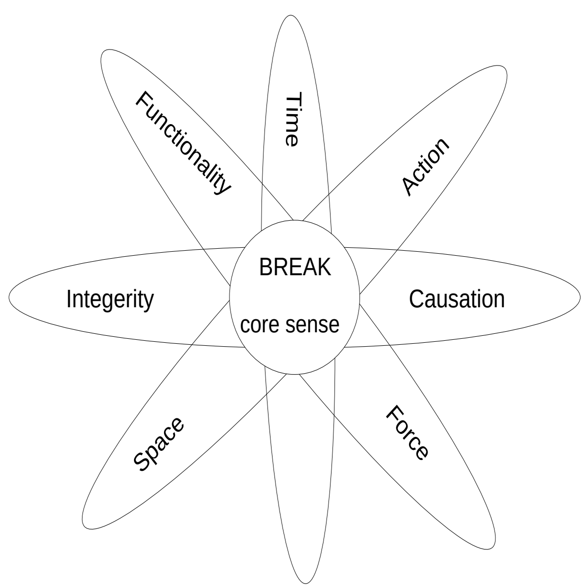



Figure 7-1: The cognitive multi-dimensional projection of core meaning of BREAK

### Other verb BREAK usage types

Other entries in the BREAK usage type list can be viewed as extended usage types. Each usage type is related to the core meaning. In the following, each usage type is analyzed to find out its correct position with respect to the core meaning in the human cognitive hierarchy.

> 2. To (cause to) become separated from the main part suddenly or violently, but not by cutting or tearing: *to break a branch off a tree. A large piece of ice broke away from the main mass.*

This usage type differs from the core meaning in the change-of-state domain. The usages in this type identify a concept: The physical object separates into two parts, with one part's main characteristics being unchanged.

> 3. To (cause to) become unusable by damage to one or more parts: *he broke his wristwatch by dropping it. This machine is broken and must be repaired.*

This usage type adds one cognitive domain to the word meaning – the functionality domain. A concept that "The physical object loses its function" is the PROFILE of the usage type. This concept relates to the integrity domain by the common sense knowledge that "A physical object that separates into several parts will always lose its function as a whole".

> 4. To (cause to) become, suddenly or violently: *The prisoner broke free/loose. The box broke open when it fell. They broke the door down.*

The PROFILE of this usage type is the concept of "sudden change of state". This concept can be viewed as the supernode of the concept "separate into several parts".

5. To open the surface of: *to break the skin/the soil.*

The PROFILE of this usage type is the concept of "the physical object surface is separated but the object is not separated". This concept can be viewed as the subnode under the concept of "change in integrity", and the sibling node of the "separate into several parts".

6. To disobey; not keep; not act in accordance with: *to break the law/ a promise.*

This usage type switches the physical object to an abstract order imposed on human behavior. The violation of the order is analogical to the physical object losing its function. The PROFILE is the action that makes this violation i.e. "to disobey", "not keep" etc.

7. To force a way (into, out of , or through): *He broke into the shop and stole $100.*

This usage type switches the physical object to a construction. The construction has its own integrity that prevents any illegal access. The PROFILE is the concept of "violation of the integrity".

8. To bring under control: *to break a horse/ a child's spirit.*

This usage type is similar to the type 3. Some abstract entities such as the spirit of the child or the horse lose their functionality.

9. To do better than: *to break a record in sports.*

The PROFILE of this usage type is the concept of "change of state to a better one". The concept is a subnode under the concept of "change of state".

10. To ruin: *If that young man tries to marry my daughter, I'll break him!.*

This usage type switches the physical object to an animate entity. “to separate an animate entity into several parts” is a kind of “to make the animate entity unfunctional”.

> 11. To destroy as an effective force: *We broke the enemy at the battle of Harlow Fields.*

This usage is similar to type 10. The animate entity is switched to an organization of the animate entities. The sense is under “lose of functionality”.

> 12. To make known (esp. something bad): *Break the bad news to him gently, please.*

This usage type is similar to type 6. There is a tendency to keep the bad news unknown, the meaning identifies the “violation of this tendency”.

> 13. To interrupt (an activity): *We broke our journey to Rome at Venice. The bushes will break his fall. Let's break for a meal and begin again afterwards. The red flowers break the green of the picture.*

When the continuity of an event is analogical to the integrity of a physical object, the “separate into several parts” is analogical to “make the event discontinuous”. These two concepts are under the same supernode concept: “violation of the tendency”.

> 14. To (cause to) come to an end: *to break the silence with a cry. The cold weather at last broke at the end of March.*

This usage type is the extension of type 13. A continuous event lose its continuity.

> 15. To come esp. suddenly into being or notice: *as day breaks. The storm broke. The news broke.*

The PROFILE of this usage type is the concept "suddenly come into being" which is under type 13. This usage type specifies the state change is from "nothing" to "exist", from "unnoticed" to "noticed".

> 16. To fail as a result of pressure from inside or outside: *His health broke. He may break under continuous questioning.*

This usage type is similar to type 8 and 10. It identifies a concept "some abstract property no longer holds". This can be under discontinuous.

> 17. To (cause to) change suddenly in direction, level, loudness, etc: *The ball broke away from the person trying to hit it. His voice broke with strong feeling.*

This usage type identifies the concept "sudden change in direction, or level or loudness" and take the "change of state" as its super concept.

> 18. To discover the secret of: *She broke their CODE. The police broke the case and caught the criminal.*

This usage type is similar to type 12. The PROFILE is the concept of "doing something to stop the tendency". Just as CODE has the tendency to stay secret, the case has the tendency to stay unsolved. The analysis of 'break' sense can also be found in (Palmer and Polguère, 1992).

### 7.1.2 Observations

From the above analysis, the word BREAK is projected to several cognitive domains. For 18 verb usage types and 5 noun usage types, the statistics of the projection is listed in the following table, where "++" means PROFILE, "+" means BASE, "−" means absence.

| | TIME | SPACE | C-O-S | CAUSE | FORCE | ACTION | FUNC. |
|---|---|---|---|---|---|---|---|
| VS1 | + | + | ++ | + | + | + | − |
| VS2 | + | + | ++ | + | + | + | − |
| VS3 | + | + | ++ | + | + | + | ++ |
| VS4 | + | + | ++ | + | + | + | − |
| VS5 | + | + | ++ | + | + | + | − |
| VS6 | + | + | − | + | − | ++ | − |
| VS7 | + | + | + | + | − | ++ | − |
| VS8 | + | + | + | + | − | + | ++ |
| VS9 | + | + | + | + | − | ++ | − |
| VS10 | + | + | ++ | + | − | + | − |
| VS11 | + | + | ++ | + | − | + | − |
| VS12 | + | + | ++ | + | − | + | − |
| VS13 | + | + | ++ | + | − | + | − |
| VS14 | + | + | ++ | + | − | + | − |
| VS15 | + | + | ++ | + | − | + | − |
| VS16 | + | + | ++ | + | + | + | − |
| VS17 | + | + | ++ | + | − | + | − |
| VS18 | + | + | ++ | + | − | + | − |
| PROFILE | 0 | 0 | 14 | 0 | 0 | 3 | 2 |
| BASE | 18 | 18 | 3 | 18 | 6 | 15 | 0 |
| Total | 18 | 18 | 17 | 18 | 6 | 18 | 2 |

Table 7.1: Statistics of BREAK cognitive domain projection

From the table we can see that most of the PROFILEs ,i.e., the information 'BREAK' expressions intend to deliver fall into the Change-Of-State (C-O-S) cognitive domain. The CAUSE domain acts as an optional PROFILE in the descriptions. The conceptual relations of these PROFILEs are shown in Figure 7-2. If we take the "functionality" as one of the state features, the "Change of functionality" can be a subnode of C-O-S. For type 6, 7 and 9, the C-O-S concepts fade to be the BASE, and the Actions emerge as the PROFILE.

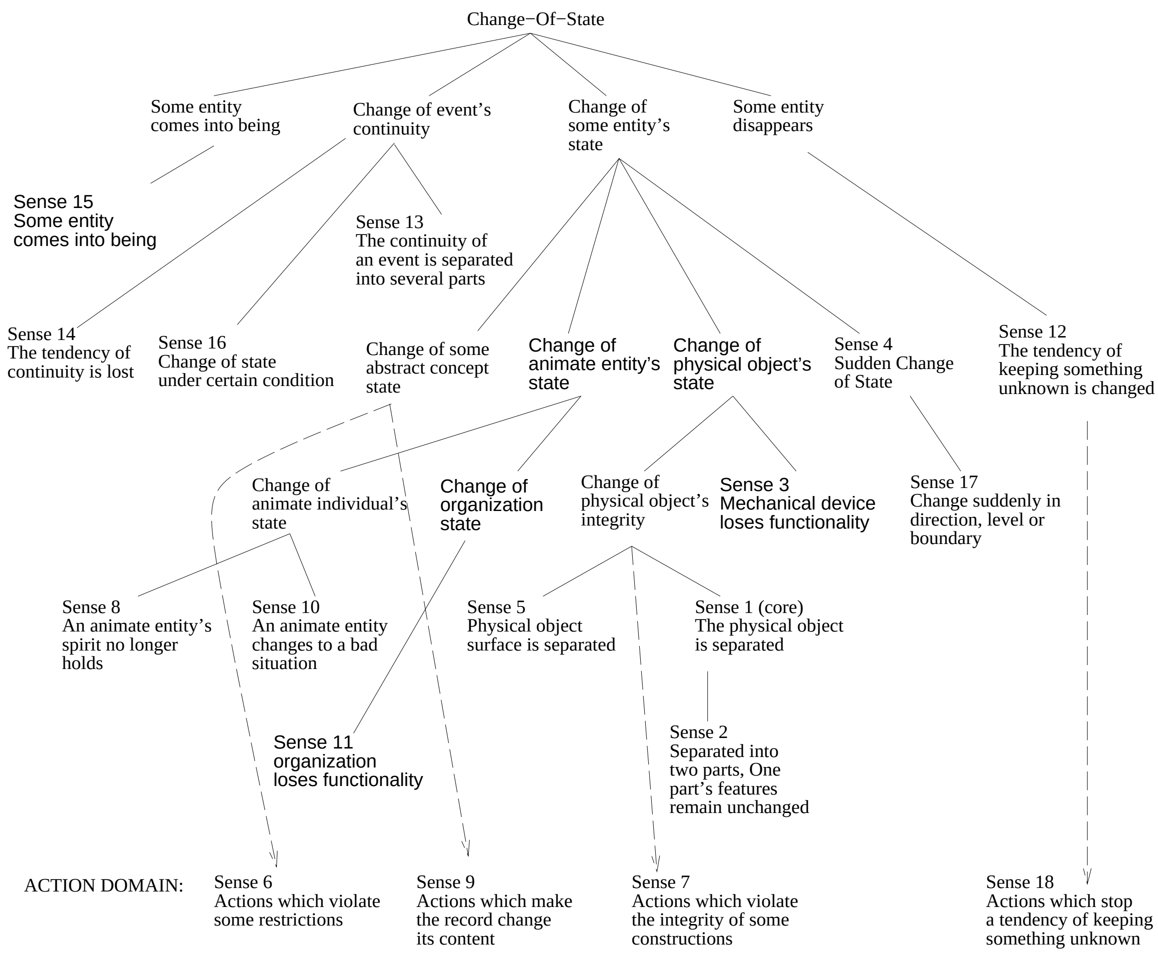



Figure 7-2: The cognitive relations of PROFILEs

The above figure is constructed according to the Is-A relations between different senses. The concept 'Change-of-state' is the top concept, while other concepts are the subconcepts of the top concept.

## 7.2 How do humans predict extended meaning?

Suppose that a person knows only the core meaning of BREAK, and all the extended meanings are unknown to him. Let us see the minimum conditions for him to make the right prediction.

### 7.2.1 The knowledge of the core meaning

First of all, the knowledge of the core meaning is obviously necessary for the prediction. If we do not know the verb's meaning at all, the only way for guessing the meaning is that the verb represents the event that has happened most often for that entity.

### 7.2.2 The knowledge of the entity

Secondly, the knowledge of the entity is also important. If we know nothing about the entity, the BASE of the core meaning will be imposed on the expression, the best guess is "the entity is a physical object" and the PROFILE is "the entity is separated into several parts". So we must have the knowledge of the entity in mind. Then more questions arise.

1. Should we know every piece of the knowledge about an entity?
   The answer for the question is NO. For instance, I do not know the number of atoms in my watch, but this does not prevent me from making correct communications using the word *watch*.
2. How much must we know about the entity?
   Generally speaking, natural language is used to communicate information in everyday life. So all the common sense knowledge about an entity should be known for understanding of the expression.
3. What knowledge is common sense knowledge?
   The answer comes from the name. Human beings have the cognitive abilities for perceiving the environment around them, i.e., common sense knowledge is the common knowledge of a human's sense about the world. Most of the knowledge about an entity is received from the cognitive processes such as vision, feeling, reasoning and many others. In the cognitive processes, different phenomena are categorized to form groups hence abstract concepts. So

if we classify the cognitive concepts into several domains, the common sense knowledge about an entity also can be projected onto several cognitive domains. For example, the watch in the FUNCTIONALITY domain is a device for telling the time, and in the PHYSICAL STATE domain, it has the shape of a circle and the position is always on the wrist.

4. Is all this knowledge equally important?

   We have said that common sense knowledge about an entity is the knowledge about that entity that has been projected onto several cognitive domains. As we have so many cognitive domains, does the knowledge in each domain has the same equal importance? Our answer is NO. An entity has its own salient features to distinguish itself from other entities. The knowledge about these salient features is more important than the features that are not possessed by the entity alone. Taking "watch" as an example, the knowledge of being functional and telling the time is more important than the knowledge of being a physical object. This allows people to think quickly.

### 7.2.3 Identifying the PROFILE

As we have the same sets of cognitive domains for verbs and nouns, the task is to find out the most probable PROFILE the expression intends to designate. Besides predicting the word meaning based on the syntactic-semantic correspondence, most of the prediction is based on the cognitive type hierarchy together with the understanding of the entities involved in the events. In the BREAK example, the following methods are used for prediction:

1. Identifying the PROFILE as the subnode of the core concept. As in type 2, the identified concept is a subnode of the core concept "separate into several parts".
2. Identifying the PROFILE as the supernode of the core concept. As in type 4, "sudden change of state" is a supernode concept of the core concept.

3. Concepts in the BASE are raised to be in the PROFILE. As in type 3, the "loss of functionality" is a BASE concept of the core meaning, but now it is changed into a PROFILE. In the noun usage of BREAK, such as "at the break", the identified concept in the TIME dimension is raised into the PROFILE of the expression.
4. Identifying a similar concept in another domain. As in type 13 and 14, the abstract notion of TIME dimension is involved, where the event is separated into several subevents in the time dimension.
5. Extending the core meaning using the above methods to a similar concept, then identifying the meaning of the nearest concept in the universal hierarchy. As in type 18, the meaning of "cause the CODE to be separated" is shifted to the meaning of "cause the CODE to be non-functional", then it is equal to "to discover the secret of the CODE".

### 7.2.4 Conditions for systems making correct inferences

As we all know, there are gaps between the system's problem space and the composed representation for natural language. If we can list every possible usage of a word in the system, the representation can trigger the inference correctly. However, the problem is that it is impossible to list all usages in full. We have to find some generalized way of composing language representation. If the generalization is a clear-cut one, once again, the solution is simple. But word meaning is well known for its dynamic, fuzzy characteristics. The usages of the word fall into prototypical fuzzy usage types. Because of generalization and the fuzziness of the word meaning, any large systems cannot avoid the problem of inference gaps between language representation and the problem domain. However, although we cannot have a fixed and precise representation for the word meaning, it is possible for us to have a fixed way of extending word meanings and make the approximate reasoning to find the correct result. Suppose state $s_1$ is one of the states in the problem space that triggers action $a_1$, but the language representation can only produce state $s_2$. In an exact matching system, action $a_1$ will never be triggered. However, if a similarity measure is given to the word meaning and the problem domain, and state $s_1$ is the nearest state to state $s_2$ that can take a legal action, then $a_1$ can still be triggered. This kind of inference fills the gap between the language representation and the problem space. The similarity mechanism will be the bridge over the gaps.

So we have the following conditions for making this bridge.

- For computer efficiency, the representation for the word meaning must be concise.
- The representation of the word meaning should be easily extended to cover as many usages as possible.
- There should be a mechanism to measure the similarities among different concepts identified by the language representations.
- The problem space of the system must be well defined. It should link naturally to the language representation.

## 7.3 Word meaning as Projection to Cognitive Domains

In natural language processing, computers are used to simulate human cognitive activities. The concepts that the language identifies and the cognitive activities the system needs to simulate, fall into the same domain, i.e., the human cognitive domain. So it is better to define the problem space and the word meaning on the same cognitive domains. In considering the word meaning, we agree with the argument in Fodor's "against definition" (Fodor et al., 1980). For the meaning of verb, it is quite difficult to give a fixed and precise definition. We also give up on the decomposition style of verb definition, since the representation scheme does not admit the fuzziness of word meaning. Moreover, a lot of information is lost during decomposition. What we are doing now is taking the verb as a whole, which cannot be decomposed into simpler primitives. However the prototypical usage types can be projected onto several cognitive domains. Besides the PROFILE that is identified by the verb, each projected domain also contains a concept that is identified by that verb. In this sense, the verb meaning is like a DOT in a multi-dimensional space. Each of the cognitive domains forms one of the dimensions. The concepts identified in each domain are just like the coordinates on those dimensions.

For example, the core meaning of the word BREAK can be projected onto the following cognitive domains:

**TIME:** The BREAK event has its beginning and ending time.

**SPACE:** The BREAK event happens in a certain space.

**CHANGE-OF-STATE:** PROFILE: The physical object in the event separates into several parts.

**CAUSATION:** PROFILE: Something causes the event happens.

**FUNCTIONALITY:** The physical object loses its functionality.

**ACTION:** Some action is involved.

This can be shown in the following figure:

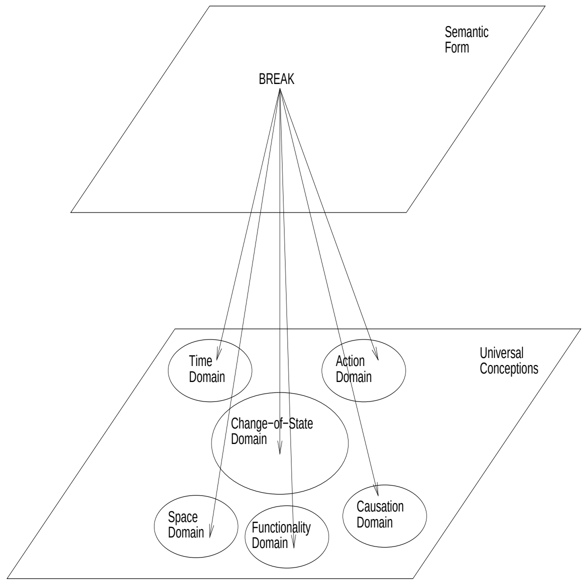



Figure 7-3: The projection of BREAK onto cognitive domains

The verb cognitive domain projection theory is consistent with Levin's verb classification according to the syntactic alternations (Levin, 1992). The semantic components identified by Levin's verb classes can be viewed as the concepts in different cognitive domains. So the projection theory can also take care of the semantic-syntactic correspondence well. In the following section, we will see that by building a universal hierarchy in each cognitive domain, the semantic-inference correspondence can also be handled well.

## 7.4 Cognitive event concepts hierarchical organization

We have said that the lexicalized cognitive event can be projected onto several cognitive domains. In one cognitive domain, those lexicalized concepts can be organized in a type hierarchy according to their relations with each other which is discussed in detail below.

### 7.4.1 Gaps between lexicalized conceptions

In human cognitive processes, not all the cognitive concepts have been lexicalized. The lexicon only identifies a small portion of the cognitive conceptions. Other cognitive concepts can be identified by the composed representation of the language expressions. Some others cannot even be expressed by the language expressions. Many such concepts can easily be identified by pictures or quantitative mathematical formulas, but they are difficult to express in words. For example, the following paragraph describes a graph.

> There is a big circle, inside the big circle, there are two half circles in the opposite directions. The diameters of these two half circles are the half of the diameter of the big circle. Then there are two small circles located at the center of the half circle, ......

It is hard for an English speaking person to imagine what the graph really is. But if the graph is shown in the following picture, people can understand it well.

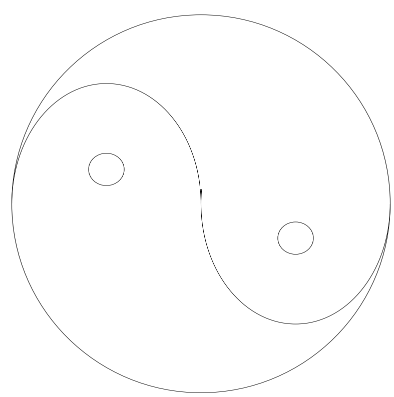


Figure 7-4: TaiJi: The symbol of Dao

The above picture is the symbol of Chinese ancient philosophy "道 (Dao)". It is lexicalized in the Chinese language as "太极 (TaiJi)". If this graph is not presented, for an English speaking person, he may get an idea about what Taiji is like from the above paragraph. But it is quite difficult for him to imagine it precisely. So all the inferences referring to the graph are based on the idea that the person gets from that paragraph. But for a Chinese speaking person who knows Taiji before, it is easy for him to recognize that the paragraph is referring to the Taiji graph. This is because there is already a conception of "TaiJi" in his mind. When he heard the paragraph, he gradually composed the picture of the graph and compared it with graphs in his memory. Finally, he found out that the described graph is very similar to the graph of "TaiJi". So he made the conclusion that the graph is a "TaiJi" or "TaiJi" like graph. This demonstrates how cultural differences can give rise to lexical gaps among languages.

Within the same culture, there are still gaps between the cognitive conceptions and the lexicalized concepts. For example, in English, "The stick is broken" is different from "The window is broken". They are two very different cognitive conceptions of "BROKEN". However, the same lexeme BREAK is used to describe them. If a system does not care about this difference, then these two different cognitive conceptions can be classified as the same state in the problem space. But if the system does care about this difference, they are two different states in the problem space, and then the links must be made from the conceptual representation of BREAK to the cognitive conceptions of these two different "BROKEN"s.

### 7.4.2 Cognitive domain type hierarchy

In human event perception, concepts do not not exist in isolation. They have relations among each other. Some of them can be classified into prototypical categories i.e. types. These concept types can be organized in a generalization hierarchy. There is already mach discussion on entity hierarchies in the world. The entities in the world form an ontological knowledge base in many computer systems. However, because of the complexity of event perception, the event hierarchical organization has not received much attention. It is complicated, because the events always involve many different cognitive domains. Obviously, it is quite difficult to express all the cognitive conceptions clearly, but fortunately, we have the language and lexicalized conceptions, which can serve as a skeleton in the hierarchical organization. The problem boils down to how to organize those lexicalized event conceptions. This is again a complicated problem. For example, a simple 'cut' event will involve the cognitive domains such as FORCE, MOTION, and CONTACT. An event conception is built upon several domains. So if we can build correctly the type hierarchy for each of these basic domains, it is possible for us to evaluate the relationship among different events. We take the following domains as the basic domains. They are: TIME, SPACE, CHANGE-OF-STATE, ACTION, CAUSATION, FORCE, MOTION, CONTACT, FEELING. This is by no means the final and precise list of basic cognitive domains. But in having this classification, we can restrict our investigation to a manageable scope. Figure 7-2 has shown the PROFILEs of different BREAK usages. The core meaning of BREAK has been identified as a concept in the CHANGE-OF-STATE domain as C1: "A Physical object separates into several parts". The second usage type identifies a concept that C2: "A physical object separates into two, one part is much larger than the other part" Here, we take the concept C1 as the subtype of the concept of C0: "An object changes its physical state". So C1 is a kind of C0, and C2 is a kind of C1. A simple similarity measure can be defined on the structure of the hierarchy: Suppose the number of nodes on the path P1 from C1 to the hierarchy root is N1, the number of nodes on the path P2 from C2 to the root is N2, Suppose P1 and P2 have N3 nodes in common, then the similarity of the concepts C1 and C2 is:

$$Sim(C1, C2) = \frac{2 * N3}{N1 + N2} \tag{7.1}$$

In Section 7.2.4, we have discussed the necessary conditions for the system to make the correct inferences. They are 1) the requirements of similarity measure and 2) the natural connection between the problem space and the language representation. Together with the above analysis, we find that the cognitive domain type hierarchical organization is a good way to address the problem. This hierarchical organization is composed of not only the concepts identified by the lexicon, but also the cognitive concepts that are the states in the problem space. So the word meaning can be easily extended to the concepts that are similar to the lexical concepts. And the distance of different concepts can be easily measured. The practical application of this theory is realized in a prototype system of lexical selection for machine translation which will be presented in the next chapter.

## 7.5 Related research

The cognitive event type hierarchy provides ways for evaluating the similarity among different concepts. It also provides ways for verbs to extend their meanings to similar concepts near the concepts identified by the usage types. There are some previous studies on word semantics that have the same goal as the current approach that deals with the dynamic and fuzzy characteristics of word meanings. In the following, I will summarize these approaches and discuss the differences between them and ours.

### 7.5.1 The cognitive semantics tradition

Geeraerts (Geeraerts, 1988) has presented a good review of the history of cognitive semantics. He summarized a set of characteristics that play a very important role in the main stream of cognitive semantics.

- Lexical concepts have vague boundaries, in the sense that they contain peripheral zones around clear conceptual centers (versus: lexical concepts are discrete, well-defined entities) – The prototypical view of categorization implies that categories may have marginal instantiations that do not conform rigidly to the central cases.

- Lexical concepts are polysemous clusters of overlapping semantic nuances (versus: the various senses of a lexical item can always be strictly separated from each other) – The vagueness that characterizes conceptual categories as a whole is mirrored by their structure.

- The distinction between analyticity and syntheticity, and that between essential and accidental attributes cannot be rigidly maintained (versus: there is a strict distinction between analytical and synthetic statements, and between essential and accidental attributes).

- Lexical concepts may be disjunctively defined (versus: lexical concepts are necessarily defined conjunctively). Within a protopically organized category, membership may be based on sufficient similarity rather than identity.

- Attributes within a category may have different degrees of salience (versus: all attributes within the definition, or examples in the extension of a concept have an equal degree of salience).

- Lexical concepts function in a flexible and analogical manner (versus: lexical concepts function in a rigid, algorithmical fashion) – The fact that category membership can be defined by similarity rather than identity entails that conceptual categories can be used in an extremely flexible way.

- lexical concepts have to be studied as a proper part of human cognition at large (versus: lexical concepts have to be studied as a part of an autonomous linguistic structure).

- There is no distinction between semantic and encyclopedic knowledge (versus: the semantic definition of a lexical concept is to be distinguished from the encyclopedic data that can be connected with the latter).

- Semantic studies cannot ignore the experiential and cultural background of the language user ( versus: semantic phenomena should be studied apart from user or culture-specific background data).

The current *Verb Projection* theory has something in common with the cognitive tradition:

- It admits that verb meaning has vague boundaries. The verb usage falls into fuzzy usage types.
- Those usage types are fuzzy clusters, they may overlap with each other.
- Attributes are not divided strictly into essential sets and accidental sets. The degree of the importance of the attributes is decided by the conceptual hierarchical structure.
- The concepts can be measured by similarity as the reflection of the conceptual organization.
- And because the experiential and cultural background of the language user decides the cognitive conceptual organization, they cannot be ignored by the semantics.

However, the *Verb Projection* theory differs from the cognitive semantics in the following ways:

- There is a rigid, algorithmical way to formalize the problem in the computer system. The data and algorithms in the system define a domain of language usage. It is possible to achieve the best performance based on this domain.
- There is a distinction between semantic and encyclopedic knowledge. As long as we correctly locate the lexicalized concepts in the cognitive hierarchical structure and find out ways for semantic forms to extend their meanings, we

have built a bridge between the semantic knowledge and the encyclopedic knowledge.

### 7.5.2 The generative lexicon

Pustejovsky's generative lexicon (Pustejovsky, 1991a) is a recent attempt to find ways to extend word meanings. In avoidence of the problems caused by the Jackendvian style of decomposing verbs *exhaustively* into primitives, he proposed that:

> I would like to suggest that lexical (and conceptual) decomposition is possible if it is performed *generatively*. Rather than assuming a fixed set of *primitives*, let us assume a fixed number of *generative devices* that can be seen as constructing semantic expressions.

Pustejovsky further proposed the *Qualia Structure* for nominals. The Qualia structure is the essential attributes of an object as defined by the lexical nominals. These attributes provide information for *Type Coercion*. The information or the knowledge about the object includes the relation between the object and its parts (Constitutive Role), its relation to the environment (Formal Role), its relation to the human being i.e. functionality (Telic Role) and cohere the object comes from (Agentive Role). The Qualia structure provides a rich scheme for representing nominals. However, his *Event Structure* (Pustejovsky, 1991b) has not been designed in a *generative* style. Our approach brings in the notion of cognitive domains, and lets the verb meaning be the structure projected onto these domains. In this way, wepropose that verb meaning can be defined in a way that is similar to the Qualia structure. And the cognitive conception hierarchical organization provides ways for extending the verb meaning.

## 7.6 Summary

In this chapter, we proposed a *Verb projection* theory about verb semantic representation. The theory differs from the cognitive semantic view that takes the verb semantics to be the same as encyclopedic knowledge. Instead, the theory separates the verb semantic form and cognitive concepts into two different levels. At the semantic form level, verb usages form fuzzy usage types according to the cognitive concepts they project onto. The verb can be extended to other usages and can be projected on other cognitive concepts as well. By making use of the hierarchical organization of the cognitive concepts, only the best interpretation is taken to be the verb meaning. So the verb meaning representation is not only the projected concept, it also includes the ways of how to extend the core meaning of the usage types and the cognitive hierarchies that provide the base for the extension. This kind of word meaning treatment has the following merits:

- It is a good mechanism for fuzzy treatment of the word meaning. There are no boundaries for verb meaning, only fuzzy usage types.
- Verb meaning can be easily extended.
- It is easy to give a similarity measure to the cognitive concepts and verb meanings.
- The hierarchy we are using here is not the type of hierarchies we discussed in Chapter 4. It not only deals with the inheritance relation for saving the storage, but also provides ways of extending word meaning.

In the next chapter, we will see how this theory is used for solving the problem of lexical selection in machine translation.

# Chapter 8

# The UNICON System

UNICON (UNIversal CONception) is a prototype system solving the inexact match between source language expression and target language expression. It attempts to show that inexact match in lexical selection is better performed when the verb is projected onto cognitive concepts in several basic cognitive domains organized in hierarchical structures. It also shows that the meaning can be easily extended to similar concepts that are on the paths in the hierarchy. The system assumes a modular approach. It views the syntactic processing and semantic processing as different processes. However, this modular view does not prevent the information from being transmitted from the later stage of processing to the early processes. A re-evaluation mechanism could be implemented on top of the system architecture. This system focuses on the semantic processing module. The input to the system is a source language argument structure that is assumed to have been generated by a syntactic analyzer. The output of the system is the target language argument structure. The syntactic realization for the target language generation is omitted. The system contains the following components:

**SEMDIS Module** This is a semantic analysis module. In this module, the internal meaning representation for the input argument structure is generated. Selection restrictions are used for disambiguation of verb meaning. The special characteristic of this module is the hierarchical conceptual organization. The selection restrictions can be relaxed along the paths in the hierarchy. The knowledge about the arguments plays an active role in the analysis.

**SELECT Module** This is a lexical selection module. After the SEMDIS process, the internal meaning representation has been composed, i.e., a concept has been identified. This module will attempt to find an argument structure in the target language that matches best to the identified concept in the system domain. The hierarchical conceptual organization in the module enables the system to evaluate the match in a quantitative measure. The search for the right word is not restricted to several fixed choices. The scope can be extended to the whole lexical base if necessary.

**EXTEND Module** This module is for learning and justifying the usage types. When a new usage is encountered, the preconditions in the verb lexeme definitions are the selectional restrictions for sense disambiguration. When a set of usage types is predefined, a particular usage may fall into one of the usage types or none of them. If it does not fall into any usage types, the sense disambiguration is that either we take the usage as a particular new usage type and define it in the dictionary or we find out the most similar usage type and classify the particular usage into that type.

**UNICON Dictionary** This is the universal conceptual knowledge base. Not only entities in the world, but also the cognitive situation perceptions are organized in a hierarchical way. This hierarchy has two important characteristics: First it can greatly improve the efficiency of access and storage. Second, it encodes the relationships among different concepts in a formal way. The similarities among different concepts can be measured as a reflection of the hierarchical structure.

**CHINESE Dictionary** This is the dictionary for Chinese. All the words are defined based on the UNICON knowledge base. Because the system is only concerned with semantic processing, the syntactic information and other pragmatic information is omitted.

**ENGLISH Dictionary** This is the dictionary for English. Again, all the words are defined based on the UNICON knowledge base.

These relations among the components in the UNICON system are shown in the following figure:

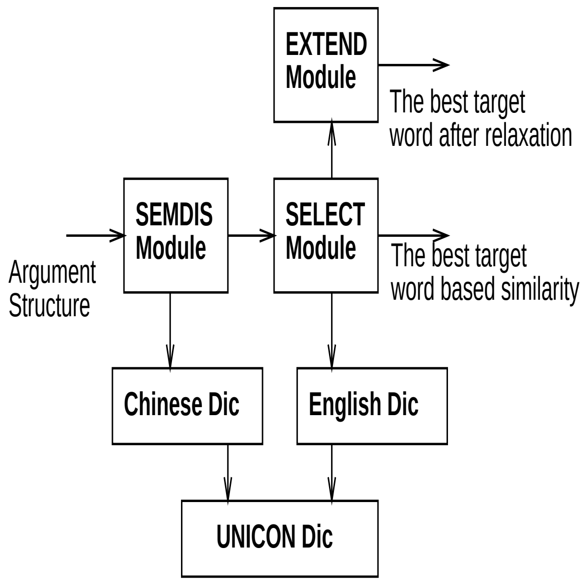



Figure 8-1: The UNICON system

In the following sections, I will explain each of these modules. First, let us see whether we do have a universal conception.

## 8.1 Do we really have universal conceptions?

Do we really have a universal lexicalized conception? The answer is NO. Anna Wierzbicka (Wierzbicka, 1992) has a good example to show this. Japanese have a particular emotion called "amae". This "amae" can mean "helplessness and the desire to be loved", "sweet", "to lean on a person's good will", "to depend on another's affection", "to act lovingly towards", "to presume upon" , "to take advantage of" ...... It is such a complex feeling that many English paraphrases can be generated. This shows that Japanese has formed a very particular lexicalized concept that cannot be found in any other languages.

In a study of children learning hehavior, Choi and Bowerman (Choi and Bowerman, 1991) have shown that English and Korean children differ in how they lexicalize the components of motion events. English characteristically conflates Motion with Manner, or Cause, and expresses Path separately. While Korean conflates Motion with Path and elements of Figure and Ground in transitive clauses for caused Motion. Their finding challenges the view that children initially map spatial words directly to the same non-linguistic spatial concepts, and suggest that they are influenced by the semantic organization of their language virtually from the beginning. The language input and cognition interact with each other in the learning and perception process. This shows that languages and cultures are the main factors in forming concepts. Different languages have different lexicalized concepts that serve as the particular categorization of human cognitive perception. Therefore different languages are defined on different categorization systems.

Then can we translate among these different systems? Do we have a way to translate "amae"? We have two alternatives. One is to translate it precisely. It will take several pages to explain what "amae" really means. The other choice is to translate it with an acceptable distortion. We need to find out the closest paraphrase for this word, such as "to take advantage of". The evaluation of the similarity of meaning becomes very important. Fortunately, human beings have universal cognitive capability. We have the same cognitive organs such as ears, eyes, and mouth. Our motor organs are the same. The perceptions are based on the same biological organizations. We live in a roughly similar environment. This provides ways for evaluating different categorization systems for different languages and makes translation possible. Even different languages have different lexicalized conceptual systems, and their construction blocks should fall into the universal cognitive concepts. The task for us is to find out those building blocks and define the lexical concepts based on them. Only in this way can we measure the similarities among different lexicalized concepts.

## 8.2 UNICON

The universal conception in the sense of cognitive perception should be a generative one. A finite number of primitives are used to cover all of these cognitive perception. Other lexicalized concepts can be composed by the construction blocks. If the concepts work well for our pratical purpose, they are not decomposed further. The universal concepts are organized in a frame hierarchy. The Frame system FRAMEWORK developed by Mark Kantrowitz at CMU is used as the Frame knowledge representation language. This hierarchy differs from other similar approaches in the way that it treats the situation concepts as equally important to the entity concepts. The situation hierarchy represents the way people classify cognitive concepts into categories.

### 8.2.1 Framework

Framework is a common-lisp portable frame-based knowledge representation language. It combines some of the best features of a variety of other frame languages. It includes a variety of tools for building, examining, and using knowledge-based systems. Framework, just as all other frame based systems, is ultimately based on Minsky's frame theory of the representation of knowledge. A frame is an object with an associated list of attributes. Objects are data structures with associated procedures. The attributes of a frame include VALUES and BEHAVIORS. Using frames we can create a network of objects representing facts, things, and other concepts, connected through a variety of links. Objects and their links are represented as frames with names, slots, facets, and values. Each object has a name (frame) and a set of slots; each slot has a set of facets; and each facet has a set of values. Values may be arbitrary lisp objects, including functions and the names of other frames. In some sense a frame is a generalized property list: it contains more than just values, and can inherit information from related frames.

Frames are implemented as nested association lists (key-value pairs) and are stored in a hash table under the frame name for efficient access:

```
      (<frame name>
 (<slot1> (<facet1> . (<value1> <value2> ...))
  (<facet2> . (<value1> ...))
  ...)
 (<slot2> (<facet1> . (<value1> ...)) ...)
 ...)
```

Slots may be used to name relations by making the name of one frame the value of the slot in another frame. For example, the :AKO slot (an abbreviation for "A Kind Of") and its inverse, :KINDSOF, link objects in the class hierarchy. Many other types of links exist, for instance, :CLASS / :INSTANCES, :TYPE / :TYPES, :PARENT / :CHILD, :PART-OF / :PARTS, :IS-COMPONENT-OF / :COMPONENTS. These relations may be used to connect the frames into a network. The primary facet is :VALUE, which is used to store the value of an attribute.

The major features used in UNICON are:

**Inheritance** The conceptual hierarchy is built by using the :AKO and :KINDOF links. The :PARTS and :PART-OF relations are also used for some conceptual relations such as the relation between the human body and the human hand.

**Frameget** The attributes of a node can be inherited from its supernodes by the function Frameget. There are also functions that find out the supernode set and subnode set for a particular node.

**Transitive closure** Framework provides a function that can find the transitive closure of FRAME by following paths specified by SLOT/FACET by traversing the paths in a breadth-first manner. This function is used to find out path length from a node to the root. The set obtained by this function is all the nodes on the path.

The general situation hierarchy in UNICON is shown in Figure 8-2:

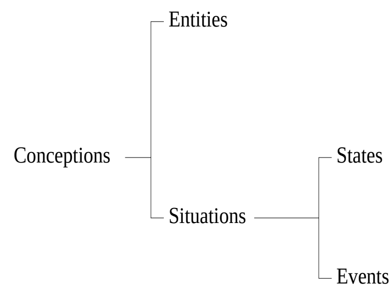



Figure 8-2: The general conceptual hierarchy

### 8.2.2 Entity hierarchy

The hierarchy for entities in the world is adopted from (Mo, 1992). The difference is that ours has extended the situation concepts as equally important to the entities. The entity hierarchy is shown in Figure 8-3.

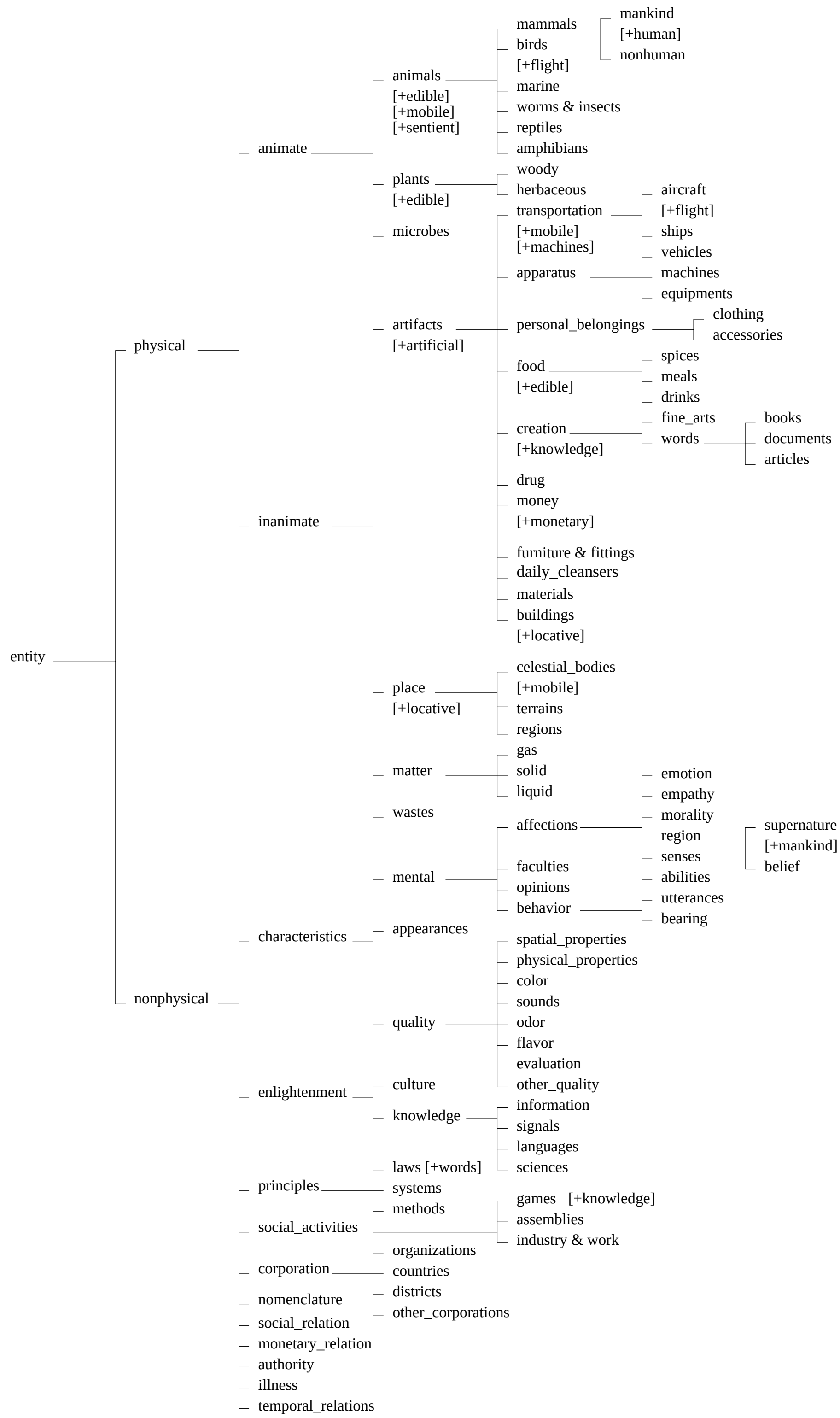



Figure 8-3: The general entity conceptual hierarchy

A representation of 'hand' is given below. The slot 'lexeme' saves the language lexemes which are the surface realization of that entity. Hand is also a part of a mammal, the :part-of slot records this. The particular slot is the 'situation-index' slot. It records the most often situations the entity involves in. It works as a quick index to the situations. In some sense, it has the same function as Pustejovsky's telic and agentive slots in Qualia structure for nominals.

```
(defclass %hand (%solid)
    (%lexeme (%chinese 手)
             (%english hand))
    (:part-of (value %mammals))
    (situation-index (value (%break-i-1a %折断)))
    )
```

### 8.2.3 Situation hierarchy

Part of our situation hierarchy is shown in Figure 8-4.

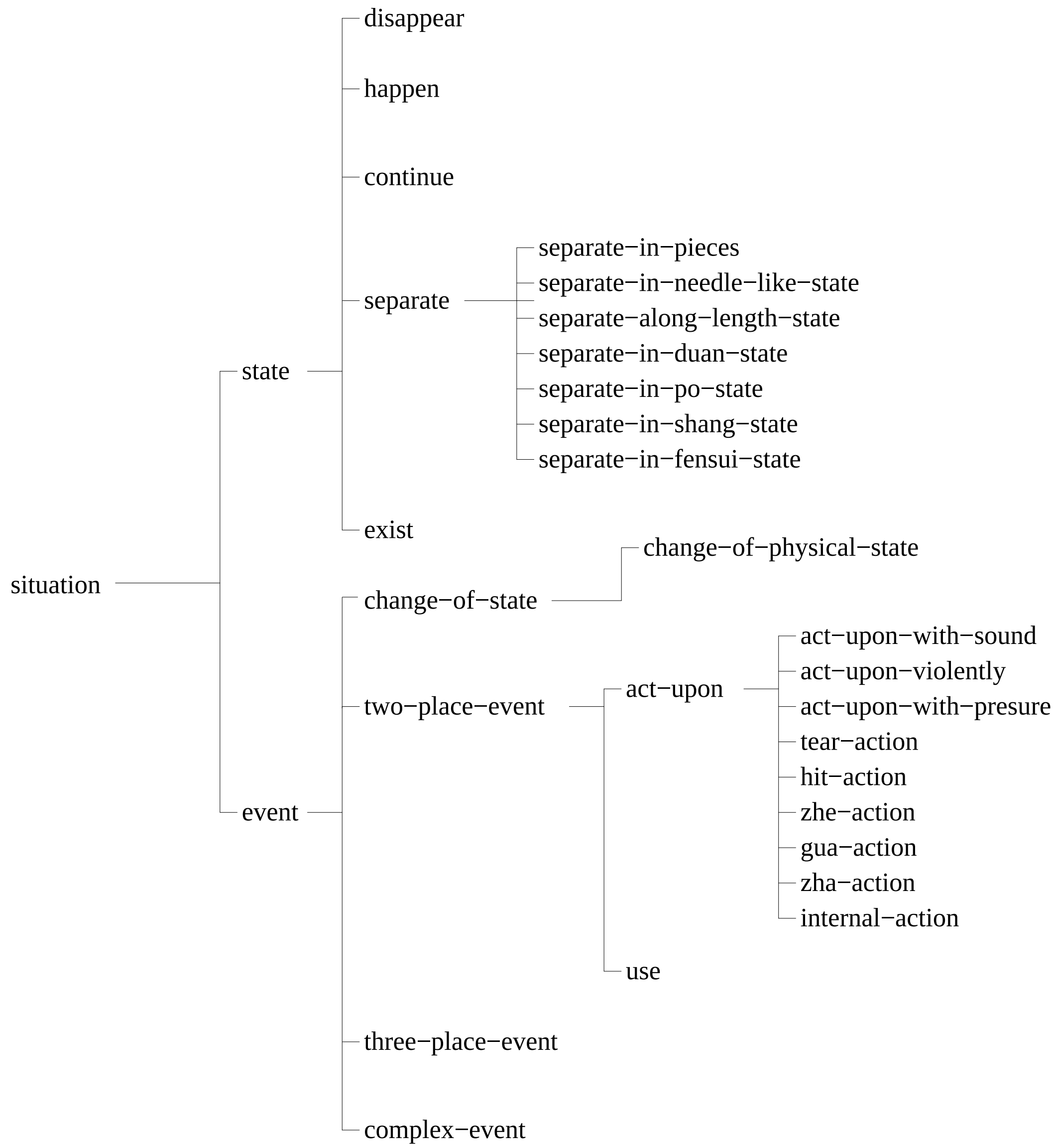



Figure 8-4: The general situation conceptual hierarchy

Situations are classified into two types, one is the state that is the basic type taking only one argument, the other is a complex type that can be decomposed into basic types according to a time. The basic state type can be further divided into 6 types according to time:

1. Type 0: To describe the entity's state at one time.
2. Type 1: To describe the entity's state without time restriction.

3. Type 2: To describe the entity's state after a time.
4. Type 3: To describe the entity's state till a time.
5. Type 4: To describe the entity's state till a time then begin the event.
6. Type 5: To describe the entity's state within a time.

Especially for situation concepts, there must be a set of conditions for the situation to happen, and there must be a set of consequences after the situation happens. So the situation concepts have four components:

**LEXEME** This is the possible realization in various languages.

**ARGUMENT-STRUCTURE** This is a list of the participants involved in the situation.

**PRECONDITION** This is the conditions for the situation to happen. It is similar to the BASE concept as discussed in Chapter 7. When a lexicalized verb concept is considered, it is the selectional restrictions imposed on that verb.

**CONCEPT** This is the situation definition. It is similar to the PROFILE concept as discussed in Chapter 7.

**CONSEQUENCE** This is the possible effect after the situation happens.

An example is shown below:

```
(defclass %change-of-physical-state (%change-of-state)
    (%lexeme (%chinese nil)
             (%english nil))
    (argument-structure (%change-of-physical-state @var1))
    (precondition (value ((%is-a %physical @var1))))
    (conception (value ((temporal-relation (@d0 @d1))
                        (@d0 (equalp (%get-feature %physical-state @var1)
```

```
                    @var2))
             (@d1 (equalp (%get-feature %physical-state @var1)
                    @var3))
             (@d1 (not (equalp @var2 @var3))))))
  (consequence (value ((%lost @var1 %function)))))
```

## 8.3 CHINESE and ENGLISH dictionaries

The word dictionaries are defined based on the UNICON dictionary. The nominals lexicon for each language are represented as given below:

```
;;; Chinese Noun
(defclass chn-noun (chn-entity)
  (syntactic-feature (cat n)))
;;; Chinese Branch
(defclass 树枝 (chn-noun)
  (vocable (value 树枝))
  (semantics (value %branch)))
;;; English Noun
(defclass eng-noun (eng-entity)
  (syntactic-feature (cat n)))
;;; English Branch
(defclass branch (eng-noun)
  (vocable (value branch))
  (semantics (value %branch)))
```

The verbs are defined as below. Each verbal lexeme has a vocable value that is its surface form. It also has a semantic value. This value identifies a unique concept. This concept is then defined based on the UNICON dictionary.

```
;;; Chinese Verb 折断 "Zhe Duan"
(defclass 折断 (chn-predicate)
    (vocable (value 折断))
    (semantics (value (%折断 @var1 @var2))))

     (defclass %折断 (%change-of-physical-state)
    (%lexeme (%chinese (折断)))
    (precondition (value ((%is-a %animate @var1)
                          (%is-a %hand @var3)
                          (%part-of @var3 @var1)
                          (%is-a %physical @var2)
                          (equalp (%get-feature %shape @var2)
                                  %line-segment))))
    (conception (value ((temporal-relation (@d0 @t0 @d1))
                        (@d0 (%use @var1 @var3))
                        (@t0 (%zhe-action @var3 @var2))
                        (@d1 (%separate-in-duan-state @var2)))))
    (consequence (value ((%lost @var1 %energy)
                         (%lost @var2 %function)))))
;;; BREAK I-1a.
(defclass break-I-1a (break)
    (vocable (value break))
    (semantics (value (%break-i-1a @var1))))
(defclass %break-i-1a (%change-of-physical-state)
    (%lexeme (%english (break-i-1a)))
    (precondition (value ((%is-a %physical @var1))))
    (conception (value ((temporal-relation (@t0 @d1))
```

(@t0 (%act-on @unk0 @var1))

(@d1 (%separate @var1)))))

(consequence (value ((%lost @var1 %function)))))

## 8.4 SEMDIS module

This module is for verb semantic disambiguation. Verb usages have been classified into usage types and stored as lexemes in the language dictionary. A source verb's argument structure is evaluated according to the preconditions of each of the lexemes. The evaluation results are a measurement of the distance from the argument structure to that particular lexeme, i.e., that usage type. The meaning of the argument structure corresponds to the nearest lexeme. The precondition for a lexeme is a set of the condition expressions for the arguments in the argument structure. The most common form of expression is:

(%is-a A B)

It is used to determine whether A is a kind of B. The other one is:

(%has-feature A F)

This expression is used to determine whether A has the feature F. Here, A is a variable that must be instantiated by the arguments in the input argument structure. B and F are instance values. Their position can be located in the UNICON hierarchy. There can be some other conditional expressions.

The special characteristic of this module is the evaluation of the conditional expressions and the treatment of the instance values in the expression. The evaluation is an operation on the hierarchy. The instance values are the nodes in the hierarchy. The particular arguments are the instances of the classes in the hierarchy. The evaluation will reflect their positions and distances in the hierarchy. For example, the sentence:

"The man cut the bread"

has an argument structure:

(cut man-1 bread-1).

man-1 is an instance of the class %man in the hierarchy. Suppose the conditional expression is:

(%is-a @var1 %man)

When @var1 is substituted by man-1, the condition expression becomes:

(%is-a man-1 %man)

This conditional expression is fully satisfied, giving a satisfied value of 1. However, if the conditional expression is: (%is-a man-1 %animal), although the IS-A relation is satisfied, %man and %animal have a conceptual distance in the hierarchy. To reflect this distinction, the satisfaction degree of the conditional expression can be defined as:

$$satisfaction\ degree = \frac{Number\ of\ common\ features}{Number\ of\ distinct\ features} \tag{8.1}$$

The equation can be complex if we consider giving weight to each feature. But it can also be simplified by only considering the :AKO path. We can define N1 as the number of nodes on the path from the root to %man and N2 is the number of nodes on the path from the root to %animal. Then the satisfaction degree can be defined as:

$$Satisfaction\ degree = \frac{N2}{N1} \tag{8.2}$$

The hierarchical structure also provides ways to relax the constraints. The instance values in the conditional expression can be relaxed to allow similar concepts near it in the hierarchy. For example, if the conditional expression for the CUT lexeme is:

(%is-a @var1 %human)

Then for the sentence:

The monkey cut the bread.

The condition is not satisfied. Suppose that this usage does not satisfy the precondition for all CUT usage types. This usage can be viewed as either a new usage type or an extended usage of existing usage types. It is possible to relax the constraints and do the evaluation again. For the above particular constraint, the instance value %human can be relaxed to its supertype %animal. This makes it possible to classify the usage into the usage type of "The man cut the bread".

The flow chart of this module is in Figure 8-5.

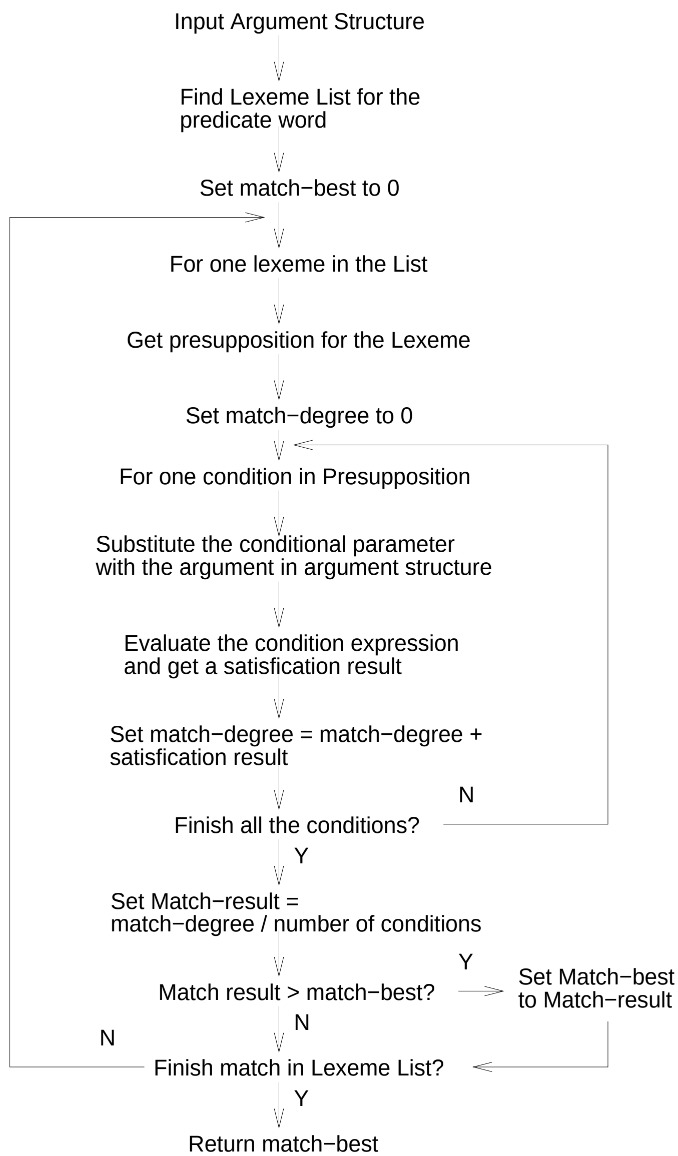


Figure 8-5: The flow chart of the SEMDIS module

## 8.5 SELECT module

This module selects the best verb in the target language that matches the internal meaning representation. As discussed in Chapter 7, each verb usage type can be viewed as a particular categorization in the cognitive space. A situation type is defined by this particular cognitive categorization. Since the building blocks for the internal meaning representations are the concepts in the cognitive domain, the matching algorithm starts from these meaning concepts. For each word that is the surface realization in the target language for these concepts, the system matches the conceptual definition of that word to the internal meaning representation. If a satisfied match has not been found, those concepts that are near the meaning concepts in the hierarchy are chosen to be the candidate concepts. Then the match is done on the lexemes that are the realizations of the candidate concepts again. The system right now only searches within three levels away from the meaning concepts. In principle, it is possible to continuously enlarge the searching area and cover the whole hierarchy.

The flow chart of this module is shown in Figure 8-6.

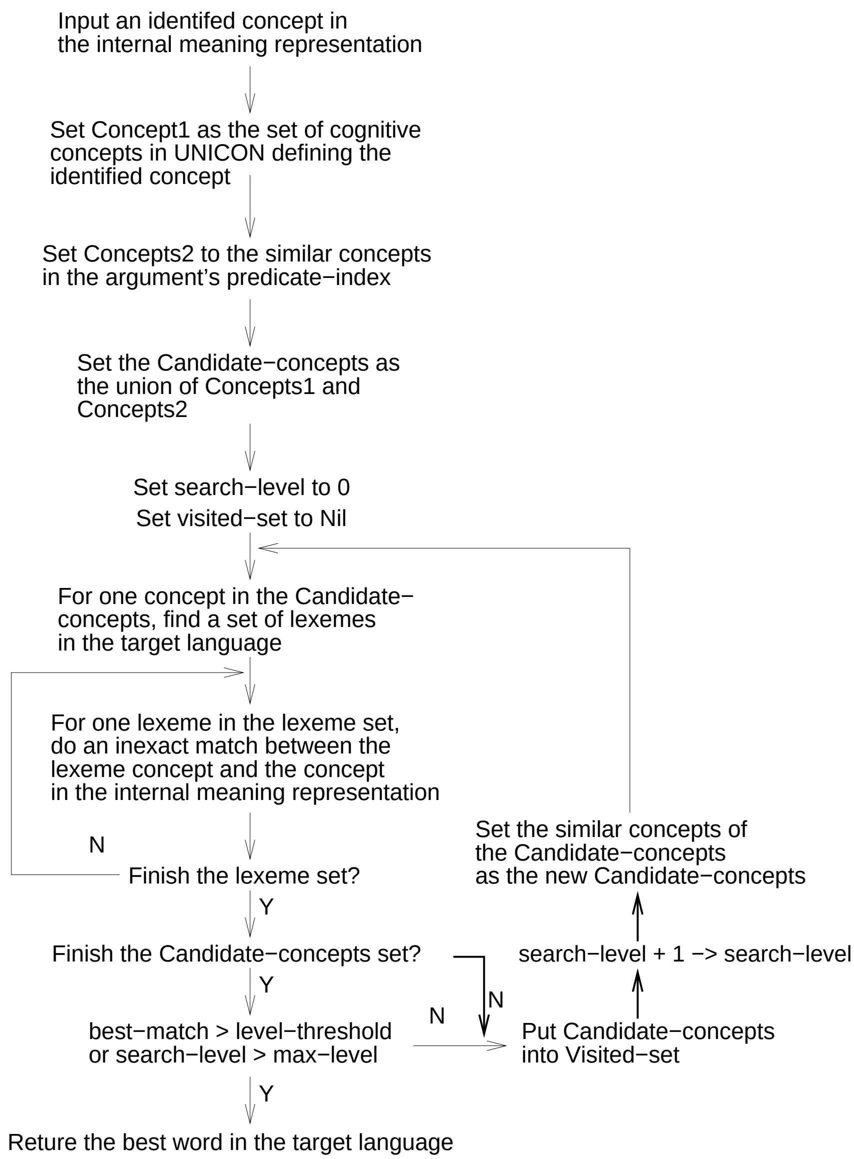



Figure 8-6: The flow chart of SELECT module

## 8.6 EXTEND Module

In UNICON, verb meanings are treated as fuzzy usage types. Most of the inputs can be handled correctly, if they fall into those fuzzy usage types. However, there are still some inputs that are outside the coverage of the usage types. These inputs can be viewed as ill-formed in the current knowledge domain. On the other hand, they can be viewed as evidences for learning new lexical knowledge. Basically, there are two cases. One is that the input usage conflicts with any of the existing usage types. If the guessed usage meaning is acceptable, the usage forms a new usage type. The other is that the input usage is the most similar to one of the usage types. The system can enlarge that usage type to cover this particular usage.

The flow chart of the EXTEND module is shown in Figure 8-7.

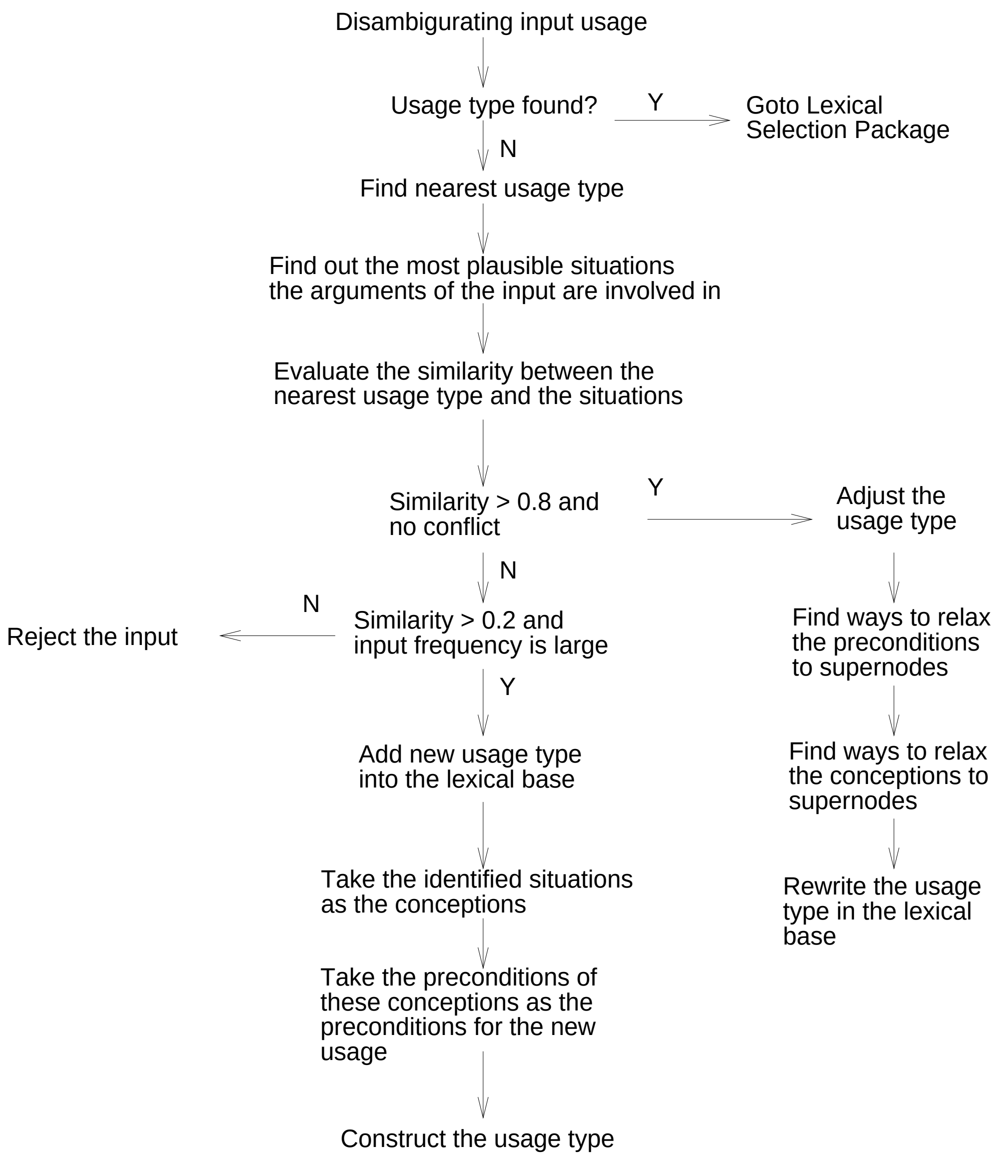



Figure 8-7: The flow chart of the EXTEND module

## 8.7 Levin's Verb classifications

Levin's verb classifications (Levin, 1992) can be the starting point of this system. We have chosen one set of her classes for building our hierarchy. The classes are:

- 12 Verb of Exerting Force: *Push/Pull* Verbs.

- 17 Verb of Throwing

  17.1 *Throw* Verbs

  17.2 *Pelt* Verbs

- 18 Verb of Contact by Impact

  18.1 *Hit* Verbs

  18.2 *Swat* Verbs

  18.3 *Spank* Verbs

  18.4 Nonagentive Verbs of Impact by Contact

- 20 Verbs of Contact: Touch Verbs

- 21 Verbs of Cutting

  21.1 *Cut* Verbs

  21.2 *Carve* Verbs

- 23 Verb of Separating and Disassembling

  23.1 *Separate* Verbs

  23.2 *Split* Verbs

- 45 Verbs of Change of State

  45.1 *Break* Verbs

  45.2 *bend* Verbs

The semantic relations among these verb classes is shown in Figure 8-8.

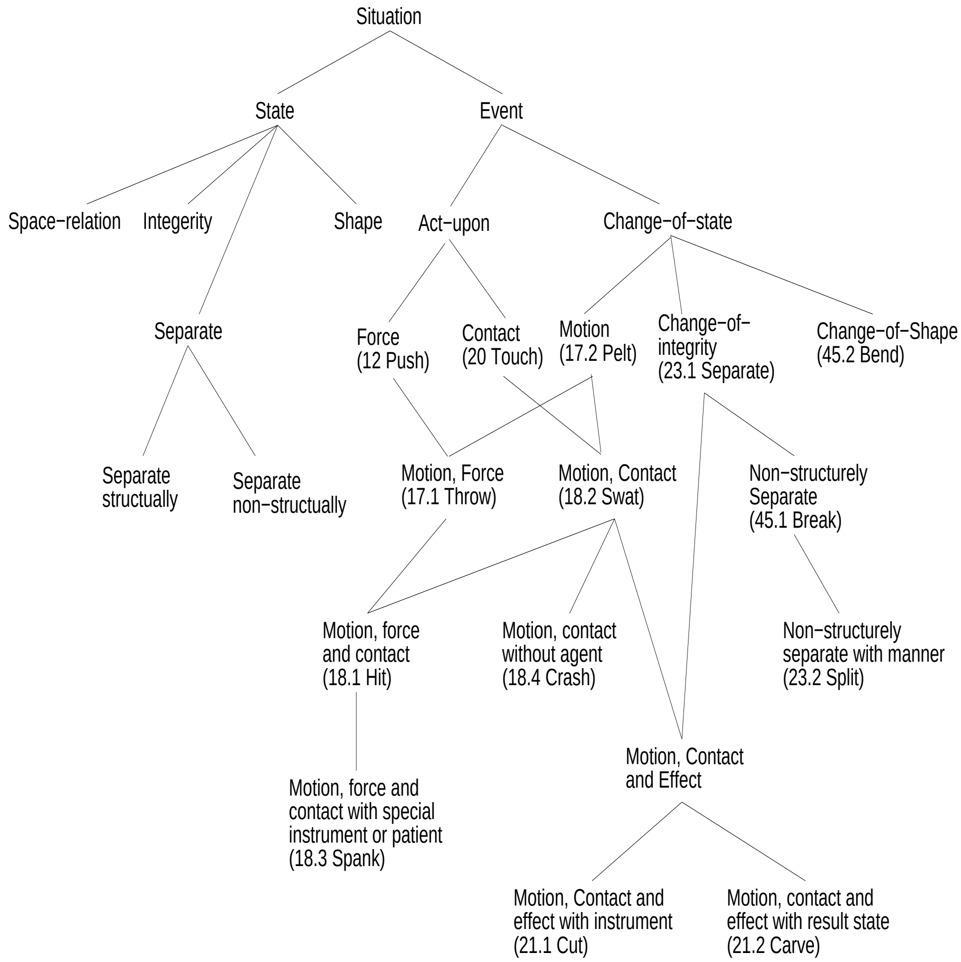



Figure 8-8: Levin's verb classifications

## 8.8 Examples

The following is a set of examples to show how the system works. First a set of symbols is set as instances of the entities. They are:

```
(setq branch-1 (make-instance '%branch))
(setq hand-1 (make-instance '%hand))
(setq man-1 (make-instance '%human-male))
(setq price-1 (make-instance '%price))
(setq top-line (make-instance '%value))
(setq barrier-1 (make-instance '%barrier))
```

(setq lang-barrier (make-instance '%lang-barrier))

### 8.8.1 Semantic disambiguation

The example sentence is "The branch broke", the syntactic analysis result is (break branch-1). Here the 'branch-1' has been set as one of the instances in the entity class %branch. In the UNICON system, eight BREAK usage types have been identified. They are classified according to (Palmer and Polguère, 1992). The prototype sentences for each of the usage types are shown below:

**separated** Some physical object is separated.

- **BREAK-I-1A** The branch broke.
- **BREAK-I-1B** Hail stones broke the roof.
- **BREAK-I-1C** John broke the table with a hammer.
- **BREAK-I-1D** The rocket broke into two parts.

**discontinue** Some continuous event becomes discontinuous.

- **BREAK-I-2** He broke the song with a solo.

**non-functional** Some devices lost their functionalities.

- **BREAK-II-1A** His watch broke.
- **BREAK-II-1B** The fall broke the watch.
- **BREAK-II-1C** He broke the paper drum.

There is a set of preconditions for each of these usage types. They are:

**BREAK-I-1A** ((%IS-A %PHYSICAL @VAR1))

**BREAK-I-1B** ((%IS-A %NATURE-FORCE @VAR1) (%IS-A %PHYSICAL @VAR2))

**BREAK-I-1C** ((%IS-A %ANIMATE @VAR1) (OR (%IS-A %PHYSICAL @VAR3) (%PART-OF @VAR3 @VAR1)) (%IS-A %PHYSICAL @VAR2))

**BREAK-I-1D** ((%IS-A %SEPARATE-STATE @VAR2) (%IS-A %PHYSICAL @VAR1))

**BREAK-I-2** ((%IS-A %ANIMATE @VAR1) (%IS-A %EVENT @VAR3) (%IS-A %CONTINUOUS-EVENT @VAR2))

**BREAK-II-1A** ((%IS-A %MECHANICAL-DEVICE @VAR1))

**BREAK-II-1B** ((%IS-A %NATURE-FORCE @VAR1) (%IS-A %MECHANICAL-DEVICE @VAR2))

**BREAK-II-1C** ((%IS-A %ANIMATE @VAR1) (OR (%IS-A %PHYSICAL @VAR3) (%PART-OF @VAR3 @VAR1)) (%IS-A %MECHANICAL-DEVICE @VAR2))

In the disambiguation process, the variables in the preconditions are set as the arguments in the argument structure. Some of them are set as @unk that means 'unknown'. In this particular example, @var1 is set as 'branch-1'. Other variables are set @unk. By using the function %is-a discussed in the last section, the degrees of precondition satisfaction for each usage type are:

| I-1A | I-1B | I-1C | I-1D | I-2 | II-1A | II-1B | II-1C |
|---|---|---|---|---|---|---|---|
| 3/7 | 0 | 0 | 3/7 | 0 | 0 | 0 | 0 |

Here 3/7 is the evaluation result of (%is-a %physical branch-1). In the UNICON dictionary, %physical is a supertype of %branch. There are four other types on the path from %branch to %physical. The reason for BREAK-I-1D having an evaluation result 3/7 is that when the @var2 does not exist, the BREAK-I-1D get the same similar measure as BREAK-I-1A.

### 8.8.2 Resolving linguistic incompatibilities

This example shows how the system resolves linguistic incompatibilities among different languages. If a verb in one language is given a precise definition, sometimes it will be hard to find a compatible verb in another language having the same definition. What the system can do is to find the most similar verb instead. The translation of the English sentence 'The branch broke' is taken as an example.

UNICON takes the argument structure of this sentence (BREAK branch-1) as the input. Processed by the SEMDIS module, 'BREAK-I-1A' is found to be the usage type the sentence falls into. After instantiating the variables in the 'BREAK-I-1A' definition, The internal meaning representation is:

Initial Concepts: ((TEMPORAL-RELATION (@T0 @D1))
(@D1 (%SEPARATE BRANCH-1)))

The task of lexical selection is to find a proper Chinese verb which realizes the internal meaning representation well. However, in Chinese, there is no single verb that carries exactly the concept of '%SEPARATE'. There is a set of verbs that describes the separation state of concrete objects, such as:

- 断 duan: Separated into several parts, each part is in line-segment shape.
- 破 po: Separated into several parts with irregular shapes.
- 散 san: Separated according to its structure.
- 碎 sui: Separated into pieces.

The separation states identified by these verbs are the specific states of the super state '%separate'. So in the UNICON dictionary, they have been organized into a hierarchical structure. When a concept cannot find a word to realize it, its nearby concepts are chosen as the candidate concepts for selecting the word. In this example, at the first loop, no word has been found in Chinese as the surface realization of candidate concept '%separate'. Then the candidate concepts are changed into its nearby concepts in the hierarchy. They are:

Concepts on path: (%SEPARATE-IN-PIECES-STATE
%SEPARATE-IN-NEEDLE-LIKE-STATE

%SEPARATE-IN-DUAN-STATE

%SEPARATE-IN-PO-STATE

%SEPARATE-IN-SHANG-STATE

%SEPARATE-IN-FENSUI-STATE

%STATE)

Concepts indexed by arguments: NIL

For each of these candidate concepts, the Chinese lexemes can be identified. For example, the concept '%separate-in-duan-state' has five Chinese lexeme realizations. They are:

断了 (duan le, to separate into line-segment shape).

打断-1 (daduan, to hit and separate the object into line-segment shape) .

断成 (duancheng, to separate into line-segment shape into ).

折断 (zheduan, to bend and separate into line-segment shape with human hands)

刮断(guaduan, to separate into line-segment shape by wind force).

In the definitions of these Chinese lexemes, there are preconditions and concepts. The UNICON system then does an inexact match between the concepts and the internal meaning representation. The variables in the precondition are replaced by the arguments and the precondition is evaluated to see whether the verb can have that meaning. In this example, the verbs have the following definitions:

duan la:

```
    (precondition (value ((equalp (%get-feature %shape @var1)
                                   %line-segment)
                          (%is-a %physical @var1))))
    (conception (value ((temporal-relation (@d0))
                        (@d0 (%separate-in-duan-state @var1)))))
```

da duan-1:

```
    (precondition (value ((%is-a %animate @var1)
```

```
                    (not (%is-a %hand @var3))
                    (%is-a %tool @var3)
                    (%is-a %physical @var2)
                    (equalp (%get-feature %shape @var2)
                            %line-segment))))
  (conception (value ((temporal-relation (@d0 @t0 @d1))
                    (@d0 (%use @var1 @var3))
                    (@t0 (%hit-action @var3 @var2))
                    (@d1 (%separate-in-duan-state @var2)))))
```

The result of the inexact match is:

| | 断了 | 打断-1 | 断成 | 折断 | 刮断 |
|---|---|---|---|---|---|
| | duanle | daduan-1 | duanchen | zheduan | guaduan |
| conception | 6/7 | 0 | 0 | 0 | 0 |
| Precondition | 3/14 | 0 | 3/7 | 0 | 0 |

The verb 断了 is taken as the word that best matches the internal meaning representation. The similarity between %separate and %separate-in-duan-state is 6/7 in the UNICON system.

### 8.8.3 Meaning extensions

The execution of the EXTEND module can be shown in the following two examples.

#### Forming new usage types

After an inference is made based on the current knowledge base, a new meaning that is different from any of the existing usage types is assigned to that usage. The usage then forms a new usage type in the dictionary. Let us take an English sentence 'The price hit the top line' as an example. The argument structure is: (HIT @UNK PRICE-1 TOP-LINE).

In the English dictionary, HIT is defined as:

```
(defclass hit ()
    (vocable (value hit))
    (semantics (value (%hit-action @var1 @var2 @var3))))
(defclass %hit-action (%act-on)
    (%lexeme (%chinese (打断-1 打破
                 打碎 打坏))
        (%english (hit)))
    (precondition (value ((%is-a %animate @var1)
                 (%is-a %physical @var2)
                 (%is-a %physical @var3))))
    (conception (value ((temporal-relation (@d0 @d1))
                 (@d0 (%act-on @var1 @var3))
                 (@d0 (%move-toward-in-space @var2 @var3))
                 (@d1 (%contact-in-space @var2 @var3))
                 (@d1 (%receive-force @var2 @var3))))))
```

Since price-1 and top-line are not physical objects, the argument structure is excluded from the HIT usage type.

```
UNICON> (FIND-CONCEPT '(HIT @UNK PRICE-1 TOP-LINE) 4)
Related lexemes are: (HIT)
preconditions are: ((%IS-A %ANIMATE @VAR1) (%IS-A %PHYSICAL @VAR2))
lexeme is: HIT match-degree= -5/7
Matched lexeme is: HIT
```

If the system does not have any knowledge about PRICE, there will be nothing it can do with this newly input usage. However, if the system has the following knowledge, the system can make a guess as the correct solution:

(defclass %price (%quality)

(%lexeme (%chinese 价格)

(%english price))

(situation-index (value (%contact-in-value %move-toward-in-value))))

In the definition of PRICE, the situations in which the PRICE is most likely to be involved in are indexed by 'situation-index'. There should be a lot of them. Here the two situations related to the example are listed. They are:

- %move-toward-in-value: Increasing or decreasing the price to one value.
- %contact-in-value: The price holds a fixed value.

Suppose in the UNICON dictionary, these concepts and the concepts in the HIT definition have the following relationship:

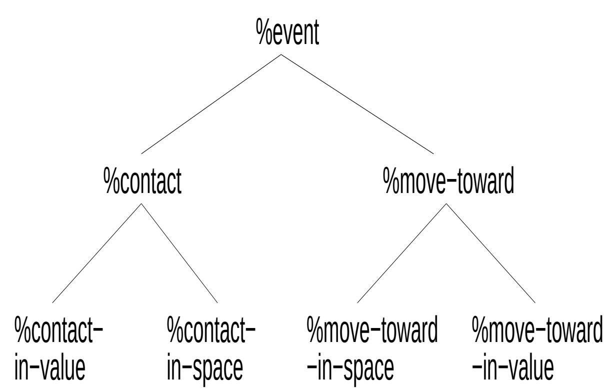



Figure 8-9: The relationship among MOTION concepts

Since the two situations %move-toward-in-value and %contact-in-value are the closest concepts to the concepts defined by the HIT usage type, they compose the most reasonable meaning of the input sentence. In this case, if this kind of inputs is not rare, a new usage type can be formed.

The preconditions of the new usage type should be the preconditions of these two new concepts.
Preconditions are: ((%IS-A %QUALITY @VAR1) (%IS-A %VALUE @VAR2))

The concept set of this new usage type are constructed by removing the unrelated parts in the old usage type and putting in the new concepts.
Conceptions are: ((TEMPORAL-RELATION (@D0 @D1))
(@D0 (%MOVE-TOWARD-IN-VALUE @VAR1 @VAR2))
(@D1 (%CONTACT-IN-VALUE @VAR1 @VAR2)))

**Adjusting the existing usage types**

In another situation, some of the constant values in the precondition of that usage type must be relaxed, as well as enlarging the scope of the concepts identified by this usage type. This enables UNICON to justify its own word definition. Let us take an English sentence 'The man breaks the language barrier' as an example. The argument structure is: (BREAK MAN-1 LANG-BARRIER).

In the SEMDIS module, the most similar usage type is BREAK-II-1C. The relations between the usage types and the input usage can be found in the system. They are shown in Figure 8-10.

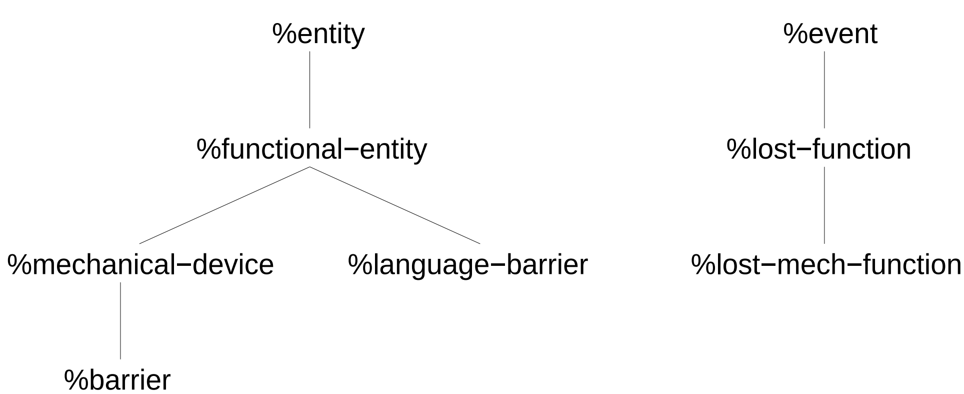


Figure 8-10: Relations among FUNCTIONAL concepts

The above knowledge provides evidence for the system to relax the selectional restriction from %mechanical-device to %functional-entity and to substitute the concept %lost-mech-function by its supernode. The usage type then covers the input usage properly.

```
UNICON> (FIND-CONCEPT '(BREAK MAN-1 LANG-BARRIER))
Old restriction is: (%IS-A %MECHANICAL-DEVICE @VAR2)
New restriction is: (%IS-A %FUNCTIONAL-ENTITY @var2)
Old conception is: ((TEMPORAL-RELATION (@D0 @T0 @D1))
                    (@D0 (%USE @VAR1 @VAR3))
                    (@T0 (%ACT-ON @VAR3 @VAR2))
                    (@D1 (%LOST-MECH-FUNCTION @VAR2)))
New conception is: ((TEMPORAL-RELATION (@D0 @T0 @D1))
                    (@D0 (%USE @VAR1 @VAR3))
                    (@T0 (%ACT-ON @VAR3 @VAR2))
                    (@D1 (%LOST-FUNCTION @VAR2)))
```

## 8.9 Experimental results

We have let UNICON run over 400 sentences in Brown corpus. The experiment involves 17 English verbs. Because UNICON is a lexical selection package, it takes the verb argument structure as its input. So once the sentence is chosen, the verb arguments are manually identified, and the verb argument structure is passed along the input to the system. One hundred sentences that contain these verbs and take concrete objects as the patients were used as the training set. The 17 English verbs in these 100 training sentences were translated into 40 Chinese verbs. All these have been encoded in the UNICON dictionary. The following is the list of these 17 English verbs and the number of the sentences.

| English verbs | Number of sentences |
|---|---|
| bang | 4 |
| beat | 6 |
| break | 8 |
| bump | 1 |
| crack | 7 |
| crush | 4 |
| cut | 13 |
| hit | 19 |
| nudge | 1 |
| pat | 4 |
| scratch | 7 |
| shatter | 2 |
| smash | 4 |
| snap | 1 |
| strike | 6 |
| stroke | 1 |
| touch | 12 |

Our 300 testing sentences are divided into two sets. Test set one has 154 sentences that are carefully chosen to make sure the verb takes a concrete object as its patient. Two tests were run on test set one. In the first test, we do not encode the meaning of the unknown verb arguments. For example, if the argument is 'sportsman' and it has not been encoded in the dictionary, all the system knows is that it is an entity. The system does not know sportsman is a kind of human. For the 154 sentences, the experimental result is as follows:

| | Correct | Wrong | No output |
|---|---|---|---|
| Number: | 89 | 14 | 51 |
| Rate: | 57.8% | 9.1% | 33.1% |

In the second test, the unknown verb arguments were given meanings. In the system, this is an instance of a concept in the UNICON conceptual hierarchy. The experimental results are as follows:

| | Correct | Wrong | No output |
|---|---|---|---|
| Number: | 153 | 1 | 0 |
| Rate: | 99.45% | 0.65% | 0 |

From these experiments, we can see that UNICON system behaves quite well for translating verbs with concrete objects as arguments. The conceptual hierarchy plays an important role in the system performance. Once the unknown English words are defined with the concepts in the hierarchy, the correct rate is improved a lot.

Our test set two contains 146 randomly selected sentences in the Brown corpus. In 30 out of 146 sentences, the English words such as 'break' behaved as nouns. These 30 setences are removed from the test. So we have 116 sentences left for the experiments. These 116 sentences include sentences with non-concrete objects, metaphors, etc. Three tests were performed. The first test is testing with unknown English words. The result is as follows:

| | Correct | Wrong | No output |
|---|---|---|---|
| Number: | 36 | 12 | 68 |
| Rate: | 31% | 10.3% | 58.6% |

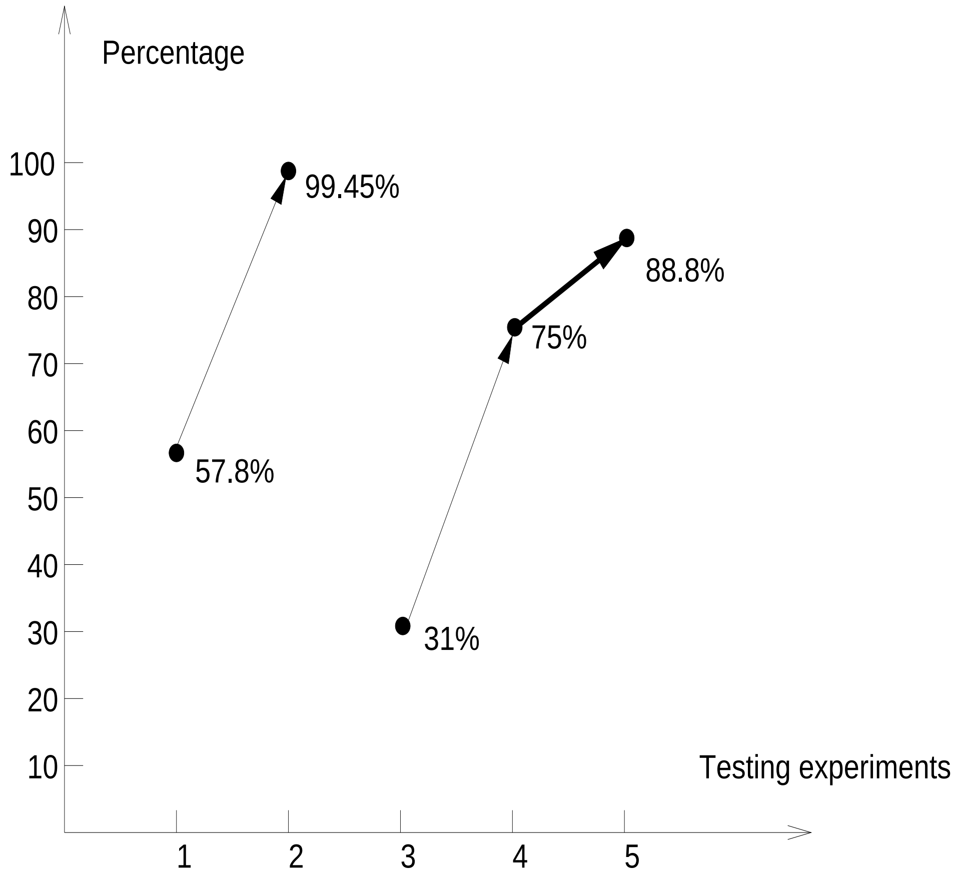
The second test is testing with the encoded English words. The results are as follows:

| | Correct | Wrong | No output |
|---|---|---|---|
| Number: | 87 | 4 | 25 |
| Rate: | 75% | 3.4% | 21.6% |

The third test is the extended selection testing. If a sentence yields no output in the second testing, the system will try to relax the constraints in the sense definition or to try to find out metaphorical sense for the input sentence. The process explained in the last section was performed. The experimental results are as follows:

| | Correct | Wrong | No output |
|---|---|---|---|
| Number: | 103 | 4 | 9 |
| Rate: | 88.8% | 3.4% | 7.8% |

From these tests, we can see that because the verbs are defined on several cognitive domains, the conceptual hierarchical structure provides a way for measuring the similarities among different verb senses, and relaxation, and metaphorical processing becomes possible. The correct rate is improved 13.8% by using this extended selection process. This is shown in the following figure.




1. Test set one , before encoding unknown arguments.
2. Test set one, after encoding unknown arguments.
3. Test set two, before encoding unknown arguments.
4. Test set two, after encoding unknown arguments.
5. Test set two, after applying extended selection process.

Figure 8-11: Testing result

## 8.10 Summary

UNICON is a small prototype system that illustrates the idea of deep hierarchical verb semantic representation and shows the benefits of this representation scheme. The system has demonstrated the following:

- In this system, the similarities among different word meanings can be measured according to the distances in the hierarchy.

- The structure of the hierarchy makes the inexact match between concepts possible.
- The structure of the hierarchy allows the lexical selection to search the candidate words in a particular order. The order is determined by the similarity between one of the semantic components of the word and the internal representation.
- The hierarchy makes it possible to relax the selection restrictions. Instances in the selection restrictions can be relaxed to their supertypes. This is an automatic justification process.
- The hierarchy provides ways for extending the word meaning to the nearest concepts along the paths. New usage types can be formed in the process.

# Chapter 9

# Conclusion

In this chapter, I will discuss the relations among knowledge, natural language processing, and the lexical selection problem in machine translation. Finally, a summary of the thesis is given.

## 9.1 Knowledge and NLP

### 9.1.1 A problem

First let us consider a problem: A computer system takes an English string as input and outputs a '0' or '1' for the following strings in the table. For all other possible strings, the system does nothing.

| input | d | e | ad | ae | dc | eb | adc | OTHERS |
|---|---|---|---|---|---|---|---|---|
| output | 0 | 1 | 0 | 1 | 0 | 1 | 0 | NO OUTPUT |

We henceforth build a computer module for the problem:

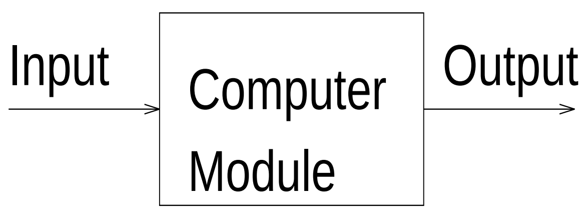



Figure 9-1: A Computer Model

The performance of the module can be evaluated by the *system coverage rate* that is defined as the percentage of the correct output when given a random set of input.

An obvious solution for this problem is a look-up table. It will achieve 100% coverage.

But suppose we have millions of input-output pairs like the following:

(dcc 0) (aadc 0) (aebb 1) (aaeb 1) (aaae 1) ...

A look-up table method would be inefficient. In a rule-based system, we have to generalize on our observation in the following two ways:

- Human intuitive reasoning
- Statistics

By observation, we can have the following rules (set 1):

- Rule 1: If the last character is 'c' or 'd', then output 0.
- Rule 2: If the last character is 'b' or 'e', then output 1.
- else no output.

It is obvious that these rules are not fine-tuned. For instance, it allows 'xyzc' to output '0', but actually 'xyzc' should have no output. The system coverage rate will fall sharply. Then we might have the following rules (set 2):

- Rule 3: If the last character is 'c' or 'd', and the first character is 'a' or 'd', then output 0.
- Rule 4: If the last character is 'b' or 'e', and the first character is 'a' or 'e', then output 1.
- else no output.

This set of rules is better than the previous one, because the system coverage is enlarged. But it is still not the best one.

Only when you know how the relations between the input and the output can be exactly described as if it is by the following finite state machine, can the system can achieve 100% system coverage.

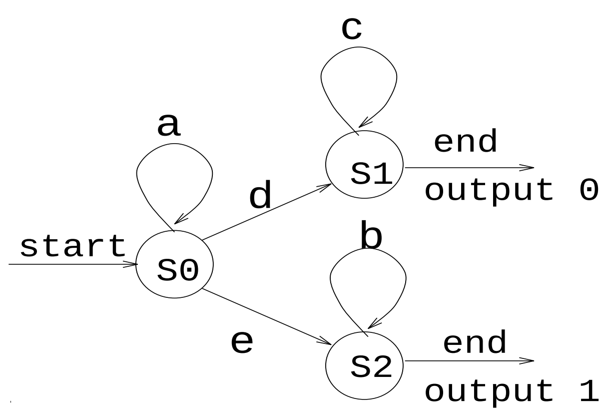



Figure 9-2: A finite state machine

### 9.1.2 Relation between generalization and system coverage

To achieve 100% system coverage, generalization of observation is very important. We must avoid over-generalization and under-generalization. Over-generalization is defined as the generalization that allows illegal input output pairs or cannot handle legal input output pairs. Under-generalization is defined as the generalization that can exactly cover the legal input output pairs, but there is still room for more generalization.

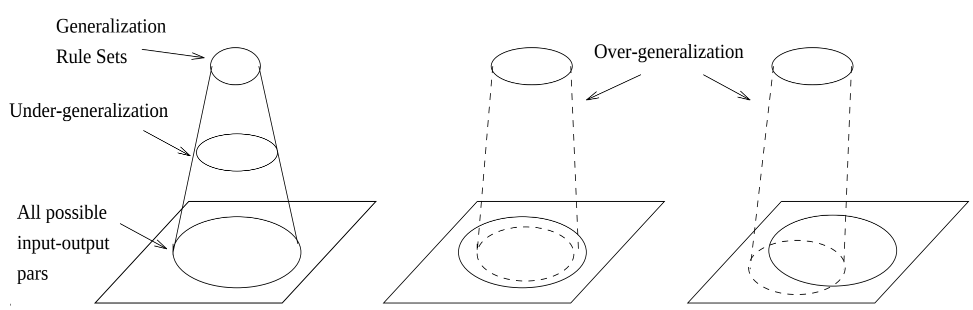



Figure 9-3: Relations between generalization and coverage

### 9.1.3 Relation among generalization, efficiency and coverage

From the following figure, we can see the relations among generalization, efficiency and coverage. Efficiency and generalization are closely related. The more generalization, the more efficient the system. However, because the generalization must avoid under-generalization and over-generalization, the degree of generalization must stop somewhere to guarantee 100 percent system coverage. For a problem that has many idiosyncrasies that are hard to generalize, there will be a tradeoff between efficiency and coverage for the system. If we want high efficiency, over-generalization will be desirable, but lots of idiosyncracies will be out of the system coverage. If the system chooses to take care of each idiosyncracy, the system efficiency will decrease.

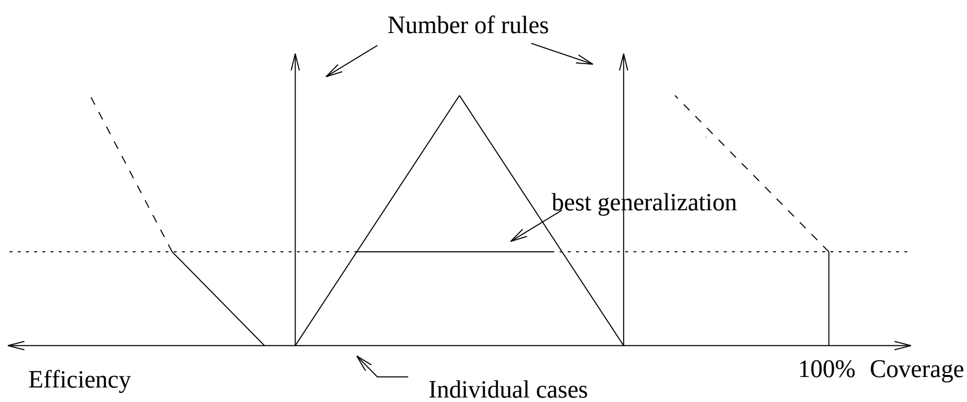



Figure 9-4: Relations among generalization, efficiency and coverage

In a word, in the rule-based system, there is no magic. You have to allow the system to handle each input-output pair, either list them explicitly in the system or cover them implicitly by rules. To achieve as close to 100% coverage as possible, the best generalization method must be identified. Some efficiency must be sacrificed.

### 9.1.4 Knowledge, Generalization and NLP

Our knowledge about natural language and the world is very important. It tells us how to generalize for a NLP problem. The knowledge base in the system is a kind of generalization model that attempts to cover all the input-output pairs. However, natural language is so complicated that it is very easy for humans to do a *over-generalization*. Chomsky is great, but his claim about 'the task of linguists is to find out a grammar which can generate all the English sentences and only the English sentences' was misleading in letting people feel that natural language can be generalized very neatly. The current trend of grammar development gradually converging to lexically based grammars demonstrates that his claim is not true. One has to take care of a lot of idiosyncracies in the lexicon. That is why GB, GPSG, HPSG, LFG, LTAG have been proposed. For the verb semantic representation, it has not reached the best generalization that an NLP system requires. There are still rooms for improvement.

## 9.2 Representation in NLP

From a purely computational point of view, machine translation is the same problem we discussed in Section 9.1. That is, letting the computer generate a output string based on input string. The difference is that the translation problem is so complicated that one cannot even list all the possible input-output pairs. The task of the system designer is to build a model that best generates those input-output pairs. From this point of view, the statistical approach and the rule-based approach are doing the same thing. The difference is that the statistical approach uses a mathematical model to do the generalization, while the rule-based approach builds a human knowledge model for the generalization. Both approaches have difficulty achieving a 100% coverage. This is because too much idiosyncracy in the language makes generalization difficult. For lexical selection problem, besides listing every possible translation pair in the dictionary, the statistical approach attempts to select the best target word based on the training probabilities and the words near the source word. The performance of the system is influenced by the size of the training corpus and the language model.

For the rule-based approach, a verb entry defines a class of verb usages. There might be some usages that are outside the coverage of these verb entries.

If the system can guess the meaning of these usages, the system coverage will be improved. So a good verb meaning representation and organization are important for system coverage. If these verbs are defined with different symbols, and there is no relation among these symbols, it will be hard for the system to measure the similarities among senses. Therefore it is hard for the system to guess a meaning. That is why we propose in this thesis the organization of these meaning definitions into a hierarchical structure. On the other hand, verb meanings may have different facets. Many verb meanings are composed of concepts each of which is related to different cognitive domain. In order to represent the relations clearly, we propose the representation of verb meaning in a multi-domain style.

## 9.3 Summary

Now we come to the summary of the thesis. We have first presented a set of experiments to show that from English to Chinese translation, some English verbs may have very large target verb translations. Simply listing the translation pairs in the bilingual dictionary is inefficient. The Chinese serial verb compound is also analyzed to show that linguistic incompatibility is important consideration in the translation process. We then look into several MT approaches, and attempt to show that none of them has taken the problem of linguistic incompatibility and handling of new usages seriously. In order to handle these problems, We first review some of the previous theories about verb semantic knowledge organization and representation. In Literature Review I, we show that the hierarchical organization of lexical knowledge is important and should be applied to verb semantic representation. In Literature Review II, we show that the generative semantics only takes care of traditional syntactic semantic correspondence. The traditional representation method is static and not adaptive to the inputs outside the coverage. We then suggest a new solution for handling linguistic incompatibility and new usage problems. We propose to have a dynamic verb semantic representation. The verb semantic is more adaptive to the new input. In Theory I, we attempt to show that the correspondence between semantics and interpretation is important. This must be addressed in the verb semantic representation scheme. In Theory II, we attempt to show that verb meanings may be composed of different facets, i.e., different concepts in different domains. Only when these concepts have been defined clearly in the system, is it possible for the system to handle the problem of linguistic incompatibility and new usages. Finally, a prototype system UNICON was built to demonstrate that by using this scheme, the problem of linguistic incompatibility and new usage can be properly handled. The experimental results are given. The experimental system UNICON only contains a very small number of verbs in English and Chinese. Future work will be to investigate how to scale up the system.

# Bibliography


Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Lafferty, J., Mercer, R., & Roossin, P. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2).

Brown, P., Pietra, S. D., Pietra, V., & Mercer, R. (1991). The Mathematics of Machine Translation: Parameter Estimation. to appear.

Choi, S. & Bowerman, M. (1991). learning to express motion events in English and Korean: The influence of language-specific lexicalization patterns. *Cognition*, 41.

Chomsky, N. (1956). *Syntactic Strcutures*. Mouton.

Church, K. (1992). Current Practice in Part of Speech Tagging and Suggestions for the Future. In Simmons, editor, *Abornik praci: In Honor of Henry Kucerau*. Michigan Slavic Studies. to appear.

Cruse, D. A. (1993). Towards a theory of polysemy. manuscript.

Daelemans, W., Smedt, K. D., & Gazdar, G. (1992). Inheritance in Natural Language Processing. *Computational Linguistics: Special issue on Inheritance: I*, 18(2).

Dong, Z. D. (1987). Language, Translation, Machine Translation. In *Proceedings of International Forum on Today's Translation*, Hong Kong.

Dorr, B. J. (1990). *Lexical Conceptual Structure and machine Translation*. PhD thesis, MIT.

Dowty, D. (1991). Thematic Proto-roles and Argument Selection. *Language*, 67(3).

EVANS, R. & GAZDAR, G. (1989a). Inference in DATR. In *Proceedings of the Fourth Conference of the European Chapter of the Association for Computational Linguistics*, London.

EVANS, R. & GAZDAR, G. (1989b). The semantics of DATR. In *Proceedings of the Seventh Conference of the Society for the study of Artificial Intelligence and Simulation of Behaviour*, London.

FILLMORE, C. (1980). The Case for Case. In BACH & HARMS, editors, *Universals in Linguistic Theory*. Holt, Rinehartm and Winston, New York.

FLICKINGER, D. & NERBONNE, J. (1992). Inheritance and Complementation: A Case Study of *Easy* Adjectives and Related Nouns. *Computational Linguistics*, 18(3).

FODOR, J. A. (1970). Three reasons for not deriving "kill" from "cause to die". *linguistic inquiry*, 1.

FODOR, J. A., GARRETT, M., WALKER, E., & PARKES, C. (1980). Against Definitions. *Cognition*, 8.

GEERAERTS, D. (1988). Cognitive grammar and the History of Lexical Semantics. In RUDZKA-OSTYN, B., editor, *Topics in Cognitive Grammar*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

GROPEN, J., PINKER, S., HOLLANDER, M., & GOLDBERG, R. (1991). Affectedness and direct objects: The role of lexical semantics in the acquisition of verb argument structure. *Cognition*, 41.

HSU, L. S., TAN, C. L., & WU, Z. (1992). Generating a Corpus for Analysis. In *1992 International Conference on Computer Processing of Chinese and Oriental Language*.

HSU, L. S., TAN, C. L., & WU, Z. (1993). Handling real world input by abduction. In *Proceedings of 1st Pacific Association for Computational Linguistics Conference*, Vancouver, British Columbia, Canada.

HSU, L. S. & WU, Z. (1992a). Chinese character prediction by recurrent network. *Journal of Computer Processing of Chinese & Oriental Languages*, 6(2).

HSU, L. S. & WU, Z. (1992b). Input pattern encoding through generalized adaptive search. In *Proceedings of the first international workshop of combination of ganetic algorithms and neural network*.

HSU, L. S. & WU, Z. (1994). Translating English 'change-of-state' verbs into Chinese serial verb compounds. In *Proceedings of International Conference on Computer Processing of Chinese and Oriental Language*, Korean.

JACKENDOFF, R. (1972). *Semantic Interpretation in generative Grammar*. MIT Press.

JACKENDOFF, R. (1990). *Semantic Structures*. MIT Press.

JACKENDOFF, R. (1991). Parts and boundaries. *Cognition*, 41.

KILGARRIFF, A. (1992). Polysemy. Technical Report CSRP 261, School of Cognitive and Computing Sciences, The University of Sussex.

LANGACKER, R. W. (1988). A view of Linguistic Semantics. In RUDZKA-OSTYN, B., editor, *Topics in Cognitive Grammar*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

LEVIN, B. (1977). *Predicate-Argument Structures in English*. MA Thesis Proposal.

LEVIN, B. (1987). Approaches to Lexical Semantic representation. In LEVIN, B., editor, *Readings for Lexical Semantics*. Northwestern University.

LEVIN, B. (1992). English Verb Classes and Alternations: A Preliminary Investigation. Technical report, Department of Linguistics, Northwestern University, 2016 Sheridan Road, Evanston, IL 60208.

LEVIN, B. & RAPPOPORT, M. (1991). Wiping the slate clean: A lexical semantic exploration. *Cognition*, 41.

LONGMAN (1978). *Longman Dictionary of Contemporary English*. Longman.

MO, R.-P. (1992). A Conceptual Structure for Parsing mandarin – Its Frame and General Applications. Technical Report TR No. 92-04, Chinese Knowledge Information Processing Group, Institute of information Science, Academia Sinica, TaiWan.

NIRENBURG, S., CARBONELL, J., TOMITA, M., & GOODMAN, K. (1992). *Machine Translation: A Knowledge-Based Approach*. Morgan Kaufmann Publishers.

PALMER, M. (1990a). Customizing verb definitions for specific semantic domains. *machine Translation*, 5(30).

PALMER, M. & POLGUÈRE, A. (1992). A preliminary lexical and conceptual analysis of BREAK: a computational perspective. To be published.

PALMER, M. & WU, Z. (1993). When to stop analyzing a lexical meaning? In *Position paper on Lexical Semantics, General Lexicography and MT Lexicography for workshop on Machine Translation at Utrecht, The Netherlands.*

PALMER, M. S. (1990b). *Semantic Processing for Finite Domains*. Cambridge University Press.

PINKER, S. (1991). *Learnability and Cognition*. The MIT Press.

PUSTEJOVSKY, J. (1991a). The Generative Lexicon. *Computational Linguistics*, 17(4).

PUSTEJOVSKY, J. (1991b). The syntax of event structure. *Cognition*, 41.

RUHL, C. (1989). *On Monosemy: A Study in Lexical Semantics.* Academic Press.

SIMMONS, R. F. (1973). Semantic networks: their computation and use for ud-erstanding English sentence. In SHANK & COLBY, editors, *Computer Models of Thought and Language.* W. H. Freeman.

SISKIND, J. M. (1992). *Naive Physics, Event Perception, Lexical Semantics, and language Acquisition.* PhD thesis, MIT.

TAN, C. L., HSU, L. S., & WU, Z. (1992). On Self-organized approaches to NLP. Technical report, Department of Information System & Computer Science.

WIERZBICKA, A. (1992). Back to definition: Cognition, semantics, and lexicography. In *Asia International Lexicography Conference*, Manila, Philippines.

WINOGRAD, T. (1972). *Understanding Natural Language.* Academic Press, New York.

WU, D. (1991). *On Serial verb Construction.* PhD thesis, Department of Information and Computer Science, University of Maryland.

WU, Z., HSU, L. S., & TAN, C. L. (1992). A Survey on Statistical Approaches to Natural Language Processing. Technical Report TRA4/92, Department of information system and computer science, National Universoty of Singapore. Submitted to *computational lingistics.*

WU, Z. & PALMER, M. (1994). Verb Semantics and Lexical Selection. In *Proceedings of ACL94*, New Mexico State University, USA.

Wu, Z., Palmer, M., Hsu, L. S., & Tan, C. L. (1993). Toward the Similarity-based Fuzzy Lexicon. Technical Report TRE4/93, Department of Information System and Computer Science, National Universoty of Singapore.