AN INVESTIGATION INTO THE USE OF ARGUMENT STRUCTURE AND LEXICAL MAPPING THEORY FOR MACHINE TRANSLATION

by

SHUN HA SYLVIA WONG

A thesis submitted to the
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
The University of Birmingham
September 1999

Abstract

In recent work on the Lexical-Functional Grammar (LFG) formalism, argument structure (a-structure) and lexical mapping theory have been used to explain many linguistic behaviours across languages. It has been suggested that the combination of c-structure, f-structure and a-structure might form a suitable architecture for Universal Grammar. If this suggestion is valid, the LFG formalism would be a suitable linguistic model for Machine Translation (MT). This thesis reports on the investigations carried out on using a-structure and lexical mapping theory for aiding various sub-tasks in MT. The two investigations described in this thesis are the abilities of a-structure and lexical mapping theory to: (1) aid different kinds of lexical and structural disambiguations involving verbs and prepositions, and (2) act as a suitable medium for carrying out source-to-target language transfer. Based on the results of these investigations, this thesis also gives an evaluation of how well a-structure and lexical mapping theory can improve the existing models of linguistic-based MT.

Acknowledgements

I would like to express my gratitude to the following people and institutions:

Dr. Peter Hancox, my supervisor, for his supervision, encouragement and for inspiring me to initiate this research,

My parents, the School of Computer Science and the Committee of Vice-Chancellors and Principles of the Universities of the United Kingdom (CVCP) for their financial support,

Michal Konečný, my husband, for his emotional and technical support,

The anonymous reviewers of my papers from PACLIC 12, AICS'98 and PACLIC 13 for their invaluable comments,

Dr. Ela Claridge and Dr. William Edmondson, members of my thesis group, for their advice and encouragement,

Dr. Alex Alsina for his advice on a-structures,

and finally, Keith Marlow and all the research students in the School of Computer Science, University of Birmingham between the years 1995–1999 for their friendship and for listening to my moans.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cross-regional business and cultural interchange have had a continuous growing importance in the history of mankind. However, due to the fact that the fundamental means of communication (i.e. natural language) among mankind differs across regions, communication between people coming from different parts of the world has always been a difficult activity. One way to ease this difficulty is to learn the language used in the communicating region before starting the communication process. However, language learning is difficult and time-consuming. If one would like to read one article written in a foreign language only, it is quicker and more cost-effective to get an expert in both languages to rewrite the article using his / her own words. This process is called translation.

Translation has been regarded by translators as a repetitive, monotonous and thus boring job. Since translation requires a profound knowledge of two languages, people who are qualified for this job are rare and thus there has been shortage of translators. This raised the idea that if there is an international language which is not ambiguous and can be understood by every human being, there would be no need for translation. Since the 17th century, much effort was put in developing this kind of unambiguous international language. However, as time went by, though many different proposals of an international language were developed (e.g. Esperanto (Hutchins & Somers 1992)), they were inadequate to achieve their aim. With the advance in technology and the development of digital computers, researchers then realised that if the translation process could be automated, the problem of the lack of translators would be solved. Therefore, in early 1950s, researchers started to work extensively on automated translation, i.e. Machine Translation (MT).

After 40 years of work on MT, there is still no very fruitful result in this field which would adequately support the translation of highly ambiguous texts. Though some successful systems have been invented (e.g. Systran, Météo), their success was mainly due to their limited applicability. One difficulty of MT lies in the fact that different languages use different ways to express the same meaning (e.g. in terms of different usage of words, and of different phrase and

sentence structures). An idea which can be expressed easily in one language might require a complicated construction to be expressed in another language. For example, something which can be expressed in one word in one language sometimes can only be expressed by the use of several words in another language. Moreover, in some cases, even the use of a long list of words cannot express the original meaning precisely; as a result, only an approximation of the original meaning can be expressed in the other language.

One might say that natural language is a system for naming and describing everything that exists and happens in the world (e.g. a state, a physical or abstract object and an action) by the use of words. Therefore, different natural languages should be different representations of nearly the same range of meanings. However, knowing a dictionary-style description of an object does not necessarily mean that the properties of the object can be fully realised. In order to fully understand the actual object behind the description, one needs to have some idea about the world. For instance, if one is to describe the object *'cloud'*, he / she can say:

> *"It is a mass of water vapour that floats in the sky and it is usually white or grey in colour."*
> [Collins COBUILD English Language Dictionary (Sinclair 1987)]

However, without looking at the actual object before, it is still difficult for the listener to understand what a cloud is. Therefore, without knowing the underlying meaning of a description (e.g. a sentence, a word or a phrase), it is very difficult to map the description from one language to another appropriately.

In addition, natural languages and their use are highly irregular. There exists implied and even inverted (e.g. in sarcasm) meaning in a word, phrase or sentence depending on which circumstances they are used in. This kind of hidden meaning is so difficult to pursue that even a native speaker might fail to understand it in certain circumstances. The lack of knowledge about the actual meaning of words, phrases and sentences defined in the language domain makes it very difficult for computer to produce a good piece of translation. As a computer is a machine which is good at exact symbolic processing but not as effective in terms of representing and processing multi-dimensional information like a human brain, it is difficult to program a computer to translate the meaning of a piece of text from one language to another without any distortion.

Although MT is not an easy task, it is not impossible and is worth pursuing. The fact that different real-life MT systems have been developed and are continuously under expansion proves that MT is useful to our every-day life and worth pursuing further. The question is *how far MT can go and how this could be achieved.* It is believed that identifying the existing problems of MT and finding ways to solving or alleviating these problems can improve the ability of MT systems to perform real-life MT tasks. These are, in fact, the general aims of this thesis.

1.1 Problems of Machine Translation

Machine Translation (MT) has always been viewed as a difficult computing task. There is even a point of view that MT is impossible and thus a waste of time. In 1964, a US government sponsored group called the Automatic Language Processing Advisory Committee (ALPAC) was formed to evaluate the possibility of MT. It was stated in the ALPAC report: "there is no immediate or predictable prospect of useful MT" (Hutchins & Somers 1992, page 7). However, was this claim really valid, meaning that all researchers should stop any activity on MT?

Since the publication of the ALPAC report, the on-going research and activities in MT carried out in various countries throughout the world apart from the U.S. and the results obtained from them showed that MT is not a waste of time. However, there are currently many problems in the area of MT which make real-life applications of MT quite limited. If these problems can be solved, MT will be more reliable and widely used in different sectors throughout the world. This section aims to unveil some of the major problems in MT. Before we look into the details of some of these problems, we are going to look at an overview on the importance of these problems to MT. This is to help the reader of the thesis to understand the incentive behind this study.

1.1.1 Why are problems in MT vital to the application of real-life MT systems?

One vital problem obstructing good quality MT is the presence of ambiguities both within and between natural languages. As natural languages continue to evolve, the chance for these ambiguities to arise increases continuously. Therefore, it seems to be impossible to solve the problem of ambiguity completely. It is, however, possible to improve the storage and processing methods of current MT systems so as to resolve some of the ambiguities, and, as a result, obtain a higher translation quality. By improving the translation quality continuously, MT can be more reliable to real-life applications, and thus able to ease the work load of human translators more extensively.

Translation has its importance in everyday life, especially when it is used in the business sector or used for producing government publications in bilingual countries. It is more important to the business sector when a company decides to go into international trade. Though up to this moment (and perhaps in the near future), the output translation quality of MT might not be able to support the translation of important documents like business contracts[1] without a considerable amount of human editing, relatively less important and more regular documents like technical and scientific journal articles, user support manual and consumer information about

[1] Due to the inadequacy of MT systems in resolving ambiguities arising in natural languages, inappropriate source-to-target translation will be likely to appear in the resulting output. The misuse of words or phrases in a business contract increases the chance of suffering from misjustice by either party when legal action is taken against the detail agreements stated in the contract.

a product on its packaging can be handled by MT systems with acceptable translation quality. However, the fact that MT systems in general take a very long time to process a medium to long piece of normal text[2] makes businesses quite reluctant to use MT. After shortening the general processing time of MT systems, the overall processing speed of the MT systems will still be acceptable even though the source texts are quite ambiguous and the system needs more time for analysis. As a result, MT can then be employed to cater for real-life translation needs.

1.1.2 What makes MT so difficult?

The problems of MT can be categorised into two major domains: the linguistic domain and the computational domain. The linguistic domain defines the language theories about syntactic, semantic and pragmatic information of words and sentences in each natural language used in the human world. The computational domain embeds the issues concerned with the computer processing involved in MT, e.g. processing speed, representation of word meaning, storage of the dictionary and mechanical analysis of source sentences by computer.

The simplest way to perform MT is to find out the corresponding target language equivalent for each word present in the source text one by one. As shown in Figure 1.1, this method is simple

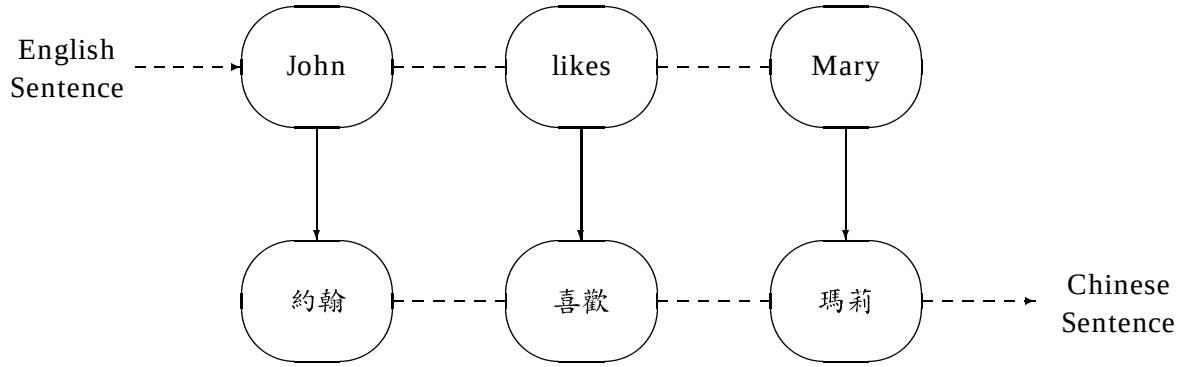


Figure 1.1: A Word-for-Word Translation

and straight-forward: no syntactic analysis is required on the source language sentence and the source-to-target language equivalents obtained build up the required target language sentence without the need of further processing. Though the word-for-word translation method works fine in translating the English sentence "*John likes Mary.*" to Chinese, if this method is used to translate the same English sentence to, say, Japanese, the output sentence obtained will be

[2]The term 'normal texts' here refers to the reports or articles which can be found in everyday life journals. These kind of texts are normally less ambiguous than the literature, e.g. novels or poems, and thus relatively easier to translate by computers.

syntactically incorrect[3]. If an ambiguous source sentence is to be translated by the simple word-for-word translation method, the output translation might seem like a piece of junk text to a target language native speaker, or worse still, convey the wrong meaning. Therefore, a more detailed processing is required for effective MT.

A more effective MT approach contains three parts: source language text analysis, source-to-target language transfer and target language text generation. This approach allows a more thorough analysis of the source language text so as to help resolving the ambiguity within it before the required source-to-target language translations are looked up during the source-to-target language transfer. This method also allows the reorganisation of selected target language words so that the output sentences will conform with the target language grammar. Even though this three-stage MT method can give rise to a better MT output, due to the complexity of natural languages, there still exist many problems which affect the effectiveness of MT systems. The following sections attempt to show briefly how and why the three-stage MT method is inadequate in catering for real-life translation needs.

1.1.3 Linguistic Problems

Natural languages are highly irregular and continually evolving. Due to this nature of natural languages, it seems impossible to program a computer to cope with the processing of the complete domain of any natural language. With continuous research on Natural Language Processing (NLP) since the 1960s, many computational linguistic formalisms have been invented (e.g. General Phrase Structure Grammar (GPSG) (Gazdar, Klein, Pullum & Sag 1985), Categorial Grammar (Wood 1993), Case Grammar (Fillmore 1968), Lexical-Functional Grammar (LFG) (Bresnan 1982*b*) and Head-driven Phrase Structure Grammar (HPSG) (Cooper 1994)). These formalisms enable the syntax (e.g. sentence and phrase structure and word category) and some of the semantic properties of natural language to be successfully represented and processed by computers. However, even though some of these formalisms have attempted to capture the semantic information of language, none of them is adequate for capturing the various properties of words required for solving all kinds of ambiguities. Thus, many problems of semantic processing of words or sentences still remain unsolved.

The major cause of these linguistic problems is the ambiguity of some words and phrases, i.e. a word or a phrase with more than one meaning. Some examples of these words and phrases are homographs, phrasal verbs, idioms in the English language. For instance, homographs are words with the same spelling but different meanings when they are used in different circumstances. Some of the ambiguity caused by this kind of words can be resolved by analysing the

[3] According to the English grammar, the order of appearance for the subject (S), verb (V) and object (O) within a simple sentence is SVO; whereas that in the Japanese grammar is SOV. A direct word-for-word translation from English to Japanese will result in a Japanese sentence with word order in SVO form.

structure of a sentence and find out what syntactic role a homograph is playing in the sentence[4] (cf. Chapters 3 and 5). However, there exist many homographs which are in the same syntactic category, e.g. 'river *bank*' versus 'money *bank*' and the noun 'ball' which can mean a kind of social gathering, as in "John went to a *ball* with Mary.", or a round object for sports, as in "John kicked the *ball* to Mary.". For this kind of disambiguation, a more sophisticated method to analyse the source language sentence is required.

A word can possess more than one meaning and causes problems of ambiguity in an MT system. When words are used in conjunction with each other, even though each of these words possesses only one meaning, they can also become ambiguous to an MT system. For instance, the phrasal verb 'look up', as in "John *looked up* a word from the dictionary.", versus the occurrence of the verb 'look' and the adverb 'up', as in "John *looked up* and prayed.". Some of these ambiguities can be resolved by analysing the syntactic information within the sentence. The raw input of a text-based MT system is a sequence of words. In the English language, this sequence of words are separated by white spaces which makes it easier for an MT system to determine where a word ends and when a new word starts while analysing the linguistic information carried in each sentence. In the Chinese language, however, a sentence is formed by a sequence of words which do not have any white space in between. Most Chinese words are made up of one to four Chinese characters. Without white space to indicate the beginning and end of a word, it is much more complicated for an MT system to determine the structure of a Chinese sentence, let alone to process the detailed linguistic information within it so as to produce a target language translation.

Due to the differences in terms of linguistics and usages between different natural languages, even though a source language sentence is not ambiguous, it can also cause problems in producing the target language sentence. For instance, though, similar to the English auxiliary verb 'be', the Chinese word '被' is commonly used to signify passive in Chinese sentences, this Chinese word is different from the auxiliary verb 'be' both syntactically and semantically. Due to this difference, the translation between English and Chinese passive sentences with ditransitive verbs poses a problem to an MT system. The English sentence "*Mary was told a matter by John.*" does not have an obvious equivalent in Chinese as it is ungrammatical to say[5]:

*	瑪莉	被	約翰	告訴	了	一	件	事情。
	Mary	*bei*	John	tell	ASPECT MARKER	one	QUANTIFIER	matter

The ambiguities arising in a source language text and the difference between the source and target language grammars, ways of expression and understanding, etc. pose many problems to an MT system in translating sentences from one language to another. However, it is believed that

[4]For instance, by knowing the syntactic category of the word 'yank' in the sentence "Don't yank my string.", one can disambiguate the meaning of this homograph (i.e. 'to pull with a lot of force suddenly') from the noun version of this word which means 'American' (as in "John is a yank.").

[5]A more detailed discussion on the problem arising in transferring this kind of sentences will be given in Chapter 6.

Word	Meaning
bug	• small insect • virus / germ • defect of a computer program • small hidden microphone
program	• a set of instructions for controlling the computer operation • work plan • a booklet for displaying the order of a performance
garden	• a piece of land where plants, e.g. flowers, trees and vegetables, are grown

Table 1.1: Different meanings of some nouns

some of these problems can be resolved and at least alleviated if there is a systematic method to process the meaning of a sentence in NLP, or more specifically, MT. However, the existing methods are often either inadequate to alleviate the problem of ambiguity or they are too complicated to be used in real-life MT systems. The aim of this study, therefore, is to investigate a method which is relatively easy to implement and at the same time more adequately aids the required disambiguation process in an MT system.

1.1.4 Meaning Representation

During translation, one way to determine which meaning of a word should be used while resolving ambiguity is verifying the various meanings of the word with the meanings of other words in the sentence and then eliminating the meaning(s) that are impossible in that context. For instance, consider the following sentences:

1. I have a bug in my garden.
2. I have a bug in my program.

Table 1.1 shows the meanings of the nouns in these sentences. After the verification of the different meanings of *'bug'* with *'garden'* and *'program'*, the meaning of 'bug' in the first sentence can either be a small insect or a small hidden microphone; whereas in the second sentence, the noun 'bug' is more likely to mean a defect (as a computing term) or a small insect depending on the meaning of 'program'[6].

With the knowledge of the real world, the above meaning verification is fairly easy to do. For example, a bug (being a defect of a computer program) is an *abstract* computing term which

[6]In British English, the word which means "a booklet for displaying the order of a performance" is spelt as 'programme'. The spelling 'program' tends to be used in American English. In this study, the spellings in both American English and British English are considered.

cannot exist in the domain of a physical non-computing object like a garden. Therefore, 'bug' — as a small insect — is chosen for "*I have a bug in my garden.*". However, without the understanding of the objects in the real world, it is very difficult for a computer to differentiate the different meanings of a word so as to choose an appropriate translation during MT processing. The computer is suited for symbolic processing and pattern matching. However, unless a computer can learn to process abstract descriptions like a human brain and/or to 'experience' the real world like a human being, it is impossible for a computer to perform natural language translation like a human translator[7]. During NLP, the more common way to program a computer to realise the real world is by tagging the various semantic properties of objects and actions for pattern verification. For instance, the noun 'garden' is often tagged with the features: $[-animate]$ and $[+physical]$. Although semantic tagging allows a computer to have more knowledge on word meaning than merely knowing the syntactic information of a word, there does not exist a systematic method for this kind of semantic tagging. If too many different semantic features are used, the resulting system would become very complicated and thus difficult to maintain. If only a couple of semantic features like $[\pm animate]$ and $[\pm physical]$ are used, it would not be sufficient to perform many disambiguations. In addition, the analysis of the tags between different words quite often requires the verification of different combinations of tags, which, as a result, makes it very complicated to program.

1.2 Motivation and Aims of the Research

Theoretical linguists have put in tremendous amount of on-going effort to formalise natural languages. As a result of this on-going effort, many linguistic formalisms which are suitable for computation (e.g. Case Grammar (Fillmore 1968), LFG (Bresnan 1982*b*) and HPSG (Cooper 1994)) have been developed. These linguistic formalisms provide effective guidelines on extracting and representing the linguistic information of sentences. One question that is often asked by computational linguists is: "*Can these results of formalisation be readily applied onto the development of MT/NLP systems?*". Rohrer (1986) seemed to suggest a positive answer to this question:

> "*If linguists can write their grammars in a formalism whose mathematical properties are well understood, then the programmer will have fewer problems implementing the grammar.*"
>
> [Rohrer (1986, page 354)]

Inspired by this idea and the fact that LFG has been used as a linguistic backbone of recent MT research and MT systems, this research studied the ability of using a relatively new exten-

[7] Human beings learn the meaning of words in a multi-media format, i.e. vision, odour, feeling, sound and taste. Though it is possible for a computer to recognise shape and sound, it still does not help a computer to understand the relationship between different objects. Thus the shape and acoustic information about an object is insufficient to help a computer to resolve ambiguity effectively.

sion of the LFG formalism — a-structure and lexical mapping theory — to aid MT processing, especially in aiding different kinds of disambiguation processes.

A-structure and lexical mapping theory are the result of many observations on linguistic behaviours of various natural languages and they have also been used for explaining these different linguistic behaviours, e.g. Bresnan & Kanerva (1989), Alsina & Mchombo (1993), Huang (1993), Bresnan & Zaenen (1990) and Alsina (1996*a*). Bresnan (1994) suggested that a-structure, together with c-structure and f-structure form a suitable architecture for Universal Grammar. Linguistic-based MT systems require ways to capture the similarities between languages so as to facilitate the transfer of sentences between different languages. If this relatively new extension of the LFG formalism (i.e. a-structure and lexical mapping theory) really forms a suitable architecture with the traditional LFG formalism and the fact that LFG has been proven to be suitable for acting as a linguistic backbone for much research on NLP and MT, it would be suitable for aiding the translation of sentences from one language to another, regardless of the language pair. This investigation studied the ability of a-structure and lexical mapping theory to act as a suitable medium for carrying out source-to-target language transfer in MT processing.

A-structure describes the structure of an event in terms of the necessary participants involved (Bresnan 1995). Different event structures should, at least to some extent, correspond to different meanings of the verb (Carlson 1984). As the arguments of an event structure reflect the real-world participants of an event, instead of some abstract representation of sentences, information captured in a-structures should be relatively language independent. Thus this information can be used for differentiating the different senses possessed by a verb. This research investigated the applicability of a-structure and lexical mapping theory in aiding lexical and structural disambiguation involving verbs, prepositions and adverbs.

In terms of computational implications of this investigation, this thesis reports the results of the study of applicability, effectiveness and problems in using a-structure and lexical mapping theory for MT. This thesis also proposes some modifications to the existing a-structure representation as described by Bresnan & Kanerva (1989), Alsina & Mchombo (1993), Huang (1993), Alsina (1996*a*), Bresnan (1994), etc. and additional features to improve the lexical mapping process between a-structure arguments and syntactic functions of sentences. These modifications are made so as to facilitate the disambiguation and transfer processes in MT.

In terms of linguistic implications of this research, this thesis reports the finding that a-structure is **not** universal across languages. This thesis also proposes a new representation of Chinese word '被' in LFG terms which would sufficiently predict and describe the unusual linguistic behaviours of this Chinese word, and thus facilitating the translation of sentences involving this Chinese word. Furthermore, this thesis describes an extension to the existing representation of a-structure which enables a-structure to account for the presence of prepositional phrases in sentences.

1.3 Organisation of this Thesis

Chapter 2 gives a brief introduction to the history of MT. It also briefly describes the characteristics of some of the well-known practical MT systems and gives an evaluation of these systems. Furthermore, this chapter introduces and evaluates some of the contemporary approaches to MT. This chapter aims to give the reader some background knowledge on the research carried out in MT. By introducing and evaluating some of the well-known approaches which dominate the recent research in MT, this chapter also aims at revealing some of the problems in MT which have not yet been solved. If the reader is familiar with the current literature and problems of MT, he/she might like to skip this chapter.

LFG is one of the contemporary linguistic formalisms which is suitable for computation. The LFG formalism forms the backbone of the investigation reported in this thesis. Chapter 3 introduces some of the fundamentals of this linguistic formalism: the representation of linguistic information within sentences in terms of c-structure, f-structure and s-structure and how to construct these structures for a given sentence. LFG has proved to be suitable for supporting the research on NLP and MT. This chapter also reviews and evaluates some of the MT approaches inspired by this linguistic formalism. The aim of this chapter is to allow the reader to have some knowledge on the LFG formalism. Through discussing the strengths and weaknesses of the existing approaches to applying LFG to MT, this chapter also aims at revealing what needs to be done in order to improve the existing methods for MT. Again, if the reader is familiar with the current literature and problems of applying LFG to MT, he/she might like to skip this chapter.

As stated in the title of this thesis, this research investigated into the use of argument structure (a-structure) and lexical mapping theory for Machine Translation. A-structure and lexical mapping theory — being a relatively new extension of the LFG formalism — are the focus of this research study. Chapter 4 reviewed the theory of a-structure and the lexical mapping theory as described by various theoretical linguists (e.g. Bresnan & Kanerva (1989), Alsina & Mchombo (1993), Huang (1993), Bresnan & Zaenen (1990), Alsina (1996*a*) and Bresnan (1994)). It also describes the extension proposed to the existing a-structure representation. With thematic roles forming the basic building blocks of a-structures and the observation that the existing literature on thematic roles does not provide clear and sufficient guidelines to define a universal set of thematic roles, this chapter specifies and illustrates the set of thematic roles and their meanings as used throughout this study. Owing to the similarities between a-structures in LFG and case frames in Case Grammar, a-structures are sometimes being mistaken as a variant of Case Grammar. This chapter also attempts to clarify this confusion.

The investigation carried out in this research focuses on exploring whether a-structure and lexical mapping theory can be used to alleviate some of the unsolved problems in MT and finding out the effectiveness of this method. Chapters 5 and 6 report the results of the investigations carried out in this study.

Looking at the results obtained in this investigation, Chapter 7 concludes that the method to MT proposed in this thesis is effective in alleviating the problem of disambiguation and it also solves the problem in transferring English ditransitive passive sentences to / from Chinese. This chapter also discusses the potential problems arising in the proposed approach to MT and gives the reader an insight into solving these problems. Furthermore, this chapter sketches some possible future developments and future research for this investigation. After reading this thesis, the reader should have gained sufficient knowledge to continue this research work.

Chapter 2

Machine Translation

Long before the term Natural Language Processing (NLP) was invented, there had been work on the use of machines to aid translation. The term *Machine Translation* (MT) refers to the use of a machine for aiding or performing translation tasks involving more than one human language. Bearing this definition in mind, the work on MT was in fact started in the 17th century when the use of mechanical dictionaries were first suggested. The machine translation (MT) 'systems' invented in those days were merely mechanical dictionaries for aiding human translation. The whole translation process very much relied on human effort. Though the MT 'systems' invented in the 17th century were referred to as mechanical dictionaries, they were not aiming at merely providing the meaning of words in the lexicon. They were aiming at forming an unambiguous language based on logical principles and iconic symbols which allow people to communicate with each other without the fear of misunderstanding (Hutchins & Somers 1992). Since then, the research on MT has focused on producing different proposals of this kind of unambiguous languages.

Without the support of a suitable device to perform MT before the invention of digital computers, MT techniques did not advance very far. Much research on MT in those days was word-for-word based, i.e. given a word in one language, the MT 'system' produced an equivalent in another language.

After the invention of digital computers, MT researchers started to explore the possibility of using digital computers for more sophisticated MT. The first discussion of 'real' MT systems using computers was offered by Weaver in 1949 in which he suggested the use of universal features and the underlying logic of language for performing MT. Since then, much research carried out in this area has been focused on developing MT systems for performing automated translation rather than aiming at producing very sophisticated mechanical/electronic dictionaries for aiding human translation. With continuous research on MT since the invention of digital computers, researchers found that the processing power of MT systems could be improved if more sophisticated techniques were used to analyse and process source and target language texts.

As a result, the development of techniques for processing natural language texts, i.e. Natural Language Processing (NLP), became another research focus. In fact, the advances in NLP have helped the advance of MT in recent decades. The various techniques and linguistic formalisms used in NLP improve the analytical power of MT systems in processing source texts. Nowadays, MT is often treated as a subset or an application of NLP research. Though the materials reviewed in this chapter are mainly related to MT, some of the techniques reviewed can also be applied on other disciplines of NLP.

In the light of the problems in MT caused by ambiguous words and / or sentences, this chapter reviews and analyses some of the MT systems developed in the past as well as the more recent approaches to MT. A discussion of some of the strengths and weaknesses of these approaches is also presented in this chapter so as to shed light on how MT can be improved.

2.1 Different Kinds of Ambiguities

If any word within a natural language has only one interpretation (i.e. having one syntactic, semantic and pragmatic analysis), MT would become a fairly simple task. An MT system could obtain the target language translation by simply analysing each word within a source language sentence and generating the target sentence according to the target language grammar. However, this is often not the case with any natural language. A word not only can have more than one interpretation, it can also combine with other constituents within a sentence to form other interpretations. For instance, a word may appear in more than one syntactic category, e.g. the word ‘ships’ can be a noun or a verb (Ingria, Boguraev & Pustejovsky 1992, page 342). A word may combine with other word(s) to form a new lexical unit, e.g. the phrasal verb ‘fish for’ as in “John fished for invitations.”. Even within the same syntactic category, a word can have more than one meaning, e.g. the noun ‘saw’ can mean a tool for cutting, or a short, well-known saying or proverb. The existence of ambiguous words makes it more difficult for an MT system to capture the appropriate meaning of a source sentence so as to produce the required translation. This section briefly discusses some of the different kinds of ambiguities occurring in natural languages which affect the effectiveness of MT systems.

2.1.1 Lexical Ambiguity

Lexical ambiguity is the ambiguity occurring due to the existence of different meanings of a lexical unit. One famous kind of lexical ambiguity is caused by *homographs*. A homograph is a word (i.e. a sequence of characters) with more than one meaning. For instance, the English word ‘saw’ is commonly used as the past tense of the verb ‘see’, but it can also mean a tool for cutting, the action of using this tool for cutting as well as a short, well-known saying or proverb. Some homographs have only one meaning within a single syntactic category. For instance, as a noun, the word ‘minute’ means *a unit for measuring time*; as a verb, it means *to*

make a written record of what is said or decided during a meeting; as an adjective, it means *tiny*. It is relatively easy to disambiguate this kind of homographs. Simply analyse the syntactic category of their occurrence in a sentence, the appropriate meaning can then be obtained. Knowing the syntactic category of a word does not always help the disambiguation of homographs. It is because some homographs have more than one meaning even when they are used in the same syntactic category. For instance, the noun 'ball' can mean a dance party or a round object for sports. With this kind of homographs, one way to disambiguate them is to consider their semantic properties in relation to the semantic properties of other words in the sentence. For instance, the meaning of 'ball' in the sentence "John kicked a ball." must be a round object for sports because the verb 'kick' requires physical contact with a physical object, but a dance party is an abstract event which cannot be *kicked*. Even comparing the semantic properties of words within a sentence does not always help. As pointed out by Hutchins & Somers (1992, page 87), with a sentence like "*When you hold a ball, . . .* " in which both senses of the verb 'hold' (i.e. to grasp and to organise) can be used with the different senses of the noun 'ball', it would be difficult to obtain the appropriate meanings unless the later part of the sentence provides more clue to disambiguate these senses.

The problem of translating homographs is concerned with the analysis of monolingual constituents (i.e. source language words). MT involves the processing of more than one natural language. Sometimes even when a source language word is unambiguous, it may still cause a problem when it is translated to another language. For instance, some ergative verbs in English do not have an exact equivalent in Chinese. Though it is possible to translate them to Chinese, an MT system would need to choose the appropriate translation from two possibilities depending on the source sentence. For instance, when the verb 'sink' subcategorises a subject in a sentence, it is translated to '沉' in Chinese; when it subcategorises both a subject and a object in a sentence, it is translated to '弄沉'. The translation of this kind of English ergative verbs can be supported by syntactic analysis, i.e. checking how many syntactic functions a verb subcategorises. However, for an English verb like 'tell' which is not ergative but it also has more than one Chinese translation (i.e. either '告訴' or '說'), syntactic information alone cannot aid the translation of this verb adequately (cf. Section 5.2).

2.1.2 Structural Ambiguity

Structural ambiguity is concerned with the syntactic representation of sentences. It occurs when more than one syntactic structure can be associated with a sequence of words. For instance, the noun phrase (NP) 'with a telescope' in the sentence "John saw Mary with a telescope." can either be interpreted as part of the NP 'Mary' (as in Mary was holding a telescope) or as a prepositional phrase (PP) of the verb phrase (VP) 'saw Mary with a telescope' (as in John saw Mary *through* the telescope). Each of these interpretations results in a different translation of the sentence. With this kind of structural ambiguity which human translators would find

difficult to disambiguate without the knowledge of the actual event, it is very unlikely that computers can perform the required disambiguation.

The disambiguation of some structural ambiguities, however, can be done by a computer. For instance, the words 'tape measures' can be a noun with the noun 'tape' functions as an adjective (as in "The *tape measures* are all sold out."); these words can also appear as a verb (i.e. to measure) following a noun (as in "The tape measures five inches long."). The occurrences of 'tape measures' in these sentences can be disambiguated after processing other constituents within each of the sentences. Syntactic processing alone is adequate to perform this kind of disambiguation. However, with the word sequence like 'stand by' which can either be interpreted as a phrasal verb (as in "John stood by Mary.") or a verb followed by a preposition (as in "John stood by a lamp-post.") and either structural interpretation can be used in a sentence, using syntactic information alone will not be able to produce the appropriate target language translation[1].

2.2 Different Kinds of MT Systems

The history of MT is dominated by two generations of MT systems. *First generation MT systems* refer generally to the ones which were constructed before 1960s. These systems employed a direct approach to MT which was mainly based on word-to-word and/or phrase-to-phrase translations. Amongst the more famous first generation direct MT systems are the Georgetown system (as described by Kay (1973)) and Systran[2]. As discussed in Section 1.1.2, a simple word-to-word translation cannot resolve the ambiguities arising in MT. A more thorough analysis of source language text is required to produce better translation. As the major problem of the first generation MT was the lack of linguistic information about source text, researchers therefore moved onto finding ways to capture this information. This gave rise to the development of the indirect MT systems which are generally regarded as *second generation MT systems*. This section reviews the characteristics of the different generations of MT systems and explains how these systems attempted to tackle the problem of ambiguity.

2.2.1 Direct MT systems

A direct MT system simply translates source language texts to the corresponding target language texts word-for-word or phrase-to-phrase. Then the resulting target language words are reorganised according to the target language sentence format. In order to improve the output quality, some direct MT systems perform some morphological analysis before the bilingual

[1] Chapter 5 gives a more detailed discussion on this problem.

[2] Though Systran is regarded as a direct MT system, the continuous modifications and developments carried out on Systran made it a less typical example of the direct MT approach (Hutchins & Somers 1992). A brief discussion on the characteristics of Systran is given in Section 2.3.1.

dictionary lookup but they rarely analyse the sentence structure of the source language text. Direct MT systems were developed in the 1950s. In those days, computers were very primitive and had a very long processing time. This explains why direct MT systems are very primitive and do not analyse the linguistics of sentences before performing the translation.

Owing to its primitive nature, the direct MT approach is very straight-forward and easy to implement. It supports the translation of source language sentences which have both matching source-to-target language words and similar structures as the target language sentences. However, as very little, if any, effort has been put into disambiguating source language sentences, this approach does not support the translation of ambiguous sentences. This approach also fails to translate sentences to a language which has very different syntactic structures and / or different use of words / phrases from the source language.

The main problem of the direct MT approach is that it does not analyse the linguistic information nor the meaning of source sentences before performing the translation. Without this information, the resulting MT system cannot resolve the ambiguities that arise in the source sentence and / or during the translation. Thus, this approach fails to translate any seemingly ambiguous sentences (e.g. "John fishes for invitations."). As a result, the first generation MT systems cannot provide a quality translation of the source text.

2.2.2 Indirect MT Systems

Owing to the fact that linguistic information helps an MT system to disambiguate source sentences and to produce better quality target language translation, with the advance of computing technology, MT researchers started to develop methods to capture and process the linguistics of sentences. This was when the era of indirect MT systems started.

There are two kinds of second generation MT systems: transfer-based and interlingual systems. As shown in Figures 2.1 and 2.2, the structures of these systems are fairly similar. The module '*Source Texts Analysis*' aims at capturing the required linguistic information of the source sentences for aiding the translation. The transfer-based approach uses the information obtained from the analysis module directly to lookup the corresponding target language words.

The interlingual approach involves the use of an intermediate language (i.e. an interlingua) for the transfer — with the source language text translated to the interlingua and the interlingua translated to the target language text. The interlingua chosen tends to be an artificial language, e.g. Esperanto[3], which is, both lexically and structurally, more regular and consistent than natural languages and could capture the characteristics of any natural language in a relatively

[3]In addition to using an artificial language like Esperanto as an interlingua, a modified version of an established artificial language could also be used. This was the case with the DLT (i.e. Distributed Language Translation) project, in which a modified and restricted form of Esperanto was used to build the DLT prototype (Hutchins & Somers 1992, Chapter 17).

precise way. The use of an interlingua enables an MT system to perform the translation without looking back at and referring to the original source language texts. After translating the

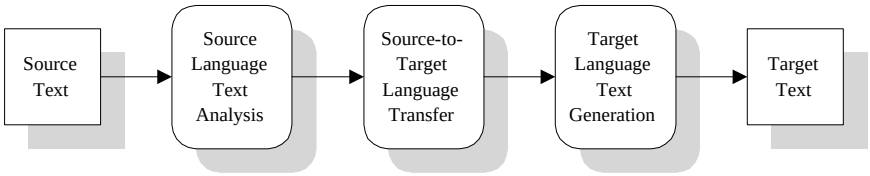


Figure 2.1: Typical building blocks of a transfer-based MT system

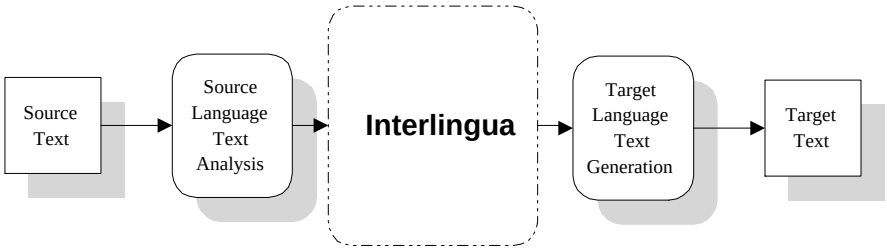


Figure 2.2: Building blocks of an interlingual MT system

source language words to their target language forms, the job of the '*Target Texts Generation*' module is to synthesise the resulting target language words to form the target sentences.

One advantage of the transfer-based approach is that it allows the source language text to be analysed according to what is required for facilitating its translation to a target language. Thus, much less effort, if any, would be wasted in analysing the unnecessary features of the source language sentences. The interlingual approach, however, is more time-consuming as a lot of processing time is consumed in the 'double-transfer'. It also allows a double chance — during both the from and to interlingua translations — for ambiguity to occur. However, if a multilingual MT system is to be built, this approach would reduce the time and effort needed to produce a transfer module for each language pair (as required in the transfer-based approach) in the system (cf. Figure 2.3). As this research is not focussed on multilingual MT, the transfer-based approach appears to be more effective than the interlingual approach. Therefore, the transfer-based model was chosen for this research.

The system structures of both the transfer-based and interlingual approaches allow a systematic analysis and processing of the linguistic information of sentences. However, due to the fact that it is a difficult task to capture the meaning of sentences for aiding the translation (cf. Section 1.1.4), many of the existing indirect MT systems tend to perform syntactic analysis on sentences only. As discussed in Section 2.1, syntactic information alone is insufficient

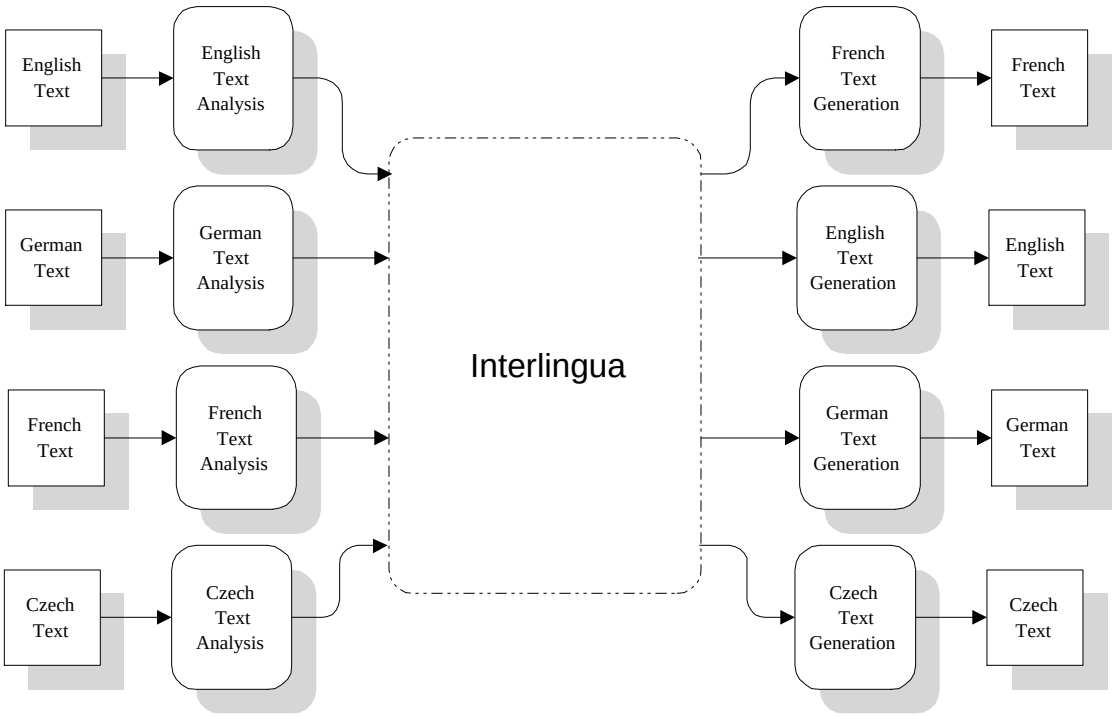


Figure 2.3: Building blocks of a multilingual MT system using the interlingual approach

to resolve some lexical and structural ambiguities. Therefore, the translation power of these systems is fairly limited.

2.3 Practical Use of some MT Systems

During the history of MT, some practical MT systems have been designed and built — the most well-known of these being: Systran, Météo, Eurotra[4] and more recently the ECS (Her, Higinbotham & Pentheroudakis 1994). Of all these MT systems, only Systran and Météo are well-known to have practical use. When comparing the size of these two practical MT systems, Systran is much bigger than Météo. This section reviews and evaluates the characteristics of these systems.

2.3.1 Systran

As mentioned in Section 2.2, Systran is a first generation direct MT system. Systran can handle translation between 23 pairs of languages (some of which are bidirectional) and it is in fact one

[4]Descriptions of Systran, Météo and Eurotra are based on Hutchins & Somers (1992).

of the few successful MT systems which is well-known to be widely used and is still under further expansion after 30 years of its initiation (Hutchins & Somers 1992, Chapter 10).

Unlike most direct approaches to MT, Systran carries out some degrees of analysis of the source texts. Although the design of Systran is based on the direct approach to MT, the stages of translation involved in Systran make it resemble a transfer-based MT system. Within Systran, the analysis, transfer and synthesis tasks are divided into a series of specialised modules. Unlike many practical MT systems, Systran employs an empirical approach and it treats real-life texts as its prospective data (Yang & Gerber 1996). The analysis of source text carried out by Systran does not follow any conventional linguistic model or framework. Instead, it invented and used a linguistic model of its own. To a certain extent, this allows a more flexible way to program and carry out the linguistic analysis; and only the necessary linguistic information is captured and processed. However, without the guidelines provided by and the support of a complete and well-proven linguistic model or formalism, Systran only carries out partial analysis on the source sentences. For technical texts (i.e. the kind of texts that Systran aims at dealing with) which are relatively regular and less ambiguous, such a partial analysis of the source text might be adequate to carry out the necessary disambiguation; however, with texts which are more ambiguous, a more thorough analysis (i.e. an analysis involving more kinds of linguistic information, e.g. syntactic, semantic and even pragmatic information, of the source texts) would be required.

Due to the lack of support from a formal linguistic formalism, Systran was strongly criticised by Hutchins & Somers (1992, page 186):

> "*It still lacks a coherent linguistic theory at its base; many routines are designed empirically for specific problematic constructions associated with particular words in particular languages; there is little generality in lexical and structural transfer; the main burden of the translation process is carried by information contained within the large bilingual dictionaries. As a consequence, methods are inconsistent, coverage and quality are uneven, and modifications of lexical information can often have unexpected consequences.*"

Though Systran does not follow any formal linguistic theory, the fact that Systran has been in well-known practical use since it was first developed 30 years ago shows that the linguistic analysis carried out by Systran apparently seems to be sufficient for translating the kind of real-life texts that it aims at, i.e. technical texts. However, the method to extract linguistic information from the source texts (i.e. one module is used to extract or identify only one kind of linguistic information from the source texts) used by Systran is repetitively redundant and clumsy. In order to finish the analysis, the source text has to go through nine modules (cf. Yang & Gerber 1996). Sometimes a decision made on a lexical item in one module might affect the decision made about another lexical item in an earlier or later module; and often more than one kind of analysis is required in order to analyse a source sentence appropriately. However,

with the use of a coherent linguistic theory, a complete and cohesive analysis on the source texts could be done in one go systematically.

Systran performs translation by the use of several bilingual dictionaries for handling the word-to-word and phrase-to-phrase translation between each language pair. The system construction for each language pair is independent of other language pairs, so each language pair might have a different system architecture. Therefore, even though Systran supports translation between more than two languages, it is not a true multilingual MT system but a collection of several mono-directional MT systems. The independent system construction for each language pair confined the system development to one language pair only, so during the development process the characteristics of other languages could be ignored. As a result, some of the problems caused by multilingual translation (e.g. the difficulty and complexity in defining the right amount of analysis for each source language so as to facilitating the transfer[5] will not occur in Systran, and thus Systran does not need a sophisticated analysis module for processing the source language texts thoroughly during the translation process.

2.3.2 Météo

Météo is another MT system which is well-known to have practical use. Generally regarded as a second generation transfer-based sublanguage MT system, Météo has been used by the Canadian government for translating weather reports from English to French since May 1977 (Hutchins & Somers 1992). Although the size of Météo is very small and it cannot support any translation apart from the weather reports from English to French, it produces translations with an accuracy of more than 90% without any human intervention. In the history of MT so far, not many practical MT systems can support quality output similar to Météo. The success of Météo is mainly due to:

- the limited scope of translation (i.e. weather report only);
- the limited number of language pairs (i.e. from English to French only); and
- the similarity of the source and target languages (i.e. English versus French).

A limited scope of translation helps Météo to avoid some of the problems mentioned in Section 2.1. For instance, as Météo is designed for translating weather reports only, the chance for an idiom to appear in a normal written weather report is so low that it can be *safely* ignored[6]. If an MT system is designed for translating any kind of natural language texts, it will have to cope with many difficult linguistic problems (cf. Section 1.1.3) within the natural language domain. The resulting MT system will be too complicated to build. Though English and French are not from the same language family, throughout the language history of English and French, the

[5] cf. Hutchins & Somers (1992, Section 4.1)

[6] Section 2.5.1 gives a further explanation on how a limited scope of translation can avoid some of the problems in MT.

two languages, to a certain extent, have been influenced by each other (as England and France are geographically and economically very close to each other). It would be less likely for the problem of different realisations of word meanings and different ways of meaning expressions to occur between English and French.

Although Météo is generally regarded as a second generation MT system, in terms of modularity, it does not have a clear separation between the analysis, transfer and generation modules. It gained the reputation of a second generation MT system because its analysis, transfer and generation modules are conceptually independent of each other. As the language pair handled by Météo is English to French only, a highly modular second generation MT architecture would not have a significant beneficial effect on Météo.

Unlike many conventional second generation MT system, Météo transfers the lexical items of a source sentence before carrying out the syntactic analysis on the sentence. This is made possible by the fact that weather reports contain highly regular text with limited use of words, and thus less problems would be caused by lexical ambiguity. The transfer is simply done by looking up each source lexical item from three dictionaries: Idioms dictionary, Place-names dictionary and the main dictionary.

2.3.3 Discussion

The evolution of different generations of MT approaches has proceeded gradually throughout the history of MT. As demonstrated by Systran and Météo, features of different generations of MT approaches blend in with each other. By looking at the domain of real-life practical MT systems, it is difficult to draw a line between the different generations of MT systems. More often, MT system developers employ the approach(es) which is / are most suitable to cater for a particular use of the required MT system. However, there is no doubt that second generation MT approaches provide a more systematic and thorough way to carry out MT. As illustrated earlier, the success (in the sense of real-life application) of Systran and Météo cannot be directly attached to the evolution of the second generation MT. The failure of most of the historical MT systems often were due to the obsession with building a very large and sophisticated MT system which can produce an output translation resembling the one done by human translators[7]. When the domain of translation is large, there will be a much greater chance for linguistic ambiguity to arise in an MT system. The problems arising in MT will be more difficult to solve. As a result, some researchers moved onto another approach to MT: the sublanguage approach, which will be discussed later in this chapter.

The overall approaches of the transfer-based and interlingual techniques are fairly similar — they both have a separate module for source language analysis and target language generation. In fact, the transfer-based approach in certain ways (when the use of the interlingua is ignored)

[7]The failure of Eurotra is a very good example for the problems of a very large and sophisticated multilingual MT system.

can be viewed as a more straight-forward (as the translation is done from source language to target language directly, without the need to use an intermediate language) and simpler (as the MT system does not need to deal with one more language in addition to the source and target languages) version of the interlingual approach. The more specific research on the interlingual approach is to establish, discover and / or refine a good interlingua[8] for mediating the translation. This is regarded as a more tedious work when compared with the transfer-based approach because:

- the choice of interlingua is quite often task- or system-oriented,
- there do not exist many good candidates to be chosen as an interlingua[9].

As a result, the interlingual approach is of no interest to this research study. As identified in Section 1.1.2, there are still a lot of unsolved problems in MT and most of them affect the source-to-target language transfer seriously. It is believed that a possible way to resolve some of these problems is to improve the method used for source-to-target language transfer. This research study therefore focuses on the improvement of the transfer module as defined in the transfer-based approach.

2.4 Methods of Transfer

The aim of transfer in MT is to get the most appropriate translation of a source text into the corresponding target language equivalent. Transfer in MT is not a simple word-to-word lookup in a bilingual dictionary (cf. Section 1.1.2), it involves grouping of some neighbouring words (e.g. when transferring the idiom 'to kick the bucket') and non-neighbouring words (e.g. when transferring the phrasal verb 'hand over' in '*hand* Hong Kong *over*') words and the selection of appropriate source-to-target transfer units. Good transfer is normally done based on the context of a sentence. Appropriate translations are obtained according to the meaning of the words used in the source sentence, and this is what makes transfer more advanced than simple dictionary lookup.

Météo is a successful example of transfer-based systems. According to Hutchins & Somers (1992, Chapter 12), it carries out the required translation in three major steps:

1. Transfer the source English text to the appropriate target equivalent(s)
2. Analyse the syntactic structure of the input English sentences and eliminate the inappropriate transfer units with the help of some semantic features

[8] Since natural languages are highly irregular, they are not suitable to be used as an interlingua. More regular artificial languages are more suitable to be used as interlinguas. Artificial languages can be either created by system developers (so as to cater for more specific needs) or selected from one of the existing artificial languages, e.g. Esperanto.

[9] Even for the use of Esperanto — a well-known artificial language which can be used as an interlingua, there are no good examples of MT systems which were developed successfully with the use of this artificial language.

3. Generate the required target sentences according to the French syntactic and French morphological rules

Météo obtains all the possible source-to-target language translations at the start of the system processing and discards the inappropriate translations as the processing goes along. In general, this transfer method increases the volume of data stored by the system during the system run-time, especially when the source language text is highly ambiguous. However, being a small scale sublanguage system which aims at translating weather reports from English to French only, the chance for Météo to encounter ambiguous text is low. Thus, even though it performs dictionary lookup before analysing the syntax of a sentence, the total volume of data carried in the system during system run-time does not affect its efficiency significantly. Furthermore, during the time when Météo was developed, computers were not as efficient and powerful as nowadays. Having two dictionary lookups (i.e. a monolingual one during parsing and a bilingual one during the transfer) would prolong the processing time even more. This was probably the reason why Météo chose to have only one dictionary lookup at the beginning of the processing.

Météo has a very limited scope of translation. This makes it immune to the problem of handling a large amount of data even though it performs dictionary lookup before analysis. However, nowadays the need for such a small scale MT system is less significant. With a reasonably large scale MT system, the input text is more likely to be more ambiguous. This means that the chance for a lexical item to have more than one translation is higher. Moreover, the chance for a word to combine with other word(s) to form a modified meaning is also higher. As there is a need to keep track of all the possible translations for each word and / or phrase in the source translation, if the source-to-target language transfer is carried out before sentence parsing, a large amount of data will be required to be handled by the system during the whole MT process. As a result, the system processing will be slowed down dramatically.

As the disadvantage of carrying out transfer before source language analysis outweighs its advantage, most transfer systems carry out the transfer after sentence parsing. After sentence parsing, more linguistic information about the sentence has been obtained, which can then be used to identify the appropriate role of each word in the sentence during the target translation selecting process. In addition, since the actual selection process for obtaining appropriate translations is in the middle of the whole MT processing, there is no need to carry the full collection of potential translations at the very beginning of, and throughout, the system processing until the inappropriate ones can be discarded. As a result, the data handled at each point of the system is reduced to minimum. The quality of the transfer in this method relies very much on the information obtained during the parsing. More thorough parsing done on the source sentences leads to an easier identification of appropriate translation units and thus results in a higher output quality. Therefore, if both the syntactic and semantic information of a sentence is captured, the quality of the translation can be improved. However, so far in the history of MT, the systems built had either moderate to extensive amounts of syntactic processing

or simple semantic processing (e.g. the experimenting system built by Wilks (1976)), but none of them combined a thorough syntactic and semantic parsing together in analysing the source language texts. This is probably due to the fact that most systems are domain specific and thus some of the difficult problems of disambiguation can be ignored as they will not have a chance to come up in the confined domain of translation.

As mentioned earlier, a good transfer should be done based on the meaning of the source language text. However, as discussed in Section 1.1.4, to capture the meaning of a word or a sentence is not an easy task. Therefore, some MT researchers began to explore alternative methods to aid the transfer without the need to analyse the meaning of the source language text. Some researchers found out that syntactic information about sentences can help the disambiguation process (cf. Section 3.2). Thus, in order to get round the difficulties in capturing the meaning of sentences, some MT system developers simply introduced some generalised phrase structures (cf. Section 3.2.3) which attempt to represent the different cases in which a word or a phrase can be used. However, these generalised phrase structures are established based on syntactic information only, there is no consideration of the semantic properties or the actual meaning of the word / phrase used in a particular case. Therefore, in many cases, they are inadequate to aid the disambiguation process[10].

Another well-known vital problem in MT is the problem of its speed[11]. More thorough analysis of both syntactic and semantic information of source text often results in longer overall system processing time because it is very difficult, if not impossible, to greatly reduce the time required for the transfer and sentence generation. Even though reducing the amount of analysis done on the source text will decrease the chance to get a high output quality, the resulting system would be more efficient and thus more suitable for catering for the real-life business needs. As average to good MT systems with thorough syntactic analysis can translate technical texts with output quality of 70–90%[12] (which is quite acceptable), and more thorough analysis would probably be able to increase the output quality by at most a few percent, most commercial MT systems tend to give up this 'costly' trade-off.

2.5 Alternative Approaches to Machine Translation

In addition to the classic Machine Translation (MT) approaches discussed above, in recent decades, researchers have discovered new approaches to MT. This section reviewed some con-

[10] cf. Section 3.2 for a more detailed discussion on the inadequacy of using syntactic information alone for the transfer.

[11] Though in the various articles on different MT systems, the processing speed is not discussed, it is a well-known fact that MT systems consume a very long time to carry out the translation and it is nearly impossible to produce a quality translation over some ambiguous texts in a very short processing time. Therefore, the previous work tends to be done on improving the translation quality instead of aiming at achieving both.

[12] With a limited sublanguage system like Météo, the output quality can even be boosted to more than 90%.

temporary approaches to MT.

2.5.1 Sublanguage Approach

Work on NLP often tries to investigate formal and adequate ways to generalise language. However, with the existence of many different kinds of ambiguities in natural languages, it is very difficult to successfully generalise the complete domain of language. A possible way to attempt to generalise language is by restricting the scope of language for processing into a smaller and more domain-specific area, e.g. language used in computing journals, manuals for specific products or weather reports. The use of the sublanguage approach in MT can help in narrowing down the scope of translation so that the size of the resulting MT system will be relatively small and of a manageable size. An example of a well-known sublanguage MT system is Météo which was developed for translating weather reports from English to French only. The sublanguage approach seems a more realistic way to develop MT systems for commercial purposes.

By restricting the usage of a natural language to a particular domain, some of the potentially ambiguous words or phrase structures will no longer be ambiguous. For instance, even though homographs (e.g. '*shower*') exist in weather reports, the problem of ambiguity caused by them will not occur in Météo. Thus, if the word 'shower' appears in a weather report, it would immediately be understood as *a brief fall of rain* (Horby 1984) instead of any other meanings associated with this word (e.g. *a device which produces a spray of water for washing our body* or *a way to wash our body by standing under a spray of water* (Sinclair 1987)). As a result, other possible translations of the word 'shower' can be safely ignored by Météo.

The sublanguage approach makes the development of MT systems relatively easy by ruling out some of the ambiguities in the source texts, and it enables the resulting MT systems to have a better chance to success. However, instead of tackling the problems of ambiguity, this approach simply avoids them. Although the uses of some sublanguages may over-lap to a certain extent (Ananiadou 1990, page 10), it does not mean that when linking all successful sublanguage MT systems together, the resulting MT system will work. A sublanguage MT system often can only produce satisfactory translations when it is used in its domain. For instance, Météo works well in translating weather reports, but it will fail to translate news reports. In order to make the resulting system work, amendment of how to distinguish a word with more than one meaning or use will be needed. Thus, the system will still have to cope with the problems of ambiguity.

Although sublanguage MT cannot resolve ambiguity, the idea of distinguishing the appropriate translation of a word by the use of the domain of the language to which a sentence belongs can facilitate the lexical selection process during the transfer. For instances, as suggested by Her et al. (1994, page 206), the two meanings of the word 'bug' can be distinguished by the f-structure reproduced in Figure 2.4[13]. When the word 'bug' is used in the computing domain,

[13]The Chinese words '毛病' and '蟲子' in Figure 2.4 pronounce 'mao2bing4' and 'chong2zi' respectively in Mandarin Chinese.

$$ec\_bug :: \left[\begin{array}{ll} \text{WORD} & \left\{\begin{array}{l}\left[\begin{array}{ll}\text{FORM} & \text{‘毛病’}\\ \text{TECH} & \text{COMPUTER}\end{array}\right]\\ \left[\begin{array}{ll}\text{FORM} & \text{‘蟲子’}\\ \text{TECH} & \text{GENERAL}\end{array}\right]\end{array}\right\} \\ \backslash\text{STD-T-TRAN} & \end{array}\right]$$

Figure 2.4: A dictionary entry for transferring ‘bug’ suggested by Her *et al.* (1994)

it refers to *a defect of a computer system*; whereas in general, it means *an insect*. From the example given by Her et al. (1994), it is unclear about how the domain (e.g. technical or general) of a sentence is categorised during system run-time. This categorisation might be done by human pre-editing. Despite this doubt, it is in fact possible to use domain tagging to aid the source-to-target language transfer without the need of a human-aided pre-editing process if the tagging is used as part of the semantic information of a homograph. For instance, consider the following sentence:

(1) I have a bug in my program.

The words ‘bug’ and ‘program’ are two separate homographs in American English. If each of the translations of ‘bug’ is tagged with ‘TECH: GENERAL’ and ‘TECH: COMPUTER’ as suggested by Her et al. (1994) and those for ‘program’ are tagged similarly (i.e. with ‘TECH: GENERAL’ refers to the booklet which contains information about a play or a concert; and ‘TECH: COMPUTER’ refers to the set of instructions in a computer system), when these taggings are matched with each other during the transfer, the two realisations of the sentence (1), i.e.:

1. I have a bug$_{[insect]}$ in my program$_{[booklet]}$.
2. I have a bug$_{[defect]}$ in my program$_{[set\ of\ instructions]}$.

can be obtained successfully. This method is similar to the use of semantic primitives in Wilks’s Preference Semantics (PS) system (as described by Whitelock & Kilby (1995)). Like the semantic primitives in Wilks’s PS system, it might not be too difficult to define the appropriate taggings if the dictionary is relatively small. However, if the dictionary is of a reasonable size for practical MT purposes, it would be more difficult to define the appropriate taggings without having to introduce and to maintain a large number of different possibly unrelated taggings. Note that this method has another limitation. With homographs whose target translations have very subtle differences, this method might not work. For instance, it would be difficult to differentiate the translation of the verb ‘break’ in “John broke a vase.” from that in “John broke a tree.” to Chinese (i.e. ‘打碎’ versus ‘打斷’).

2.5.2 Statistics-based Approach

The use of statistical data for MT has been suggested since the age of first generation MT. However, in the history of MT, this approach has not been employed in many practical MT systems. This is perhaps mainly due to the low translation quality produced by this approach.

As reported by Somers (1990), in the pure statistics-based approach to MT, NO linguistic knowledge of the source and the target languages is required to perform the translation. The translation is based on the statistical data on which source language word unit is translated to which target language word(s) and how often this translation occurs. This statistical data is obtained from an analysis of a vast amount of bilingual texts. Different probabilities are extracted from the bilingual texts automatically by a computer, i.e. the probability of a source sentence to occur in the texts, the probabilities of a source word to be translated as one, two, three, etc. target words, the translation probabilities for each word in each language and the probabilities of the position of each source word in a sentence which is not in the same position of the target word in the target sentence (Arnold, Balkan, Humphreys, Meijer & Sadler 1994). These probabilities are vital to the translation process as they are the sole information for calculating how a source sentence should be translated to the target language form.

Without going through any further analysis on a source sentence, a statistics-based MT system performs the translation directly based on the statistical information. For instance, if the analysis shows that *'blowing snow'* in English is always translated as *pouderie* in French, whenever the MT system comes across with the words 'blowing snow' in English, without analysing the underlying linguistic functions of these two words, it will translate them to *pouderie*. If there are more than one target language equivalents for a source language word, the frequency of each translation will be used for calculating the probability of the use of each translation[14]. While translating an ambiguous word, the probabilities for the current and neighbouring words in a sentence will be combined and used for resolving the ambiguity (cf. Arnold et al. 1994, pages 201–204).

Unlike the traditional rule-based approach to MT[15], statistics-based approach allows an MT system to find the best match available through the maximisation of the calculated probabilities. As the translation probabilities are obtained through the analysis of a vast real-life bilingual corpus, the statistical data obtained should contain translation information which covers a good variety of different cases for each use of word or phrase. Thus, the resulting MT system would have a good chance to find a match or even the best match. Even if an exact match of

[14]For example, as described by Hutchins & Somers (1992), the statistical data obtained from the corpus used by the statistics-based MT built by IBM show that '*the*' in English has 0.610 chance to be translated as '*le*' in French and 0.178 chance to be translated as '*la*'.

[15]The translation process in a rule-based MT system often relies on matching the set of pre-defined hand-crafted rules (often encoded with some linguistic information) in the MT system. If no matching rule is available at any instance, the MT system often will fail and cease to produce a translation for the source sentence which is currently under processing.

a translation is not listed in the bilingual corpus, the MT system can still use the translation probabilities to approximate a possible translation.

Provided that a good corpus of bilingual texts is available, the statistics-based approach offers a fast and less costly approach to MT as both the extraction of statistical data and the actual MT process are done electronically. While acquisition from experts in linguistics can be a difficult and costly task, this approach completely avoid this problem as it does not require any knowledge of linguistics. However, the translation performance of this approach is rather poor. According to the work carried out by a team at IBM, about 48% of the translations produced were either the same as or preserved the meaning of the official translations (Somers 1990); out of 100 short test sentences, only 39% of the translations are correct (Arnold et al. 1994). If this approach is used in real-life MT tasks, a lot of post-editing on the resulting translations will be required which makes this approach very costly.

In some cases, even shallow linguistic analysis (e.g. for obtaining syntactic structure of sentences) can achieve similar result, e.g. the knowledge of the syntactic category of the homograph '*yank*' within a sentence is sufficient to determine its translation appropriately. However, if there is a need to calculate the combined probability of the neighbouring words, the translation process for this kind of relatively simple translation becomes unnecessarily more complicated (in terms of the amount of mathematical calculation involved).

Further work carried out by the IBM team showed that the introduction of simple linguistic information to this approach improved the results to 60%. This seems to suggest that, while statistics-based approach to MT is plausible and has its potential to produce an efficient MT system, some level of linguistic information is still required in order to produce a high quality translation. Therefore, as pointed out by Arnold et al. (1994), one possible extension to the research on this approach is to investigate the effective use of both statistical data and linguistic analysis to MT.

2.5.3 Example-based Approach

The example-based approach is another contemporary approach to MT which relies on the use of a bilingual corpus. As suggested by its name, this approach collects examples of translation pairs from the bilingual corpus and translates the source text by using the best match from these examples together with a word-for-word translation. The neighbouring words in the source text are often used to aid the determination of the best match. For instance, according to an example given by Arnold et al. (1994, page 200), the different translations to English for the Japanese word '*sochira*' are 'this', if the source string involves the Japanese word '*desu*' (meaning '*be*') or '*miru*' (meaning '*see*'), and 'you', if the source string involves the Japanese word '*okuru*' (meaning '*send*'). To translate an input string like '*sochira ni tsutaeru*', the translation 'you' would be chosen for the Japanese word 'sochira' as the meaning of the word 'tsutaeru' (meaning '*convey*') is closest to 'okuru' (meaning '*send*'). This approach resembles how a human

translator performs translation, i.e. searches for the best translations to fit the word or phrase concerned from all the available translations for this word or phrase.

A pure example-based approach does not require the aid of any linguistic information. The translation is purely done by locating the best matching example and sometimes calculating how close the located example matches with the source language string (i.e. words, phrases or sentences, depending on how the system is developed). As the translation carried out by the example-based approach is based on string matching, it does not support a translation by approximation like the statistics-based approach. If a suitable match is not available from the examples listed in the system, this approach would fail to produce a translation, unless a default translation is pre-defined in the system. Another possible problem, as suggested by Arnold et al. (1994), is that there might be several examples, each of which matches part of the source language string, or where the examples match the source string do not cover the entire string; in this case, as all the matching examples would need to be taken under consideration when calculating the best match, the resulting calculation involved would be very complicated.

While the linguistic-based approach generalises linguistic information in terms of grammar rules, to a certain extent, the example-based approach can be viewed as a more detailed way to specify real-life linguistic information about sentences. In terms of selecting the best translation for an individual word / phrase unit within the source language texts, this approach is more realistic and flexible as real-life data, instead of some pre-defined grammar rules produced as a result of the research in theoretical linguistics, are used to generate the translation. However, in order to produce a complete target language translation for a source language sentence, a considerable amount of complicated calculation is involved, especially when a complete match is not available. In terms of generating a target language sentence, the use of simple grammar rules seems to be a more straight-forward approach to accomplish this task.

2.6 Conclusion

This chapter very briefly introduced the history of Machine Translation (MT). It also reviewed and discussed some of the computational and linguistic issues of existing MT systems. A lot of problems exist in the domain of MT which makes good quality MT a very difficult task. It is believed that any attempts at alleviating some of these problems would benefit the field. So far, we have reviewed a number of problems in this domain. However, this study does not intend to find a solution to all of these problems. Rather, by identifying and studying the various existing problems in MT, a clearer idea on what needs to be done in order to improve the quality of MT is obtained. As it was discussed in Chapter 1, most of the problems in MT are caused by ambiguous sentences and versatile uses of words / phrases. Section 2.4 of this chapter illustrated that pure syntactic analysis alone cannot provide sufficient information to perform source-to-target language transfer effectively. Though the second generation MT systems em-

ploy a modular approach to tackle MT, many of these systems mainly focus on extracting and using syntactic information about source language sentences to perform translation as it is relatively easy to program an MT system to perform syntactic analysis on sentences compared with any higher level of linguistic analysis. This relatively low level of linguistic analysis is insufficient to produce good quality translations for these MT systems.

Somers (1990) pointed out that the problems of second generation MT are mainly due to the fact that the developers did not put in enough effort to find out the actual problems of MT with different language pairs. After studying some of the existing problems in MT, it is observed that the lack of an adequate level of linguistic analysis on the source language texts makes the resulting MT systems incapable of resolving some of the ambiguities arising in the source texts. A systematic introduction of a higher level of linguistic analysis is believed to alleviate this problem.

This chapter also gave a brief discussion on several contemporary alternative approaches to MT. The sublanguage approach allows the development of effective practical MT systems with high-quality translation output for translating texts in a very small and confined domain. The statistics-based and example-based approaches take advantage of an effective use of a vast collection of bilingual corpus for developing efficient MT systems which require very little involvement of human effort. Each of these alternative approaches to MT has its own limitation(s). For instance, the sublanguage approach cannot handle general-purpose MT. With the use of the statistics-based approach, in order to produce good translation, a certain level of human editing on the final output is required. The main problem of these approaches is that they did not have a good algorithm to solve the problems of MT adequately. The sublanguage approach does not attempt to solve the linguistic problems in MT, but simply ignores them. Therefore, if it is to be applied on a wider scope of translation, the readability and correctness of the output translation cannot be guaranteed. Though the statistics-based approach attempts to solve the problem of ambiguity and the problem of slow processing speed by an effective use of statistic data. Without the support of linguistic information, the translation quality supported by this approach is not good enough for practical purposes. Though, it is observed that the introduction of linguistic information to the statistics-based approach would produce more fruitful results for this approach. This research study is not intended to find a good combination of linguistics and statistics to aid MT.

The alternative MT approaches reviewed in this chapter cannot provide an adequate solution to the difficult problems in MT. While the problems of MT remain unsolved, the success of these alternative approaches to MT is limited. If a less domain-specific MT system is required, these approaches will either lead to a complete failure or produce a relatively low quality output. Therefore, there is still the need to identify and to find ways to solve the problems in MT. One seemingly more promising solution to the problem of ambiguity is to analyse the source language texts systematically according to the guidelines provided by a formal computational linguistic formalism. With the knowledge of the linguistic information carried in the source

language texts, it is believed that the problem of ambiguity would be alleviated. This study is aiming at exploring a new way to introduce a higher level of linguistic information to aid MT tasks.

Chapter 3

Lexical-Functional Grammar (LFG)

Non-computational grammars of natural languages are descriptive in nature. The grammatical information within these grammars is often not defined in a manner which is suitable for computation. In addition, there are often no obvious links between different kinds / levels of linguistic information. Thus these grammars are not readily applicable to Natural Language Processing (NLP). A better-structured grammar which is suitable for computational representation is needed to linguistically describe natural languages. This can be achieved by symbolising the linguistic information of each word within the natural language domain and defining relationships between these symbols. There are many linguistic formalisms which are computationally viable and can be used to form the linguistic backbone of research and development of NLP systems. LFG is a good example of these formalisms in terms of its suitability for aiding NLP.

Though linguistic formalisms attempt to represent the characteristics of natural languages, they are different from abstract interlinguas. Linguistic formalisms tend to analyse and represent natural languages in terms of the linguistic roles of the words or phrases within a sentence. The representation often is done by tagging each component of a sentence with one or more corresponding role / feature markers within the framework of the sentence defined by the particular linguistic formalism. When transferring sentences from one language to another, the original source language words would still form a base to select the corresponding target language words. This differs from the interlingua method. As suggested by Hutchins & Somers (1992), an interlingua is an intermediate 'meaning' representation and this representation:

> "*includes all information necessary for the generation of the target text without 'looking back' to the original text. The representation is thus a projection from the source text and at the same time acts as the basis for the generation of the target text; it is an abstract representation of the target text as well as a representation of the source text.*"
>
> [Hutchins & Somers (1992, page 73)]

When compared with an abstract interlingua, a linguistic formalism like LFG aims at representing the concrete linguistic functions and behaviours of words within sentences instead of

being an abstract representation of sentences.

LFG was developed with the goal of serving as the grammatical basis of a computationally precise and psychologically realistic model of human language (Sells 1985*a*). LFG is a descriptive, model-based approach to grammatical analysis. This makes LFG suitable for acting as the linguistic backbone of NLP systems and LFG has been used extensively in the area of NLP (especially in MT) for more than ten years. Many results prove that LFG is quite adequate and easy to use in serving the needs of different kinds of NLP systems. LFG also provides a systematic way to analyse the different levels of linguistic information encoded in sentences. As research and development on the LFG formalism are not confined to the English language, but are conducted based on the observation of many different languages from different families, LFG has proven to be suitable for applying in different languages, e.g. Kudo & Nomura (1986), Kaplan, Netter, Wedekind & Zaenen (1989), Nyberg & Mitamura (1992) and Her et al. (1994). Owing to these characteristics, LFG is chosen as the main linguistic formalism for this research.

This chapter gives a brief introduction to some aspects of the LFG formalism. It also discusses some of the previous applications of the LFG formalism in Machine Translation (MT).

3.1 The LFG Formalism

Lexical-Functional Grammar was developed by Kaplan and Bresnan in the late 1970s. An early version of the formal principles of this grammar was first extensively described in 1982 (Bresnan 1982*b*). Since then, a considerable amount of continuous research on improving and extending this formalism has been conducted which makes LFG (when compared with other formalisms) fairly mature and stable for describing the various characteristics of natural languages.

Unlike the mainstream of generative grammar, LFG advocated the view that syntax is not only expressible in terms of phrase structure trees (Sells 1985*a*). In fact, Kaplan (1989) pointed out that a phrase structure tree alone is insufficient to represent the syntactic information of a sentence because the more abstract relations between grammatical functions and features cannot be conveniently expressed in a phrase structure tree. Thus, the model of syntax proposed in LFG is not purely tree-based. Instead, LFG uses two different structures to represent different aspects of syntax. The external structure of a sentence is represented in the form of a tree structure named constituent structure (c-structure); whereas the internal structure is represented in the form of an acyclic graph structure named functional structure (f-structure). Kaplan (1989) suggested that LFG is a linguistic formalism which attempts to deal with different levels of linguistic representations, i.e. syntactic, semantic and pragmatic, in a coherent manner. In addition to representing syntactic information by means of c- and f-structures, LFG uses two other structures, i.e. the predicate argument structure[1] (a-structure) and the f-structure look-alike se-

[1] cf. Chapter 4

mantic structure (s-structure) (Halvorsen & Kaplan 1988), to express the thematic and semantic information of a sentence. The different structures in the LFG formalism are related to each other hierarchically by means of structural correspondences.

3.1.1 Constituent Structure (c-structure)

The c-structure in LFG represents the external structure of a sentence in the form of a phrase structure tree. It shows the syntactic categories and the linear order of constituents. It also shows the hierarchical grouping of words in a sentence, i.e. how each phrase within the sentence is formed by the combination of words in the sentence and how these phrases combine to form the sentence itself. The hierarchical grouping of words in a sentence is governed by phrase structure rules which are in the form of the context-free grammar rules. For instance, consider the sentence:

(2) John played Mary a tune on the violin.

the set of phrase structure rules that describes the structure of this sentence is:

(3)
$$
\begin{array}{lcl}
S & \longrightarrow & NP \quad VP \\
NP & \longrightarrow & N \\
VP & \longrightarrow & V \quad NP \quad NP \quad PP \\
NP & \longrightarrow & Det \quad N \\
PP & \longrightarrow & P \quad NP
\end{array}
$$

where ‘S’ stands for ‘Sentence’, ‘NP’ stands for ‘Noun Phrase’, ‘VP’ stands for ‘Verb Phrase’, ‘N’ stands for ‘Noun’, ‘V’ stands for ‘Verb’, ‘PP’ stands for ‘Prepositional Phrase’, ‘Det’ stands for ‘Determiner’ and ‘P’ stands for ‘Preposition’. The c-structure of the sentence in (2) can be obtained by applying the phrase structure rules in (3) as is shown in Figure 3.1.

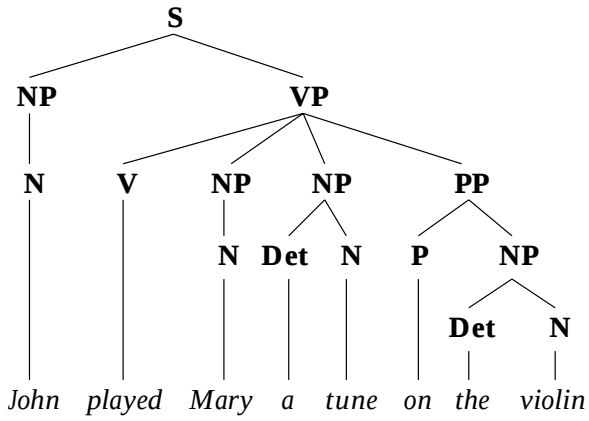


Figure 3.1: C-structure for the sentence “*John played Mary a tune on the violin.*”

C-structure displays information about the part-of-speech of each constituent in a sentence and the syntactic structure of the sentence. This information is useful for simple lexical disambiguation. For instance, the English word '*minute*' has different meanings when it is used in different part-of-speech. When this homograph is used as a *noun*, it is used as a unit for measuring time (e.g. "One *minute* has sixty seconds."). When it is used as a *verb*, it means "*to make a written record of what are said or decided during a meeting*" (Sinclair 1987) (e.g. "I *minute* a meeting."). As an adjective, the word '*minute*' means '*tiny*' (e.g. "I can see only *minute* difference between these articles."). Knowing the part-of-speech of the word 'minute' will be sufficient to determine its meaning and its translation appropriately. However, this method only works for a small number of cases: the disambiguation of the majority of homographs cannot be handled by this method.

As c-structure encodes surface syntactic information like word order and phrasal structure, it is language dependent. Although c-structure displays how each constituent is grouped to form a sentence which can aid analysing source language sentences or generating a target language sentences in an MT system, it is language dependent and it only captures the shallow syntactic information of sentences which makes it insufficient for performing the transfer of sentences from one language to another.

3.1.2 Functional Structure (f-structure)

While c-structure captures the external structure of a sentence, f-structure represents the internal structure of a sentence. This includes the representation of the higher syntactic and functional information of a sentence. The higher syntactic information of a sentence refers to the grammatical information of a lexical item such as: the word 'cats' is in *plural* form and the word 'ate' is expressed in *past* tense. The functional information of a sentence includes the information about functional relations between parts of sentences and how each part of the sentence affects each other (Sells 1985*a*, LFG research group in CSLI 1995). The relationship between some elements of a sentence is shown in an f-structure by means of the links drawn between them. F-structure also expresses the information about the kind(s) of syntactic functions that each predicator (e.g. verb or preposition) governs.

The higher syntactic and functional information of a sentence is represented in f-structure as a set of attribute-value pairs. These pairs form the nodes of the acyclic graph structure. In an attribute-value pair of an f-structure, the attribute corresponds to the name of a grammatical symbol (e.g. NUMB, TENSE) or a syntactic function (e.g. SUBJ, OBJ) and the value is the corresponding feature possessed by the concerning constituent. The value for each attribute can be an atomic symbol, a semantic form or a subsidiary f-structure (Kaplan & Bresnan 1982, Sells 1985*a*, Kaplan 1989). An atomic value is used to describe a grammatical feature of a constituent, e.g. the tense of a verb, whether a noun is of a singular or plural form, etc. (4) shows an example of an attribute-value pair with an atomic value showing the tense of the verb 'played':

(4) $\begin{bmatrix}\text{TENSE} & \text{PAST}\end{bmatrix}$

In LFG terms, a semantic form expresses the semantic interpretation of a predicate. This semantic interpretation is represented in terms of the syntactic functions a predicator governs. For instance, the attribute-value pair which encodes the semantic form of the verb 'played', as in (2), is:

(5) $\begin{bmatrix}\text{PRED} & \text{'PLAY}\langle(\uparrow \text{SUBJ})\ (\uparrow \text{OBJ})\ (\uparrow \text{OBJ2})\rangle\text{'}\end{bmatrix}$

The functional structure of a syntactic function is encoded as a subsidiary f-structure in an attribute-value pair. For instance, the f-structure representation of the NP 'John' which functions as the subject in a sentence (e.g. in (2)) is:

(6) $\begin{bmatrix}\text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{'JOHN'}\\ \text{SPEC} & -\\ \text{NUMB} & \text{SG}\\ \text{PERSON} & 3\text{RD}\end{bmatrix}\end{bmatrix}$

As an f-structure may contain subsidiary f-structure(s), the f-structure is a multi-levelled tree-like structure. Nevertheless, an f-structure is not a tree because some of the attributes that appear in different places within it can sometimes be linked with each other. For instance, in the sentence:

(7) John tried to play the guitar.

The subject of the sentence 'John' is also the subject of the complement clause "*to play the guitar*". Within the f-structure for (7), the value of the attribute 'SUBJ' will be *linked* to the value of the same attribute in the f-structure of the complement (cf. Figure 3.2). Within the same level of an

$$\begin{bmatrix}
\text{PRED} & \text{'TRY}\langle(\uparrow \text{SUBJ})\ \ (\uparrow \text{XCOMP})\rangle\text{'}\\
\text{TENSE} & \text{PAST}\\
\text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{'JOHN'}\\ \text{SPEC} & -\\ \text{NUMB} & \text{SG}\\ \text{PERSON} & 3\text{RD}\end{bmatrix}\\
\text{XCOMP} & \begin{bmatrix}\text{PRED} & \text{'PLAY}\langle(\uparrow \text{SUBJ})\ \ (\uparrow \text{OBJ})\rangle\text{'}\\ \text{SUBJ} & \\ \text{OBJ} & \begin{bmatrix}\text{PRED} & \text{'GUITAR'}\\ \text{SPEC} & \text{'THE'}\\ \text{NUMB} & \text{SG}\\ \text{PERSON} & 3\text{RD}\end{bmatrix}\end{bmatrix}
\end{bmatrix}$$

Figure 3.2: F-structure for the sentence "*John tried to play the guitar.*"

f-structure, the attribute-value pairs can appear in any order.

As mentioned in the previous section, the c-structure of a sentence is assigned by phrase structure rules. The phrase structure rules in (3) did not carry any functional information about the constituents within a sentence. Thus, they are insufficient for assigning f-structures. In order to enrich the syntactic information carried by the phrase structure rules, they are equipped with functional annotations. For instance, the functional annotations for the first rule in (3) are:

(8) S $\longrightarrow$ NP VP
$(\uparrow \text{SUBJ}) = \downarrow \quad \uparrow = \downarrow$

where the functional annotation for the NP node expresses the grammatical relation "*the f-structure which fills the value of the attribute 'subject' (*SUBJ*) of the mother of this NP node's is the f-structure of this NP node*[2]"; and the functional annotation '$\uparrow = \downarrow$' for the VP node indicates that the functional information encoded in this VP node is passed to the f-structure of its mother node. In addition to appearing in the form of the functional annotations in (8), most of the functional information appears in the lexical items[3], e.g.:

(9) John N $(\uparrow \text{PRED}) = \text{'John'}$
$(\uparrow \text{NUMB}) = \text{sg}$
$(\uparrow \text{PERSON}) = \text{3rd}$

(10) played V $(\uparrow \text{PRED}) = \text{'play}\langle(\uparrow \text{SUBJ}) \ (\uparrow \text{OBJ}) \ (\uparrow \text{OBJ2})\rangle\text{'}$
$(\uparrow \text{TENSE}) = \text{past}$

The lexical items form the terminals of the grammar rules, e.g. the lexical items (9) and (10) appear in the following grammar rules:

(11) N $\longrightarrow$ John
$(\uparrow \text{PRED}) = \text{'John'}$
$(\uparrow \text{NUMB}) = \text{sg}$
$(\uparrow \text{PERSON}) = \text{3rd}$

[2] According to Kaplan (1989, page 18), the functional annotation '$(\uparrow \text{SUBJ}) = \downarrow$' can be read as "*the matching NP node's mother's f-structure's subject is the matching node's f-structure*".

[3] In LFG, the relationship between a word and its corresponding lexical entry is 1:1. This means that a lexical item having different morphological forms would have different lexical entries for each morphological form. For instance, some of the lexical entries for the different morphological forms of the verb 'play' are:

plays V $(\uparrow \text{PRED}) = \text{'play}\langle(\uparrow \text{SUBJ}) \ (\uparrow \text{OBJ}) \ (\uparrow \text{OBJ2})\rangle\text{'}$
$(\uparrow \text{TENSE}) = \text{present}$
$(\uparrow \text{NUMB}) = \text{sg}$
$(\uparrow \text{PERSON}) = \text{3rd}$

played V $(\uparrow \text{PRED}) = \text{'play}\langle(\uparrow \text{SUBJ}) \ (\uparrow \text{OBJ}) \ (\uparrow \text{OBJ2})\rangle\text{'}$
$(\uparrow \text{TENSE}) = \text{past}$

playing V $(\uparrow \text{PRED}) = \text{'play}\langle(\uparrow \text{SUBJ}) \ (\uparrow \text{OBJ}) \ (\uparrow \text{OBJ2})\rangle\text{'}$
$(\uparrow \text{PARTICIPLE}) = \text{present}$

(12) V $\longrightarrow$ played
(↑ PRED) = 'play<(↑ SUBJ) (↑ OBJ) (↑ OBJ2)>'
(↑ TENSE) = past

The appropriate f-structure for a sentence is obtained by instantiating the functional annotations embedded in the grammar rules. For instance, the f-structure corresponding to the sentence (2) is shown in Figure 3.3.

$$\begin{bmatrix}
\text{PRED} & \text{'PLAY}\langle(\uparrow \text{SUBJ}) \ (\uparrow \text{OBJ}) \ (\uparrow \text{OBJ2})\rangle\text{'} \\
\text{TENSE} & \text{PAST} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'JOHN'} \\ \text{SPEC} & - \\ \text{NUMB} & \text{SG} \\ \text{PERSON} & 3\text{RD} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'MARY'} \\ \text{SPEC} & - \\ \text{NUMB} & \text{SG} \\ \text{PERSON} & 3\text{RD} \end{bmatrix} \\
\text{OBJ2} & \begin{bmatrix} \text{PRED} & \text{'TUNE'} \\ \text{SPEC} & \text{'A'} \\ \text{NUMB} & \text{SG} \\ \text{PERSON} & 3\text{RD} \end{bmatrix} \\
\text{ADJUNCT} & \begin{bmatrix} \text{PRED} & \text{'ON}\langle(\uparrow \text{OBJ})\rangle\text{'} & \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'VIOLIN'} \\ \text{SPEC} & \text{'THE'} \\ \text{NUMB} & \text{SG} \\ \text{PERSON} & 3\text{RD} \\ \text{PCASE} & \text{'ON'} \end{bmatrix} \end{bmatrix}
\end{bmatrix}$$

Figure 3.3: F-structure for the sentence "*John played Mary a tune on the violin.*"

With a grammatical sentence, a parser with the annotated grammar rules will generate a well-formed f-structure. However, these grammar rules can also derive an f-structure corresponding to an ungrammatical sentence. In addition to the annotated grammar rules, LFG provides three conditions to check if the derived f-structure is well-formed, thus ensuring that the corresponding sentence is grammatical. These conditions are: completeness, coherence and functional uniqueness (or consistency) (Kaplan & Bresnan 1982, Sells 1985*a*).

Completeness ensures that all the grammatical functions subcategorised by a predicator (as shown in its semantic form) in an f-structure must have a corresponding value in the same level

of the f-structure. For instance, the semantic form for the verb 'like' (as in "*John likes Mary.*") is:

(13) 'like<(↑ subj) (↑ obj)>'

To make the corresponding f-structure complete, all the syntactic functions appearing in this semantic form (i.e. a subject and an object) must be present in the f-structure. Therefore, the sentence "*John likes.*", which only possesses a subject but no object, would produce an *incomplete* f-structure.

Coherency ensures that only the grammatical functions which are governed by the predicator in the same level of an f-structure can be present, but nothing more. Under this definition, the sentence "*John likes Mary the garden.*" would produce an f-structure which is not coherent because it contains a subject (i.e. '*John*') and two objects (i.e. '*Mary*' and '*the garden*'), while the verb 'likes' only governs a subject and an object (cf. (13)).

The uniqueness condition ensures that the value of each attribute within the same level of an f-structure is unique. This means that within the same level of an f-structure, only one value is assigned to each attribute-value pair in the f-structure. For example, if an f-structure contains the attribute-value pairs (14), the uniqueness condition will not hold in this f-structure.

(14)
$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'JOHN'} \\ \text{SPEC} & - \\ \text{NUMB} & \text{SG} \\ \text{PERSON} & \text{3RD} \end{bmatrix} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'GARDEN'} \\ \text{SPEC} & \text{'THE'} \\ \text{NUMB} & \text{SG} \\ \text{PERSON} & \text{3RD} \end{bmatrix} \end{bmatrix}$$

The above well-formedness conditions for f-structures play an important role in ensuring that an f-structure is well-formed and that all necessary functional information in a sentence is captured adequately and appropriately. In addition, these conditions help to detect whether a sentence is syntactically correct — if the f-structure of a sentence is well-formed, the sentence is grammatically correct.

While the linguistic information displayed in a c-structure to a certain extent aids lexical disambiguation, the linguistic information encoded in an f-structure also helps an MT system to select the appropriate translation for some source language words, especially those which are case sensitive. For instance, unlike English nouns, nouns in the Czech language are case sensitive. There are seven different cases for each noun (either singular or plural): nominative, genitive, dative, etc. A noun in different cases might appear in different spelling. For instance, the proper noun '*Jan*' in Czech has five different spellings[4] (cf. Table 3.1) depending on which

[4]The data on the Czech language shown in this thesis are supplied by a native Czech speaker.

case it is in; whereas its equivalent in English, i.e. 'John', apart from its genitive case which has a different spelling (i.e. "John*'s*"), always has the same spelling in all other cases. When

	Case Name	**Variations**	**Example**	**English translation**
1.	nominative	Jan	*Jan* to řekl.	*John* said it.
2.	genitive	Jana	řeč našeho *Jana*	Our *John's* speech
3.	dative	Janovi	Řekl to *Janovi.*	He said it to *John.*
4.	accusative	Jana	Řekl mi *Jana* 3:16.	He told me *John* 3:16.
5.	vocative	Jane	Řekl mi: „*Jane*".	He called me: "*John*".
6.	locative	Janovi	Řekl mi o *Janovi.*	He told me about *John.*
7.	instrumental	Janem	řeceno *Janem*	said by *John*

Table 3.1: Different cases for the Czech proper noun 'Jan'

translating English sentences like "Mary killed *John*." and "*John* died." to Czech, the proper noun 'John' becomes 'Jana' and 'Jan' respectively. Using c-structure alone is insufficient to perform this kind of translation as the c-structure does not display this kind of case information. This information, however, is implicitly encoded in the f-structures in Figures 3.2 & 3.3 (i.e. being the subject of both sentences, the proper noun 'John' in these f-structures is in the nominative case). This information sometimes is even explicitly displayed in f-structures by using the attribute-value pair:

$$\begin{bmatrix}\text{CASE} & \text{NOMINATIVE}\end{bmatrix}$$

Bearing this case information, f-structure is capable of aiding the lexical selection process.

In addition to aiding the lexical selection process, the information about case encoded in f-structure is also useful in aiding the lexical disambiguation process. Consider the following Czech sentences:

(15) Jde na *trávu*.

(16) Jde na *trávě.*

Both '*trávu*' and '*trávě*' are two variants (i.e. in different cases) of the Czech noun '*tráva*' (Meaning: '*grass*'). Although these sentences are so similar (i.e. both of them start with the words "*Jde na*" and end with the same root form), they have different meanings:

(17) **Czech:** Jde na *trávu*. $\Longrightarrow$ **English:** (He) *goes onto* grass.

(18) **Czech:** Jde na *trávě.* $\Longrightarrow$ **English:** (He) *walks on* grass.

The different translations depend on the case that the Czech noun '*tráva*' (Meaning: '*grass*') has in a sentence. As f-structure contains the information about cases, it can be used to disambiguate the meaning of the sentences (15) and (16) and thus produce an appropriate translation

for each case. In addition to the information about cases, other linguistic information captured in an f-structure (e.g. whether a noun is in its plural or singular form, tenses) are also useful for selecting the appropriate target language translation.

The linguistic information in both c-structure and f-structure aid different kinds of lexical disambiguation. During the transfer, if this information can be used in conjunction with each other, the lexical disambiguation process will become more consistent and more effective. In the LFG formalism, the correspondence from c-structure to f-structure is denoted by the symbol ϕ. The details of this correspondence are defined by *co-description* in the form of simple equations. Consider the simple English sentence "*John died.*" and its corresponding c-structure and f-structure shown in Figure 3.4. The NP 'John' in the c-structure in Figure 3.4 corresponds

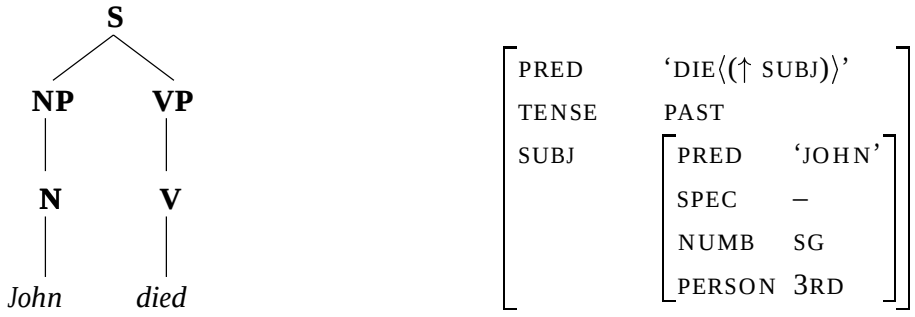


Figure 3.4: C-structure and F-structure for the sentence "*John died.*"

to the subject f-structure, SUBJ_{fs}, in the same figure. In LFG terms, this correspondence can simply be represented by the following equation:

(19) $\text{SUBJ}_{fs} = \phi(\text{NP})$

By defining which c-structure components correspond to f-structure components, a c-structure is properly linked with its corresponding f-structure. The correspondence between the c-structure and the f-structure in Figure 3.4 is shown in Figure 3.5.

3.1.3 Semantic Structure (s-structure)

While the c-structure and the f-structure capture different kinds of syntactic information of sentences, the semantic structure (s-structure) in LFG is responsible for representing the semantics of a sentence. Unlike Wilks's preference semantic system as described by Whitelock & Kilby (1995), the representation of semantic information in an s-structure does not involve the use of semantic markers for characterising the semantic properties of each lexical item. This is because the kind of semantic information captured in an s-structure does not correspond to the combination of the detailed semantic properties of each word. Rather, the s-structure shows the semantic interpretation of the sentence (Halvorsen & Kaplan 1988). For instance, the information like the number of arguments that a predicator takes and the fact that an active and

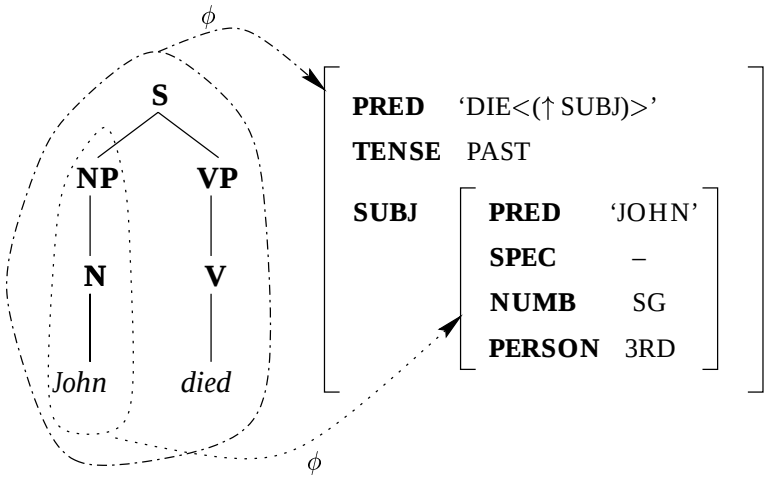


Figure 3.5: C-structure & F-structure correspondence of the sentence "*John died.*"

its passive variance are semantically identical are some of the semantic information captured in an s-structure[5].

An example of an s-structure in LFG, as reproduced from Kaplan et al. (1989, page 316), is shown in Figure 3.6. The s-structure in Figure 3.6 described the number of arguments that this

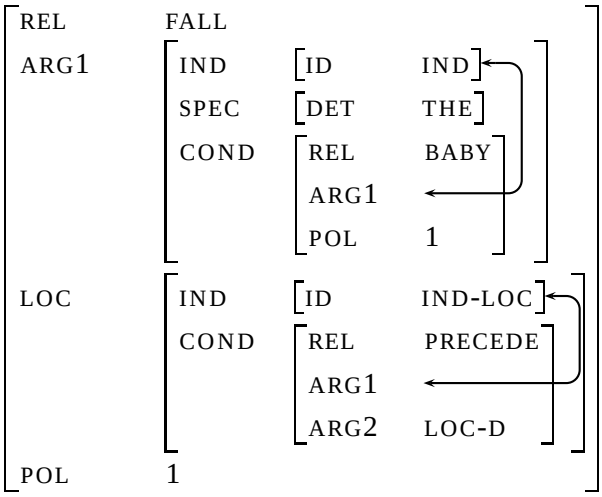


Figure 3.6: S-structure for the sentence "*The baby fell.*"

'fall'-event takes, i.e. ARG1, and who is the entity that involved in this event, i.e. [REL BABY].

[5]For details on how the LFG formalism deals with the representation of semantic information and the kind of semantic information represented by the s-structure, cf. Halvorsen & Kaplan (1988) and Halvorsen (1988).

It also captured the semantic description of the tense of this event as described by the following lexical entry (as reproduced from (Halvorsen & Kaplan 1988, page 287)):

-ed AFF $(\uparrow \text{TENSE}) = \text{PAST}$
$(\sigma\mathcal{M}^* \text{LOC}) = \sigma^*$
$(\sigma^* \text{IND ID}) = \text{IND-LOC}$
$(\sigma^* \text{COND RELATION}) = \prec$
$(\sigma^* \text{COND ARG1}) = (\sigma^* \text{IND})$
$(\sigma^* \text{ARG2}) = \text{LOC-D}$
$(\sigma^* \text{POL}) = 1$

Unlike the previous attempt to interpret the semantics of a sentence, the semantic interpretation proposed by Halvorsen & Kaplan (1988) was not based on an analysis of the f-structure, but on *co-description*, in which different kinds of correspondences (e.g. c-structure to f-structure correspondence and c-structure to s-structure correspondence) appear in a single lexical entry. Instead of relating the s-structure directly with the f-structure, Halvorsen & Kaplan (1988) projects the s-structure directly from the c-structure via the correspondence function σ. As c-structure is relating to the f-structure via the ϕ-function (cf. Section 3.1.2), f-structure can also be related to s-structure indirectly through the inversion of the mapping ϕ (ϕ^{-1}) and the composition of the mappings ϕ^{-1} and σ ($\phi^{-1} \circ \sigma$), i.e. σ'; where the inverted correspondence function ϕ^{-1} gives the c-structure node(s) corresponding to a given f-structure, e.g. given the structures in Figure 3.4, the expression $\phi^{-1}(\uparrow \text{SUBJ})$ gives the set of NP nodes (i.e. NP, N, John) from the c-structure in Figure 3.4. Figure 3.7 shows the correspondence between the c-structure, f-structure and s-structure for the sentence "*John died.*".

The representation of semantic information by s-structure was not among the early proposal on the LFG formalism by Kaplan & Bresnan (1982). When compared with the applications of the c-structure and the f-structure in various disciplines of NLP, the applications of the s-structure in NLP are less significant.

3.2 Lexical-Functional Grammar in Machine Translation

LFG is a unification-based linguistic formalism which is suitable for computation. Since the fundamentals of the LFG formalism were first introduced in late 1970s, many researchers in both theoretical and computational linguistics have explored possible solutions to various problems in describing the linguistic behaviours of natural languages in a formal manner. In addition, different means to apply LFG to MT have been examined by various researchers, e.g. Kudo & Nomura (1986), Kaplan et al. (1989) and Her et al. (1994). It was found that certain aspects of the LFG formalism make it suitable for MT processing.

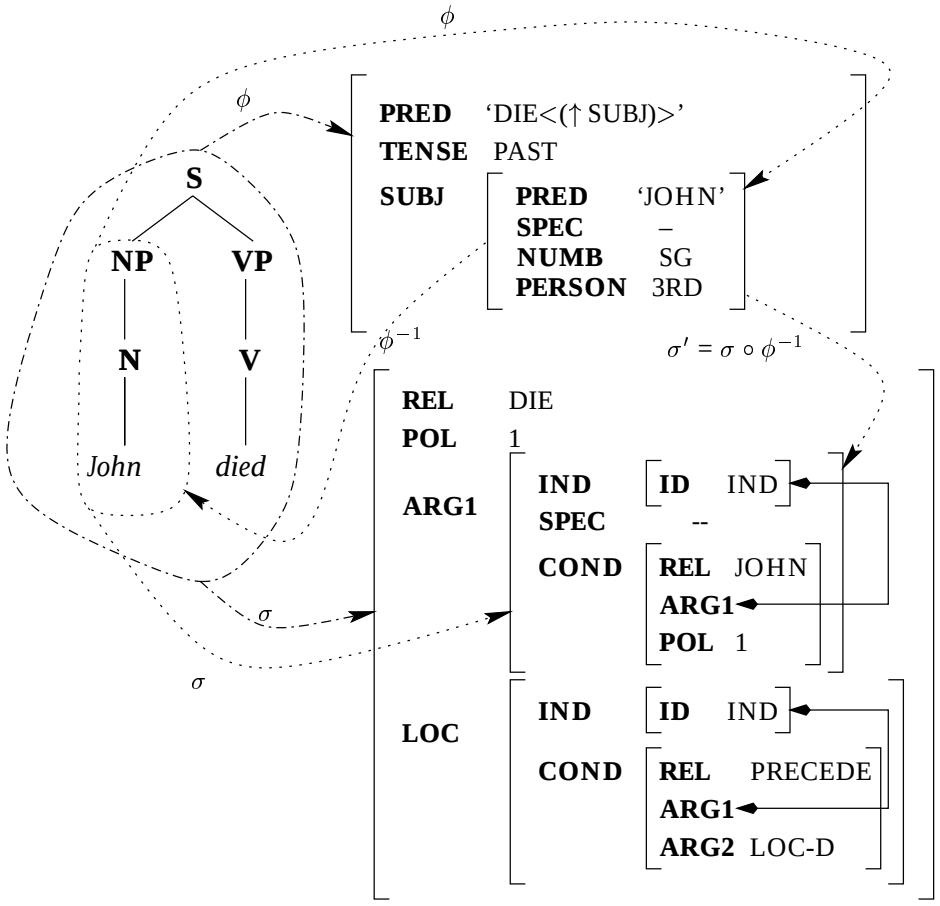


Figure 3.7: C-structure, F-structure and S-structure correspondence of the sentence "*John died.*"

3.2.1 Kudo and Nomura's Lexical-Functional Transfer

Based on the information encoded in f-structures, Kudo & Nomura (1986) performed a so-called "lexical-functional" transfer (LFT) between English and Japanese sentences. Kudo & Nomura adopted the transfer-based approach to their model of MT. They used f-structure as the intermediate representations between the sub-processes (i.e. analysis, transfer and generation) of their MT model. The whole translation process was simply the sequence of:

1. analysing source language sentences and building the corresponding f-structure for each of them,
2. using a two-way dictionary and a set of transfer rules to obtain target f-structure descriptions from the source f-structure,
3. using the resulting target f-structure descriptions to construct the target f-structure; and
4. generating the target sentence based on the target f-structure.

Kudo & Nomura's (1986) work was inspired by the simple representation of lexical and functional information of each lexicon within the LFG framework. However, since the lexical entry for each lexicon was monolingual-based, Kudo & Nomura (1986) defined a set of transfer rules to relate the source language lexicon and their target language equivalents. Those transfer rules were in the form of:

J[(LFG) schemata] <===> **E[(LFG) schemata]**

where 'J' and 'E' correspond to 'Japanese' and 'English' respectively. Examples of transfer rules in Kudo and Nomura's LFT framework are:

$$\text{J}[(\uparrow \text{SUBJ}) = \downarrow] \text{ <===> } \text{E}[(\uparrow \text{SUBJ}) = \downarrow]$$

$$\text{J}[(\uparrow \text{PRED}) = \text{'tomu'}] \text{ <===> } \text{E}[(\uparrow \text{PRED}) = \text{'Tom'}]$$

$$\text{E}[(\uparrow \text{PRED}) = \text{'play<}(\uparrow \text{SUBJ})\ (\uparrow \text{OBJ})\text{>'}] \text{ <===> J} \begin{bmatrix} (\uparrow \text{PRED}) = \text{'suru<}(\uparrow \text{SUBJ})\ (\uparrow \text{OBJ})\text{>'} \\ (\uparrow \text{SUBJ case-marker}) = \text{'ha'} \\ (\uparrow \text{OBJ case-marker}) = \text{'wo'} \end{bmatrix}$$

where the metavariables ↑ and ↓ on the right hand side must correspond to the those on the left hand side. The symbol <===> signifies that the left hand side of a rule corresponds to the right hand side and vice versa. During the transfer, the linguistic information from the source sentence (e.g. (↑ PRED) = 'tomu') was used to obtain the corresponding target sentence translation (i.e. ↑ PRED) = 'Tom').

The lexical-functional transfer was based on selecting the appropriate target f-structure descriptions, in the form of simple functional equations (cf. (9) and (10)), and solving those functional equations to form the target language f-structure. This approach is declarative and mathematically viable, thus it is suitable for computation. However, this approach to sentence transfer is heavily reliant on the use of hand-crafted transfer rules to capture the correspondence between two languages. A considerable amount of human effort and time is required to define such transfer rules. Furthermore, to capture the correspondence between two languages is not a simple task. It is a task requiring some knowledge of linguistics and a good command in both source and target languages. As a result, this approach is not readily adaptable by average programmers.

3.2.2 Kaplan et al.'s approach to MT

Kaplan et al. (1989) observed that the hierarchical representation of linguistic information within the LFG formalism provides a good means to translate sentences from one language to another. Therefore, they proposed an approach to LFG-based Machine Translation (MT) by using the linguistic information encoded in different structures defined in LFG (i.e. c-structure, f-structure and s-structure) and the correspondences between them to perform the transfer. In addition

to the correspondence functions ϕ and σ, Kaplan et al. (1989) introduced two additional correspondence functions τ and τ' to relate the source and target f-structures and the source and target s-structures respectively. Thus, the different structures in the LFG formalism for the source and target languages are inter-related as shown in Figure 3.8. Other conventional ap-

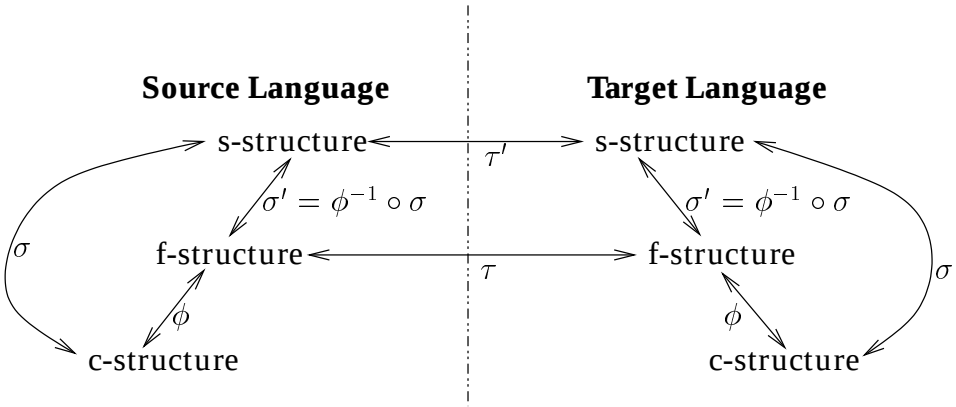


Figure 3.8: The correspondences between different structures for source and target languages in LFG

proaches to LFG-based MT, e.g. the approach proposed by Kudo & Nomura (1986), tend to perform the transfer through the use of a *description-by-analysis* approach to relate source sentence f-structures to their target language counterparts. When comparing Kudo & Nomura's (1986) approach with these approaches, Kudo & Nomura's (1986) approach allowed a more thorough and complete analysis of the source sentences through the use of the whole LFG formalism instead of exploiting only some linguistic aspects defined in it.

In Kudo & Nomura's (1986) approach to MT, the correspondences between different structures of sentences (cf. Sections 3.1.2 & 3.1.3) in different languages played a key role in relating the source and target languages. The translation process involved analysing a source language sentence and obtaining its descriptions in LFG terms. With the translation correspondences already co-described in the form of additional annotations in c-structure rules and lexical entries in the source language grammar, the descriptions for the corresponding target language sentence were obtained from the corresponding co-description embedded in the source language sentence descriptions. The resulting target sentence descriptions, which are in the form of equations, are then resolved and the solutions are then used to generate the target language sentence. This approach is very flexible as it allows the representation of a wide variety of source-to-target correspondences (Sadler 1990). This approach is also computationally viable as the descriptions for the source and target languages are all defined in a mathematical form. Kaplan et al. (1989) suggested that their approach to MT shares some fundamental features with the formalism (i.e. Functional Unification Grammar) described by Kay (1984). However, their approach, as observed by Sadler (1990), is not bidirectional. Therefore, the 'striking' feature, i.e. allowing *a* to translate to *b* only if *b* could translate to *a*, in Kay's (1984) approach to MT is not preserved in Kaplan et al.'s (1989) approach.

Kaplan et al. (1989) showed that *co-description* provides a direct means to relate linguistic information in source and target languages for a systematic transfer between different sentence structures. This method, as exemplified by Kaplan et al. (1989), provides a simple way to transfer sentences which do not have matching target sentence structures, e.g. the sentences "*Der Student beantwortet die Frage.*" in German versus "*L'étudiant répond à la question.*" in French, and "*The baby just fell.*" in American English versus "*Le bébé vient de tomber.*" in French. However, as observed by Sadler (1990), Kaplan et al.'s approach has problems in dealing with the translation involving head-switching and it has difficulty in picking out the appropriate target language units for the translation. In addition, Kaplan et al. (1989) did not address the problem caused by lexical ambiguity when transferring words from one language to another. Hence, it is not clear how they tackle the problem of lexical ambiguity in their approach to MT.

3.2.3 Her et al.'s Lexical and Idiomatic Transfer

Her et al. (1994) attempted to handle lexical disambiguation by using different semantic forms of verbs and additional selectional criteria to transfer words from English to Chinese. However, this method was syntax-oriented and the selectional criteria were developed in an *ad hoc* manner (i.e. there are no formal guidelines to govern the formation of these criteria). This method is, to some extent, inadequate to solve the problem of lexical ambiguity. For instance, the English verb 'tell' can be used to express the meaning: "*to give someone some information*" or simply, "*to deliver information*". These two meanings are expressed by different verbs in Chinese: '告訴' and '說' respectively. The translation to Chinese of the verb 'tell' depends on the meaning of the sentence, e.g.:

(20) John told Mary a matter.
約翰 告訴 了 瑪莉 一 件 事情。
John tell TENSE MARKER Mary one QUANTIFIER matter.

(21) John told a lie.
約翰 說 了 一 個 謊話。
John say ASPECT MARKER one QUANTIFIER lie.

(22) John told a story to Mary.
約翰 給 瑪莉 說 了 一 個 故事。
John to Mary say ASPECT MARKER one QUANTIFIER story.

The semantic forms for the English verb 'tell' in the above sentences are:

- TELL<(↑ SUBJ) (↑ OBJ2) (↑ OBJ)>
- TELL<(↑ SUBJ) (↑ OBJ)>
- TELL<(↑ SUBJ) (↑ OBJ) (↑ TO-OBJ)>

respectively. Her et al. (1994, page 206) suggested that the verb 'tell' should be translated as '說' if the semantic form of the verb 'tell' is 'TELL<(↑ SUBJ) (↑ OBJ)>'; otherwise it should be translated as '告訴'. This approach is capable of producing appropriate translations for the sentences:

- John told Mary a matter.
- John told a lie.
- John told a story to Mary.

However, it will fail to translate the conventional saying "*I told you.*" appropriately. With the semantic form 'TELL<(↑ SUBJ) (↑ OBJ)>', the sentence "*I told you!*" will be translated to:

(23) * 我 說 了 你!
I say ASPECT MARKER you!

in Chinese according to the method suggested by Her et al. (1994). This translation does not make any sense to a native Chinese speaker. The appropriate translation for this sentence should be:

(24) 我 告訴 了 你!
I tell ASPECT MARKER you!

in which the two-character Chinese translation of 'tell' (i.e. '告訴') is used. The semantic form of a verb was expressed in terms of the relevant syntactic functions of a predicate. As syntactic information is language-dependent and is not universal across languages, it, when used on its own, is inadequate to disambiguate the meaning of a verb. A higher level of linguistic information, which is relatively language-independent and can capture the meaning of the sentence to a certain extent, is required to improve the lexical selection.

Another problem in MT studied in Her et al. (1994) is the problem in disambiguating English idioms. The definition of an idiom varies according to the point of view taken by individual computational and theoretical linguists. Her et al. (1994) regarded words which are treated as a single dictionary unit for transfer as idioms, e.g. phrasal verbs and name phrases like 'American Airline'. However, according to Warren (1994):

> "The word 'idiom' is used to describe the 'special phrases' that are an essential part of a language. ... for example, the expression **kick the bucket** seems to follow the normal rules of grammar, although we cannot say 'kick a bucket' or 'kick the buckets', but it is impossible to guess that it means 'to die'. Phrases like **all right**, **on second thoughts**, and **same here**, which are used in everyday English, and especially in spoken English, are 'special' because they are fixed units of language that clearly do not follow the normal rules of grammar."

It can be summarised that idioms refer to a special combination of words which might not conform with normal grammar rules, i.e. in terms of both syntax and semantics, and they might possess a different meaning from their literal meaning. For instance, consider the meaning of the idiom to '*kick the bucket*' in the following sentences:

(25) John knocked down by a car and he *kicked the bucket* instantly.

(26) John *kicks the bucket* out of his way.

The phrase 'kicks the bucket' in sentences (25) and (26) means 'to die' and 'to hit the bucket with one's foot' respectively. Similar to phrasal verbs, which we will consider in detailed in Chapter 5, both the literal and non-literal meanings can be used depending on the context of the sentence. With a sentence like:

- John *kicked the bucket* yesterday.
- John suddenly *kicked the bucket*.

in which both the idiomatic and the literal meanings can be applied, they are considered as highly ambiguous.

As idioms do not need to observe regular grammar rules and they sometimes have meaning different from their literal ones, idioms should be treated as a single dictionary unit for transfer. However, as illustrated by the phrase 'to kick the bucket', the literal meaning of idioms might be more applicable in some cases. If whenever a certain combination of words that can form an idiom is encountered during MT processing, it is translated according to its idiomatic meaning, the target translation obtained might be wrong. Hence, this leads to the question:

When should an MT system consider the idiomatic meaning and when should it not?

In order to resolve the ambiguity arising in idioms, Her et al. (1994) defined a minimal form of f-structure for each idiom. If a phrase satisfies all the requirements of any of the minimal forms, the corresponding idiomatic translation will be used. This method was inspired by the property of English idioms that only a limited number of the morphological forms of a potential idiomatic phrase can possess the idiomatic meaning. For instance, the minimal f-structure[6] for transferring the idiom "*to kick the bucket*" according to Her et al. (1994, page 210) is shown in Figure 3.9. Her et al. (1994) suggested that any phrase which satisfies the minimal f-structure in Figure 3.9 is translated to the corresponding target word / phrase with the idiomatic meaning '*to die*'. Therefore, the sentence "*John kicks the water bucket.*" will be translated to its literal meaning as it does not fully satisfy this minimal f-structure. With sentences like "John kicked

[6]In the minimal f-structure suggested by Her et al. (1994), the syntactic functions appearing in the lexical form of a predicator are not preceded by the meta-variable ↑. Her et al. (1994) did not give a reason for this. One possible explanation for this might be that they put more emphasis on how many and what arguments a lexical form has instead of how the f-structure is built from the relevant lexical entries.

$$\begin{bmatrix} \text{FORM} & \text{'KICK'} \\ \text{PRED} & \langle(\text{SUBJ})\ (\text{OBJ})\rangle \\ \text{VOICE} & \text{ACTIVE} \\ \text{OBJ} & \begin{bmatrix} \text{SPFORM} & \text{'THE'} \\ \text{FORM} & \text{'BUCKET'} \\ \text{NUMBER} & \text{SG} \\ \text{ADJUNCTS} & \text{NONE} \end{bmatrix} \end{bmatrix}$$

Figure 3.9: A minimal f-structure for transferring the idiom "to kick the bucket" suggested by Her et al. (1994)

the bucket yesterday." in which both the idiomatic and the literal meanings can be applied, only the idiomatic translation would be generated.

Her et al. (1994)'s method is relatively simple and straight-forward. However, it does not handle the problem of ambiguity appropriately. With sentences like "John kicked the bucket yesterday." in which both the idiomatic and the literal meanings are valid, the system defaults to only the idiomatic translation. However, if only the context of this sentence, but not any neighbouring sentences, is considered, it is unsafe to rule out any of the possible meanings, therefore, the more appropriate solution to translating this kind of sentences would be generating two translations — one with the idiomatic meaning and the other with the literal meanings.

Furthermore, the more vital problem of Her et al.'s (1994) method is that it cannot distinguish the idiomatic use of a phrase from its literal meaning successfully at all times. For example, it is clear that the phrase 'kick the bucket' in the following sentences:

(27) John kicked the bucket *away*.

(28) John kicked the bucket *out of his way*.

can only be assigned their literal meaning. However, the method suggested by Her et al. (1994) will automatically translate it to the idiomatic meaning 'to die' because the f-structures of both sentences "are subsumed by[7]" the minimal form reproduced in Figure 3.9 (cf. Her et al. (1994, pages 210-211)):

[7]The definition of *subsumption* follows Gazdar & Mellish (1989). For instance, if A *subsumes* B, then A contains less information than B. Similarly, if B *is subsumed by* A, that means B *extends* A or B contains more information than A.

(27) John kicked the bucket away.

$$\begin{bmatrix} \text{FORM} & \text{'KICK'} \\ \text{PRED} & \langle(\text{SUBJ})\ (\text{OBJ})\rangle \\ \text{VOICE} & \text{ACTIVE} \\ \text{TENSE} & \text{PAST} \\ \text{SUBJ} & \begin{bmatrix} \text{FORM} & \text{'JOHN'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{SPFORM} & \text{'THE'} \\ \text{FORM} & \text{'BUCKET'} \\ \text{NUMBER} & \text{SG} \\ \text{ADJUNCTS} & \text{NONE} \end{bmatrix} \\ \text{ADJS} & \begin{bmatrix} \text{FORM} & \text{'AWAY'} \end{bmatrix} \end{bmatrix}$$

(28) John kicked the bucket out of his way.

$$\begin{bmatrix} \text{FORM} & \text{'KICK'} \\ \text{PRED} & \langle(\text{SUBJ})\ (\text{OBJ})\rangle \\ \text{VOICE} & \text{ACTIVE} \\ \text{TENSE} & \text{PAST} \\ \text{SUBJ} & \begin{bmatrix} \text{FORM} & \text{'JOHN'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{SPFORM} & \text{'THE'} \\ \text{FORM} & \text{'BUCKET'} \\ \text{NUMBER} & \text{SG} \\ \text{ADJUNCTS} & \text{NONE} \end{bmatrix} \\ \text{ADJS} & \begin{bmatrix} \text{'OUT OF HIS WAY'} \end{bmatrix} \end{bmatrix}$$

From the study carried out by Her et al. (1994), it is obvious that they have put in a considerable amount of effort to study the different syntactic behaviours and occurrences of the idiom "kick the bucket" in order to derive the minimal form shown in Figure 3.9. However, this is still inadequate to solve the problem of ambiguity posed by this idiom. Again, the problem of Her et al.'s (1994) method in disambiguating idioms is that it is based on syntactic analysis only. Syntactic information is often *insufficient* to disambiguate the meaning of words, phrases and / or sentences. A higher level of linguistic information of sentences is required to alleviate the problem of ambiguity.

3.3 Conclusion

LFG is one of the contemporary linguistic formalisms which is both precisely defined and symbolic, thus it is suitable for computation. This chapter briefly reviewed some of the fundamentals of this formalism. Based on human understanding of natural language, LFG defines several structures (i.e. c-structure, f-structure and s-structure) to capture various levels of linguistic information about a sentence. The idea of a computational linguistic formalism like LFG is to be able to construct and manipulate these structures without implicit understanding of the meaning of the underlying sentences. For this purpose, LFG precisely defines how the relevant linguistic behaviour of words in the language can be precisely encoded as lexical entries and phrase-structure rules. This representation facilitates the parsing of sentences and producing all of the appropriate structures. It also enables the generation of sentences from incomplete structure(s).

The representation of different kinds of linguistic information by means of different structures and the fact that these structures are related with each other make LFG suitable for using as a means to capture the linguistic information of natural language sentences. With the grammar

defined in terms of phrase structure rules and equations, to process LFG can be viewed as resolving these phrase structure rules and equations by means of unification and recursion.

The second half of this chapter reviewed some of the past attempts to build on the LFG formalism to perform MT. By observing the syntactic difference between different languages, some of these researchers (e.g. Kudo & Nomura 1986) define a huge number of complicated transfer rules to bridge the gap between different grammars and use of words in languages. These transfer rules are often derived by the description-by-analysis approach which requires profound knowledge of both source and target language and careful observation of the detail translation correspondences. Thus this method is extremely difficult, tedious and costly. Other researchers (e.g. Kaplan et al. 1989) exploit the hierarchical and inter-related representation of linguistic information in LFG to perform a systematic and flexible transfer between source and target language sentences. However, this method can result in a difficult transfer in certain cases and it did not attempt to solve the problem of lexical ambiguity. Some researchers (e.g. Her et al. 1994) attempt to derive a method to tackle the problem of lexical ambiguity by using the linguistic information encoded in f-structure and exploiting its attribute-value pair representation for introducing additional selectional criteria to aid lexical disambiguation. However, this method is syntax-oriented and the selectional criteria are developed in an *ad hoc* manner (i.e. there are no formal guidelines to govern the formation of these criteria). This method is, again, inadequate to alleviate the problem of lexical ambiguity.

To a certain extent, the transfer methods proposed by Kudo & Nomura (1986) and Kaplan et al. (1989) share some similarities as they both attempt to observe, extract and formulate the translation correspondences between the source and target languages. Though the language pairs that these researchers experimented on did not involve English and Chinese, their methods can be applied to the transfer between English and Chinese sentences. However, the weaknesses of these methods make them difficult to be pursued. If these methods can be improved in such a way that their weaknesses can be minimised, instead of restricting the use of these methods to theoretical linguists, the resulting method would be easier to be employed by computer scientists.

Despite the problems in Kaplan et al.'s (1989) approach to MT, their idea to use the combination of different kinds of linguistic information in performing the transfer has many advantages. One of which is that it enables an MT system to 'understand' the linguistics of a sentence to a greater extent. This, in turn, facilitates the analysis and the generation of sentences. However, as Kaplan et al.'s method, like Kudo and Nomura's lexical functional transfer, relies on the use of transfer rules (e.g. $(\tau \uparrow \text{SUBJ}) = \tau(\uparrow \text{SUBJ})$) to bridge the gap between the syntax of source and target language sentences and these transfer rules are defined manually, it is again not easy to pursue, tedious and costly in terms of both time and human effort involved. If this translation correspondence can be defined automatically by means of a simple computational process, the resulting MT system will be easier and less costly to build. During this research study, it is found that a relatively new extension of the LFG formalism — a-structure and lexical

mapping theory — can provide a solution to some of the weaknesses of the transfer approaches suggested by Kaplan et al. (1989) and Kudo & Nomura (1986). Chapters 4 and 6 will explain the details of a-structure and lexical mapping theory and illustrate how they can be applied to facilitate a relatively simple and straight-forward transfer.

From the inadequacy in Her et al.'s (1994) method to overcome the problem of lexical ambiguity, it is observed that this inadequacy is mainly due to the insufficient knowledge of the meaning of words. If more information about the meaning of the words in the sentences can be introduced to the MT system, the problem of lexical ambiguity would be alleviated to a greater extent. It is found that the introduction of a-structure and lexical mapping theory to an MT process also provides a solution to the inadequacy of Her et al.'s transfer method. Chapter 5 will discuss this issue in detail.

Chapter 4

Argument Structure and Lexical Mapping Theory

With the aim of improving the ability of the LFG formalism to act as a Universal Grammar for language comparison, since the late 1980s the research work on the LFG formalism has branched into the extension of the existing structural representation of syntactic and functional information (i.e. in the form of c-structure and f-structure) to include some level of semantic information (e.g. Halvorsen & Kaplan (1988) and Halvorsen (1988)). Kaplan and Halvorsen's work, as was reviewed in Section 3.1.3, is concerned with the representation of semantic information in a sentence by means of s-structure. The kind of semantic information captured in an s-structure, however, is insufficient to reveal the actual meaning of the words in a sentence (cf. Section 3.1.3). The use of c- and f-structures, as will be discussed in Chapter 6, is inadequate to capture some common aspects of different languages. A higher level of linguistic information, which is more language-independent, is required to capture more of the similarities across languages and to reveal the actual meaning of the words in sentences. This has given rise to the use of argument structure (a-structure) to represent thematic information within sentences. The lexical mapping theory (Bresnan & Kanerva 1989, Huang 1993, Alsina & Mchombo 1993) defines how thematic information represented in a-structures can be mapped onto traditional f-structures for enriching their information expressive power.

The theory of a-structure and the lexical mapping theory are the fruit of many studies on analysing and explaining the different linguistic behaviours of various languages, e.g. Bresnan & Kanerva (1989), Alsina & Mchombo (1993), Huang (1993), Bresnan & Zaenen (1990) and Alsina (1996*a*). The ability of this theory to aid various areas on Natural Language Processing (NLP) has not yet been explored extensively. The work carried out in this study, therefore, aims at experimenting with the ability of this theory to aid Machine Translation (MT). It is found that by incorporating thematic information into traditional f-structures, the ability of f-structures to act as a medium for the transfer in MT improves. This is achieved by reducing lexical and structural ambiguities in source sentences through the introduction of thematic information which

expresses the meaning of words and sentences to a greater extent than syntactic information. Furthermore, a-structure and the lexical mapping theory also provide a good means to carry out the lexical selection for verbs with multiple meanings. Details on how a-structure and the lexical mapping theory aid MT processing will be discussed in the later chapters.

This chapter introduces the concepts of a-structure and the lexical mapping theory as presented by Bresnan and other researchers as well as the adaptations made during the application of these concepts in this study. Though the use of thematic information to represent the predicate argument structures of sentences has been suggested in many independent theoretical linguistic studies, e.g. Carlson (1984), Rappaport & Levin (1988) and Tanenhaus & Carlson (1989), a well-defined method to derive an a-structure from a verb or a predicate is not highly available. In the light of the different definitions of thematic roles or case-roles given by some well-known researchers on this field, this chapter illustrates how to derive the a-structure of a verb. The concept of a-structure and the concept of Case Grammar as suggested by Fillmore (1968, 1977) in the late 1960s share some resemblance. There has been a confusion over whether the idea to apply a-structure and the lexical mapping theory to MT is in fact re-inventing the wheel which had been done some twenty years ago with Case Grammar. This chapter will attempt to clear this confusion by discussing the similarities and differences between these two concepts.

4.1 Thematic Roles

The work on classifying the arguments of predicates[1] into a small set of participant types according to the manner of their involvement in an event, characterised by a process, an action or a state, started in the mid-1960s (EAGLES 1996). The term '*thematic relations*' was used to describe this classification in the mid-1960s and 1970s (Fillmore 1968, Jackendoff 1972). Sells (1985*b*) suggested that the idea to bring thematic relations into syntactic description in a general way started in the early 1980s. Since then, there has been a considerable amount of work on defining a set of thematic roles for describing the role that each of the participants plays within an event structure, and abstracting the relationship between these thematic roles and the syntactic functions appearing in different sentences. However, since the nature of natural languages is infinite, highly irregular and continually evolving, it is very difficult to come up with a classification for the types of arguments that can satisfy every natural language predicate. Up till now, a universally accepted set of guidelines on defining the set of thematic roles and on defining what properties each thematic role in the set possesses is still not available.

Different linguists, therefore, have different interpretations of the types of participants involved in different event structures and their semantic properties. However, there is a set of thematic roles and a list of properties associated with them which are more commonly adopted. The set of thematic roles adopted in this study is the set presented by Bresnan & Kanerva (1989): *agent*,

[1] cf. Section 4.2 for a detail description on what an argument structure looks like.

beneficiary, recipient, experiencer, instrument, patient, theme and locative. This set of thematic roles forms one of the major components of the lexical mapping theory. In the light of the works on case theory carried out by Fillmore (1968), thematic relations carried out by Jackendoff (1972) and case-roles done by Givón (1984), and the definition of these thematic roles adopted in Bresnan & Kanerva (1989) as well as Bresnan (1994), the rest of this section is devoted to explaining the meaning of each of these thematic roles adopted in this study.

4.1.1 Agent

An agent is generally accepted as the *animate* participant who willfully *initiates* the action characterised by the verb. For instance, both Fillmore (1968) and Jackendoff (1972) supported this idea:

> "*. . . the typically animate perceived instigator of the action identified by the verb.*"
>
> [Fillmore (1968, page 24)]

> "*The Agent NP is identified by the semantic reading which attributes to the NP will or volition toward the action expressed by the sentence. Hence only animate NPs can function as Agents.*"
>
> [Jackendoff (1972, page 32)]

Givón (1984) gave a more detailed description on the semantic properties of an agent:

> "*The agent is always a* **conscious** *participant in an event, since he is a* **volitional** *initiator of the change*[2] *. . . In addition to being conscious . . . the agent is also the* **responsible initiator** *of the event. . . . The assumption of responsibility for initiating actions also implies having* **control** *and being subject to* **blame**."
>
> [Givón (1984, pages 88-89)]

Again, this definition also supports the idea that an agent is an animate participant who causes an event to happen. In this study, the definition of the thematic role 'agent' follows the definition given by Givón (1984). In the following sentences, the underlined NPs are examples of an agent:

(29)
- John cooked a meal.
- John knocked Mary down.
- John gave a book to Mary.
- John bought some flowers for Mary.
- John rolled the rock down the hill.
- John dropped the bowl.
- John kept a car in the garden.
- The dog barked.

[2]The 'change' refers to the action described by the verb.

Givón (1984, page 64) considered the NP '*John*' in the sentence "John died." as the agent the 'die'-event described. He gave no clear explanation on why John is considered as the agent of the event. According to one of her explanations on the semantic properties of an agent:

> *"It is the actual initiation of the act — motivated by volition and prompted by decision — which makes the agent an agent."*
>
> [Givón (1984, page 89)]

an agent has the power to decide whether or not to initiate an act. In the real world understanding, it is difficult to imagine John as the agent of the above 'die'-event because he does not have the decision as to whether or not to die. Consider the following sentences:

(30) John died. He killed himself.

(31) John died. He was killed by Mary.

In (30), John initiated a killing event and he died as a result of this event; whereas in (31), John died because Mary initiated a killing event which involved John. By considering the event described by the sentence "John died." alone, although John was an animate object when he was undergoing the dying process, it is difficult to tell if he was the initiator of the 'die'-event or whether he has control over this event. Therefore, in this study, John is *not* considered as the agent of this 'die'-event. To push this distinction a bit further, we can derive that any verb which describes an event that involves only one participant, and this participant has no control over the initiation of this event, this participant cannot be considered as an agent of the event. Other examples of this kind of verbs are 'sink' (as in "The ship sank."), 'freeze' (as in "The river froze.") and 'cook' (as in "The apples cook well.").

From the example sentences shown in the beginning of this section which involve an agent, an agent seems to appear in the subject position of an active sentence. However, not every subject serves as an agent in a sentence. For instance, the NP '*the boat*' in "The boat sank." is not the agent of the event because the boat cannot cause itself to sink. Many linguists observed that if a verb takes an agent as one of its arguments, the agent is very likely to appear as the logical subject. The mapping of a thematic role to the corresponding syntactic function is discussed in Section 4.3.

4.1.2 Beneficiary, Recipient and Experiencer

Givón (1984, page 88) regards the term 'recipient' as a synonym of the case-role 'dative'. The participant 'dative', according to Givón, is a *conscious participant* which is being *in a state* or *undergoing a change*. It also commonly registers a change of mental state, e.g. the NP 'Mary' in "John told Mary a story." and "John taught Mary a lesson.". A dative is often a *conscious goal* of the transaction in an event. Some examples of a dative given by Givón are:

(32)
- John knew the answer.
- John is angry.
- John learned a lesson.
- John informed Mary of the change.
- John gave Mary a book.
- John sent Mary some flowers.

Givón considered the case-roles 'beneficiary' and 'experiencer' as a kind of *dative*.

Fillmore also used the term 'dative' to describe the animate participant who was affected by the state or action identified by the verb (Fillmore 1968, page 24), like Givón, regarding the case-roles 'beneficiary', 'experiencer' and 'recipient' as a kind of dative. However, owing to a different classification and perception of case-roles, some of the examples cited by Fillmore bear a different case-role than the one suggested by Givón[3]. For instance, the following underlined NPs are considered as dative by Fillmore:

(33)
- John died.
- John killed Mary.
- John murdered Mary.
- John terrorised the teacher.

whereas Givón regarded 'John' in "John died." as the agent[4] and the rest of the above underlined NPs as the patient. However, the subject of verbs like 'look' and 'learn' (which Givón regarded as the dative participant of an event) is considered as an agent by Fillmore:

(34)
- John looked at the clock.
- John learnt a vocabulary.

Apart from this difference, some of the examples of a dative case suggested by Fillmore are similar to those suggested by Givón:

[3]This difference provides a good example of the fact that the classification of predicate arguments into thematic roles is not definite. Although both Fillmore and Givón used similar definitions for the terms 'agent' and 'dative', they classified the same predicate arguments into different case-roles. This suggests that the classification of predicate arguments is based on one's understanding of the thematic roles and thus no definite classification for each predicate argument is available. During the classification, so long as the role played by an argument does not violate the definition of the thematic role classified, but reflects the meaning of that thematic role, the classification would be considered as correct.

[4]cf. Section 4.1.1

(35)
- John showed Mary a ring.
- John saw Mary.
- John liked Mary.
- John knew the answer.
- John expected Mary to come.
- John forced Mary to go.

Unlike Fillmore, Givón further divided the dative case into two groups: 'dative-beneficiary' and 'dative-experiencer'. He roughly defined 'beneficiary' as the:

> "*Conscious benefiter from an agent-initiated event*"
>
> [Givón (1984, page 126)]

Givón suggested that the participant which is benefiting from a transaction is a dative-beneficiary. Since the receiver of a transaction can also be viewed as the participant who benefits from the transaction, some examples of a beneficiary suggested by Givón are in fact the receiver of the transaction involved in an event. For instance:

(36)
- John told a story to Mary.
- John showed Mary a doll.
- Mary received a book from John.
- John promised a ring to Mary.
- John asked a favour from Mary.
- John brought Mary a letter.

In addition to appearing as a mandatory argument of a predicate, a beneficiary often appears as an optional argument, as in:

(37)
- John worked for Mary.
- John cooked a chicken for Mary.
- John killed the snake for Mary.
- John made a doll for Mary.

In this study, the terms 'beneficiary' and 'recipient' are used to describe different kinds of arguments. In an event which involves something to be given / delivered from one party to another, the term 'recipient' refers to the conscious participant whom the 'something' is given / delivered to.

The term 'recipient' is used to describe the conscious participant who is the goal of the transaction identified by the verb. In other words, a recipient is the participant who *receives* in the event. For instance, the following underlined NPs are considered as recipients in this study:

(38)
- John gave Mary a book.
- John sent Mary a letter.
- John received a book from Mary.
- John brought Mary a letter.
- John sold Mary a house.
- John showed Mary a doll.
- John asked Mary a question.

The events identified by the verbs in (38) all involve the transfer of something (both physical and abstract) from one participant to another. A recipient is at the receiving end of the transaction involved.

The term 'beneficiary' in this study refers to the conscious participant who *benefits* from the result of an event. Although a beneficiary is a conscious participant of an event, it does not involve in carrying out the event, thus it does not have control over the situation described by the verb. Unlike a recipient whose involvement in an event is clearly identified by the verb, a beneficiary involves in an event in a less direct manner. For instance, the NP 'Mary' in both "John bought some flowers for Mary." and "John cooked Mary a chicken." is the beneficiary of the events described. Take the latter sentence as an example: rather than involvement in the event directly like a recipient (e.g. aiding the completion of the cooking process), Mary benefited from the result of the cooking process. She is also realised as an initiative of this event. The underlined NPs in (37) are some more examples of a beneficiary.

According to Givón (1984, page 100), an experiencer appears as the subject of verbs of cognition, sensation or volition who registers some internal or cognitive change. The object of these verbs does not register discernible impact or change. For instance:

(39)
- John saw Mary.
- The dog sensed the earthquake.
- John understood the situation.
- John knew the answer.
- John looked at the dog.
- John listened to the radio.

With verbs of emotion, it is the experiencer who experiences the emotion, as in:

(40)
- John loves Mary.
- John envied Mary's success.
- John hated Mary.

4.1.3 Instrument

The thematic role 'instrument' is generally used to describe the participant of an event which was used to cause the event to take place. The meaning of an instrument in the case-role sense suggested by both Fillmore (1968) and Givón (1984) reflects this use:

> "*the case of the inanimate force or object causally involved in the action or state identified by the verb.*"
>
> [Fillmore (1968, page 24)]

> "*unconscious instrument used by the agent in bringing about the event*"
>
> [Givón (1984, page 126)]

The instrument of an event often appears as a prepositional phrase which is marked by the preposition 'with'. For instance:

(41)
- John broke a window with a hammer.
- John opened the window with a key.
- John killed Mary with a snake.
- John filled the kettle with water.
- John covered the dog with a blanket.
- John supplied Mary with the information.
- John stabbed Mary with a knife.
- John sprayed the wall with paint.

However, an instrumental case can also appear in a sentence as the subject or be marked by other prepositions, e.g.:

(42)
- The wind opened the window.
- The rock shattered the window.

Fillmore suggested that an instrument is an inanimate object. However, this does not mean that only inanimate object can act as an instrument in an event. In fact, as pointed by Fillmore (1977), any object can function as an agent, an instrument, a patient, etc., depending on the meaning of the sentence. For instance, the snake in "John killed Mary with a snake." can be used by John to strangle Mary to death; alternately, John might have caused the snake to bite Mary in order to kill her. In both of these instances, the snake could not be realised as the responsible initiator of the killing event because it was used by John as an instrument to kill Mary.

4.1.4 Theme and Patient

Unlike the thematic role 'agent', the terms 'theme' and 'patient' are not generally adopted in different proposals on thematic relations. For instance, although the term 'patient' is widely

used in much research work involving Case Grammar, amongst the cases proposed by Fillmore (1968) for describing the general participant types appeared in different event structures, both the theme and the patient roles do not exist. The thematic role 'theme', but not patient, appears in the thematic relations presented by Jackendoff (1972); whereas patient, but not theme, is found in Givón's work (Givón 1984). Although the terms 'theme' and 'patient' are not generally used, the case or role description which is similar to the thematic role represented by patient and theme can be found in all of the above cited works.

Jackendoff (1972) suggested that every sentence contains a theme role. With verbs of motion, the theme is the participant which undergoes the motion; with verbs of location, the theme is the participant whose location is subcategorised by the verb. For instance, in the following sentences, the underlined NPs function as the theme according to the definition given by Jackendoff:

(43)
- The book fell on the floor.
- John gave Mary a book.
- John cooked a chicken in the garden.
- John put the book on the table.
- The book sat on the table.
- The book belonged to John.

According to Givón (1984), a state is an existing condition which does not involve change across time; a patient (also referred to as 'accusative') is the participant who exhibits a state or undergoes the change in state. The underlined NP in the following sentences are some examples of a patient given by Givón:

(44)
- Soon the water warmed up.
- The rock sank first.
- John painted a picture.
- They demolished a house.
- Mary cracked the pot.
- They bleached his hair.
- They moved the barn.
- John kicked the wall.
- Mary washed a shirt.
- John heated the solution.
- John murdered Mary.

Both Jackendoff and Givón did not suggest any distinction over the roles played by a theme and a patient. From the lists of thematic roles suggested by Jackendoff and Givón, they seemed to ignore the possibility that the thematic roles 'theme' and 'patient' should exist together. One possible reason for this is that the *theme* role suggested by Jackendoff and the *patient* role

described by Givón are in fact referring to the same kind of participants in an event structure. For instance, according to Givón's definition of *patient*, the theme NP '*a chicken*' in "John cooked a chicken in the garden." is also functioning as a patient since it was the participant who underwent the change in state (i.e. from uncooked to cooked). The theme NP '*the book*' in "The book belonged to John." can also be considered as a patient in Givón's terms because it was the participant who was in the state of belonging to John. Similarly, since the patient NP '*the barn*' in "They moved the barn." underwent the motion 'move', it is also a *theme* in Jackendoff's sense.

Amongst the cases identified by Fillmore (1968), the terms 'theme' and 'patient' do not exist. However, the *factitive* case and the *objective* case identified by Fillmore resembled the functions of the theme and patient roles suggested by Jackendoff and Givón respectively:

> "*Factitive (F), the case of the object or being resulting from the action or state identified by the verb, or understood as a part of the meaning of the verb.*
> *Objective (O), . . . the case of anything representable by a noun whose role in the action or state identified by the verb is identified by the semantic interpretation of the verb itself; conceivably the concept should be limited to things which are affected by the action or state identified by the verb. The term is not to be confused with the notion of direct object, nor with the name of the surface case synonymous with accusative.*"
>
> [Fillmore (1968, page 25)]

An example of the factitive case reproduced in (Malmkjoer 1991) is the NP '*a wurley*' in "The man makes a wurley.". As suggested by Fillmore, the objective case is not confined to the direct object of a sentence nor does it refer only to the NP appearing as the accusative case in a sentence. An NP that is in an objective case can appear as the subject or the object of an active sentence. For instance, according to Fillmore, the NP '*the door*' in both "The door opened." and "John opened the door." is in the objective case.

The definitions of patient, theme and the objective case given by Givón (1984), Jackendoff (1972) and Fillmore (1968) respectively seem to suggest that the thematic roles 'patient' and 'theme' describe arguments which have very similar, if not the same, semantic properties. No solid distinction between the thematic roles 'patient' and 'theme' could be drawn from these definitions. However, in the lexical mapping theory proposed by Bresnan & Kanerva (1989), the distinction between a theme role and a patient role is significant because an aspect of one of the lexical mapping principles only applies to a theme role but not a patient role.

According to Bresnan & Kanerva (1989, page 76), the argument "*of which location or state is predicated*", or the argument of "*change of location or state*" is the theme; and the argument which displays the locus of the effect is the patient. For instance, the NP 'the river' in "The river froze." and the NP 'the vase' in "John broke the vase." are regarded as the theme; the NP 'Mary' in "John kicked the statue." and the NP 'the statue' in "John kicked the statue." are the patient of these events. A more distinctive difference between a patient role and a theme role is that, unlike the theme role, the patient does not appear as an argument of an intransitive verb

(Bresnan & Kanerva 1989, page 99, Note 29). For instance, the NP 'the doorbell' is a *theme* in both "John rang the doorbell." and "The doorbell rang."; whereas the NP 'Mary' which acts as a *patient* in "John killed Mary." *cannot* appear as an argument in the intransitive form of the verb 'kick' to form "* The statue kicked.". Another distinction pointed out by Bresnan and Kanerva is that the patient, *but not the theme*, can alternate with an irresultative oblique. For instance, one can say "John kicked at the statue." (where the NP 'the statue' is the patient), but not "* *John shattered at the vase.*" where the NP 'the vase' is the theme. In this study, the definition of the thematic roles 'patient' and 'theme' follows the one adopted by Bresnan & Kanerva (1989).

4.1.5 Locative

The term 'locative' (or 'location' in Jackendoff's (1972) term) is generally used to describe the argument of a predicate which expresses the location involved in an event. Jackendoff defined the term 'location' as:

> "*the thematic relation associated with the NP expressing the location, in a sentence with a verb of location*."
>
> [Jackendoff (1972, page 31)]

Some typical examples of a locative argument can be found in verbs of location like '*remain*', '*stand*', '*put*' and '*keep*':

(45)
- John remained <u>at home</u>.
- John stood by <u>the fire</u>.
- John put a book <u>on the table</u>.
- John kept a bike <u>in the shed</u>.

According to Jackendoff, the term 'location' is not restricted to physical location only. Abstract locations, e.g. human states, expressed by adjectives are also subsumed by the term 'location'. Below are two examples of a predicator which takes a physical location or an abstract location as one of its arguments:

(46)
- *stay*
  - John stayed <u>in the room</u>.
  - John stayed <u>single</u>.
- *remain*
  - John remained <u>at home</u>.
  - John remained <u>silent</u>.

Jackendoff separated the locational arguments from the directional ones and he used three thematic relations to describe these arguments: *location, source and goal*. Examples of source and goal are:

(47)
- *source* – John came from London.
  – John bought a book from Mary.
- *goal* – John gave the book away.
  – John drove to Manchester.

As observed by Jackendoff, like locational arguments (i.e. 'location'), directional arguments (i.e. 'source' or 'goal') referred to both physical and abstract directions, e.g.:

(48)
- John went from London to Manchester.
- John went from elated to depressed.

The case-role 'locative' presented by Givón referred to a similar kind of arguments as the thematic relation 'location' suggested by Jackendoff:

> "*Concrete point of spatial reference with respect to which the position or change-in-location of another participant is construed.*"
>
> [Givón (1984, page 127)]

Givón gave some more examples of the locative case-role:

(49)
- John spread the cream on the cake.
- John sprayed the paint on the wall.
- John sticked a drawing on the board.
- John took the dog from Mary.

The definition of the locative case suggested by Fillmore presented a different perspective of the meaning of the term 'locative'. In addition to describing the location of concern involved in the event described by a verb, the term 'locative' is also used to describe:

> "*the case which identified the location or spatial orientation of the state or action identified by the verb.*"
>
> [Fillmore (1968, page 25)]

For instance:

(50)
- Singapore is hot.
- It is hot in Singapore.

Fillmore pointed out that locational and directional arguments do not conflict with each other. The definition of the thematic role 'locative' adopted by Bresnan refers to the argument of a predicate which indicates either the location or the direction:

> "*The term* LOCATIVE *will be used to subsume a broad range of spatial locations, paths, or directions, and their extensions to some temporal and abstract locative domains . . .*"
>
> [Bresnan (1994, page 75)]

The definition of 'locative' presented by Bresnan covers a much broader range of 'location' than the ones suggested by Jackendoff and Fillmore. For instance, the underlined NPs in:

(51)
- John ate breakfast at Tiffany's.
- John ate breakfast at seven-thirty in the morning.

would be considered as a kind of locative argument by Bresnan. The thematic role 'locative' adopted in this study follows the definition given by Bresnan (1994).

4.2 Argument Structure

In the LFG formalism, an a-structure shows the participants involved in an event characterised by a single action, state or process (Bresnan 1995). The participants of an event refer to the entities that take part in the event. They can be either animate or inanimate, physical or abstract objects, etc. For instance, the event structure characterised by the verb '*cook*' as in the sentence:

(52) "John cooked Mary a chicken in the garden."

described three participants involved in this event: *John*, *Mary* and the *chicken* that John cooked. These participants appear in an a-structure in the form of the role each of them played in the event. There are eight different roles identified in the a-structure theory presented by Bresnan & Kanerva (1989): agent, beneficiary, recipient, experiencer, instrument, patient, theme and locative, and they are called *thematic roles*. Although thematic roles do not carry sufficient semantic information to describe the exact meaning of words or phrases that they are assigned to, as discussed in Section 4.1 there is a list of semantic properties associated with each of them. These semantic properties allow the identification of what role each participant plays in an event structure. For instance, in the sentence "The key opened the door.", instead of being regarded as an agent, the NP '*the key*' would be considered as the instrument of this 'open'-event because the key is an inanimate object and it cannot act as an agent who initiates the action described by the verb 'open'[5]. To a certain extent, these semantic properties also help to restrict the kind of words that can appear as a valid argument of a verb.

The event structure described by the verb 'cook' in (52) is represented by the a-structure:

(53) cook<agent beneficiary theme>

An a-structure contains two parts: a head and a list of arguments. The head corresponds to a predicator (PRED) of a sentence which possesses a semantic form and forms the head of the event (e.g. the verb 'cook' in (53)). The list of arguments which the predicator takes are enclosed in a pair of angled brackets and are expressed in terms of thematic roles (e.g. the thematic roles 'agent', 'beneficiary' and 'theme' in (53)). These thematic roles are ordered according to the

[5] cf. Sections 4.1.1 and 4.1.3

thematic hierarchy (Bresnan & Kanerva 1989, Bresnan & Zaenen 1990) (cf. Section 4.3.1). The arguments appearing in an a-structure are the least required participants for characterising the event. If any of the thematic roles appearing in an a-structure cannot be used to describe the role played by a syntactic function in a sentence, either this sentence is ill-formed or it is expressing an event which is different from the a-structure. In the sentence (52) "John cooked Mary a chicken in the garden.", although the prepositional phrase (PP) "in the garden" forms an adjunct to this sentence, the role that it plays (i.e. locative) is absent from the a-structure (53). This is because in the event that the ditransitive verb 'cook' is describing, the role 'locative' is not thematic and it does not aid the characterisation of this 'cook'-event. The removal of the PP "in the garden" from (52) will not affect the well-formedness of this event structure nor change the nature of this event. Adjuncts are often not regarded as an argument of a predicate, thus they tend not to be included in an a-structure.

Although adjuncts play a less important role in characterising an event structure, they introduce additional information about the event, e.g. the location and the duration of an event, to the sentence. This information is often useful to an MT process in producing target translations and thus it must be made available to the process of creating the target sentence. In this study, enriched a-structures are formed by appending sub-structures to an a-structure to represent these kind of adjuncts. For instance, in the following example, the sub-structure "in<locative>" is added to the event structure (53) to become:

(54) cook<agent beneficiary theme> in<locative>

The thematic role 'locative' in (54) is governed by the preposition 'in'. Though the sub-structure "in<locative>" has similar representation as an a-structure, it is *not* an additional a-structure to (53) because the preposition 'in', unlike the verb 'cook', cannot be used to express an event, and thus it cannot form the head of an event structure. In the rest of this thesis, the name "a-structure" refers to this enriched a-structure.

4.2.1 How to establish the a-structure(s) for a verb?

A-structure describes the structure of an event in terms of its participants. A verb often can be used to describe different event structures involving different participants. Thus, more than one a-structure can be associated with the same verb. For instance, consider the following sentences with the verb 'open':

(55)
a. John opened the window.
b. John opened the window with a key.
c. The window opened.
d. The key opened the window.

The participants involved in the events described in (55) are agent (i.e. John), theme (i.e. the

window) and instrument (i.e. the key). Based on the roles played by these participants, different a-structures are used to characterise the event structures described in (55):

(56) a. open<agent theme>
b. open<agent theme> with<instrument>
c. open<theme>
d. open<instrument theme>

While deriving an a-structure from a sentence, the focus is put on the event described by the verb (i.e. the state, process or action identified by the verb). Each participant of an event is derived with respect to the role it plays in the event. Consider the sentences with verbs like 'load' and 'spray':

(57) John loaded the truck with hay.

(58) John loaded hay onto the truck.

(59) John sprayed the wall with paint.

(60) John sprayed paint on the wall.

The sentences (57) & (58) and (59) & (60) appear to be two pairs of paraphrases: both (57) and (58) or (59) and (60) describe the same event with the same participants, i.e. a loading event involving the participants 'John', 'the truck' and 'hay' in (57) and (58), or a spraying event involving 'John', 'the wall' and 'paint'. Based on this similarity, one might start to draw a conclusion that same a-structure should be used to describe both (57) and (58) or (59) and (60). However, according to the definition of the thematic roles presented in Section 4.1, the NPs 'the truck' and 'hay' in (57) are the theme and the instrument of this event respectively; whereas the same NPs in (58) are the location and the theme respectively. Same applies to the NPs 'the wall' and 'paint' in (59) and (60) respectively. Therefore, the a-structures for the verbs 'load' and 'spray' in (57), (58), (59) and (60) are:

- load<agent theme> with<instrument>
- load<agent theme> onto<locative>
- spray<agent theme> with<instrument>
- spray<agent theme> onto<locative>

respectively. In fact, as pointed out by Rappaport & Levin (1988), the meaning of the sentences (57) and (58) has a subtle difference. The sentence (57) "John loaded the truck with hay." implies that the truck is fully loaded with hay; whereas the sentence (58) "John loaded hay onto the truck." does not have this implication — some hay is loaded to the truck, but the truck is not necessarily be fully filled with hay. The difference between the a-structures of (57) and (58) reflects that the two events described in these sentences are not identical.

4.3 Lexical Mapping Theory

According to Bresnan (1995), the function of an a-structure is to act as a link between lexical semantics and syntactic structures. This link is established by mapping the thematic information represented in an a-structure to the corresponding syntactic functions within a sentence. The lexical mapping theory formulates some constraints on how to carry out this mapping. This was done by observing and accounting for the relationship between syntactic and thematic structures. According to Bresnan & Kanerva (1989), the lexical mapping theory comprises four components:

1. hierarchically-ordered thematic structures,
2. classification of syntactic functions,
3. a set of lexical mapping principles for governing the mapping between each thematic role to the corresponding syntactic function, and
4. two well-formedness conditions on lexical forms.

These four components together govern the mapping between the thematic roles in an a-structure and the syntactic functions that appear in a sentence. The rest of this section explains what these components are and illustrates how the lexical mapping can be done.

4.3.1 Thematic Hierarchy

The thematic roles specified within an a-structure are ordered according to the thematic hierarchy:

(61) agent > beneficiary > recipient[6] / experiencer > instrument > patient / theme > locative

where the sequence 'X > Y' means the thematic role 'X' is hierarchically higher than the thematic role 'Y'. As pointed out by Alsina (1996*b*), the proposal to organise the thematic roles in a hierarchical order was motivated by the observation that the position of a thematic role in relation to the other arguments in an a-structure is more relevant to the regularities holding across a-structures involving different thematic roles than the specific thematic role of an argument[7]. That is to say, in generalising the regularities holding across a-structures, abstracting the hierarchical order of thematic roles is more relevant than defining what role an argument plays in an event structure. This is because what role an argument plays in an event structure is dependent on the event. It is impossible to derive a generalisation like: "the first argument of an event structure must be an agent", because this kind of regularity, as exemplified at the end of Section 4.1.1, does not exist across a-structures.

[6] In other citation of this thematic hierarchy (e.g. Alsina & Mchombo (1993) and Huang (1993)), the term 'goal' is used instead.

[7] cf. Alsina (1996*b*, pages 35–36)

The thematic hierarchy reflects the universal hierarchy of thematic roles and it reflects the relative prominence of thematic roles characterised by any given predicator. The thematic role appearing in the left-most position of an a-structure is the most prominent role within the a-structure (i.e. the highest thematic role, $\hat{\theta}$). The prominence of a thematic role decreases from left to right within an a-structure. For instance, in the a-structure (53) 'cook<agent beneficiary theme>', the agent is the most prominent role and the theme is the least prominent role. The most prominent argument in an a-structure is referred to as the logical subject[8]. Many researchers have contributed to the development of the universal hierarchy of thematic roles (e.g. Jackendoff (1972) and Givón (1984)). The thematic hierarchy shown in (61) follows the one presented by Bresnan & Kanerva (1989).

The thematic hierarchy (61) was said to be universal across languages. This suggests that if two a-structures in different languages bear the same thematic roles, these thematic roles should appear in the same order. However, as observed by Huang (1993), the data from the Chinese language do not support this claim. Although the a-structures of the English verb 'give' and its Chinese counterpart '送' have the same arguments (i.e. agent, recipient and theme), these arguments are ordered differently:

'give<*agent recipient theme*>' *versus* '送<*agent theme recipient*>'

Hence Huang suggested a different thematic hierarchy for Chinese[9]:

(62)

agent > beneficiary / malificiary > instrument > *patient* / *theme* > *recipient* / *experiencer* > locative

Note that in the above thematic hierarchy, unlike the thematic hierarchy for English (cf. (61)), the thematic roles 'patient' and 'theme' are hierarchically *higher* than the thematic roles 'recipient' and 'experiencer'. The difference between the thematic hierarchies (61) and (62), as it will be discussed in Chapter 6, facilitates the transfer of passive sentences from English to Chinese.

4.3.2 Classification of Syntactic Functions

There are four kinds of syntactic functions identified in the lexical mapping theory: subject (SUBJ), object (OBJ), object$_\theta$ (OBJ$_\theta$) and oblique$_\theta$ (OBL$_\theta$) (where the subscript θ refers to the specific thematic role associated with the syntactic function, e.g. object$_{theme}$, oblique$_{beneficiary}$ and oblique$_{agent}$). By observing the manner of these syntactic functions appearing in different sentences, these syntactic functions are therefore grouped according the features $[\pm r]$ and $[\pm o]$, where '**r**' stands for *thematically restricted* and '**o**' stands for *objective*. If a syntactic function is

[8] cf. Alsina (1996*b*, page 36)

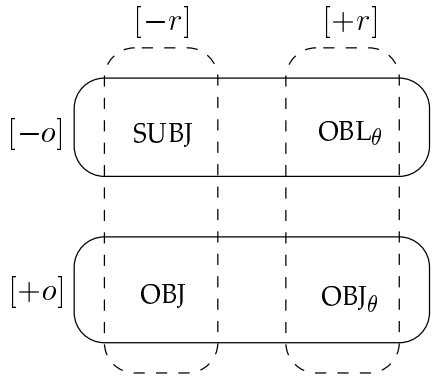
[9] Huang (1993) introduced the term 'malificiary' to describe the participant who suffers from the event. Huang also used the term 'goal', instead of 'recipient', to denote the participant who receives in the event.

classified as $[+r]$, its thematic role is fixed. A syntactic function which is classified as objective (i.e. $[+o]$) means that it is functioning as a kind of object (i.e. direct object or indirect object) within the sentence. The syntactic functions are classified as:

(63) $\text{SUBJ}\ \begin{bmatrix}-r\\-o\end{bmatrix}$ $\text{OBJ}\ \begin{bmatrix}-r\\+o\end{bmatrix}$ $\text{OBL}_\theta\ \begin{bmatrix}+r\\-o\end{bmatrix}$ $\text{OBJ}_\theta\ \begin{bmatrix}+r\\+o\end{bmatrix}$

From this classification, these syntactic functions can be grouped in the following way:

(64)



4.3.3 Lexical Mapping Principles

Through the use of the features $[\pm r]$ and $[\pm o]$, Bresnan & Kanerva (1989) presented three kinds of principles to associate thematic roles with partial specifications of syntactic functions in verbal argument structures. These principles are:

- the intrinsic classifications of some thematic roles,
- the effect of morpholexical operations on the arguments of an a-structure, and
- the default classifications of the arguments of an a-structure.

The application of these principles to an a-structure is subject to the preservation of information. This means that the application of a lexical mapping principle can only add feature(s) to an argument of an a-structure: it cannot alter or delete the existing feature(s) of the argument. These principles govern the assignment of the $[\pm r]$ and $[\pm o]$ features to each of the thematic roles appearing in an a-structure. After this assignment of features, the thematic roles in an a-structure can then be mapped to the corresponding syntactic functions in a sentence through simple feature matching.

Intrinsic Role Classifications

After observing the behaviours of some thematic roles across languages, Bresnan & Kanerva (1989) found that some thematic roles, similar to the syntactic functions SUBJ, OBJ, OBJ_θ and OBL_θ, can be classified as non-objective (i.e. $[-o]$) or not thematically restricted (i.e. $[-r]$). These classifications of some thematic roles — named *intrinsic role classifications* — assign features to

thematic roles which are common across languages. Bresnan & Kanerva (1989) suggested that the thematic roles 'agent', 'theme', 'patient' and 'locative' are *intrinsically* classified as:

(65) *agent* $[-o]$ *theme / patient* $[-r]$ *locative* $[-o]$

This means that cross-linguistically the thematic role 'agent' does not appear as an object and the thematic role 'theme' or 'patient' is an unrestricted function. Cross-linguistically, the thematic role 'locative' also does not appear as an object; it either appears as a subject or an oblique in a sentence[10].

The intrinsic role classifications presented by Bresnan & Kanerva (1989) are restricted to a limited number of thematic roles (i.e. agent, theme / patient and locative). If an a-structure subcategorises thematic roles other than agent, theme / patient and locative, these classifications would fail to assign appropriate features to these thematic roles. When compared with Bresnan & Kanerva (1989), the version of intrinsic role classifications presented by Alsina & Mchombo (1993) is more complete. Instead of classifying in terms of thematic roles, Alsina & Mchombo (1993) presented the intrinsic role classifications in terms of the nature of the arguments within a-structure. These classifications apply to not just some, but all arguments within an a-structure.

According to Alsina, there are three kinds of arguments that a predicate can take:

> *"An external argument is the kind of argument that, in accusative languages like English, must map to the subject function. An internal argument is the kind of argument that can be alternatively assigned to an object or to a subject function. Arguments that are neither internal nor external, which we can call indirect arguments, can only be expressed as oblique functions."*
>
> [Alsina (1996*a*, page 8)]

Instead of presenting the arguments in an a-structure of a predicate in terms of thematic roles, an a-structure can be represented in terms of this classification of arguments. For instance, the a-structures for "John gave Mary a book.", "John broke the vase.", "The vase broke." and "John told a story to Mary." can be written as:

- give<Ext.θ Int.θ Int.θ >
- break<Ext.θ Int.θ >
- break<Int.θ >
- tell<Ext.θ Int.θ θ >

respectively; where 'Ext.θ', 'Int.θ' and 'θ' stand for 'external argument', 'internal argument' and 'indirect argument' respectively. Alsina and Mchombo observed that the intrinsic classification of a thematic role is sensitive to whether the thematic role is an internal argument or

[10] However, Bresnan & Kanerva (1989) pointed out that a locative role can be classified as $[+o]$ (i.e. objective) intrinsically if it is introduced by an applicative morpheme, and some verbs in Chichewa take locative objects.

not. An internal argument can be a theme, a patient, an applied argument[11] or the causee in direct causation (Alsina & Mchombo 1993). An internal argument is classified as $[-r]$ (i.e. non-restricted). The internal arguments which are hierarchically lower than the recipient can sometimes be classified as $[+o]$ (i.e. objective). This classification (i.e. recipient $> \theta$) applies to many languages (e.g. English and Chicheŵa) but Chinese because Chinese has a different thematic hierarchy (cf. the thematic hierarchies (61) and (62)). In Chinese, the internal arguments which are hierarchically lower than the *theme* (i.e. theme $> \theta$) are sometimes classified as $[+o]$ instead. This classification allows some internal arguments (e.g. one of the arguments in the a-structure "give<Ext.θ Int.θ Int.θ >") to be an object$_\theta$. The thematic roles which do not appear as an internal argument are classified as $[-o]$ (i.e. non-objective).

The intrinsic role classifications presented by Alsina and Mchombo can be summed up as:

(66)

External & Indirect Arguments	**Internal Arguments**	
θ	θ	recipient $>$ θ
$\vert$	$\vert$	$\vert$
$[-o]$	$[-r]$	$[+o]$

Morpholexical Operations

Morpholexical operations change the arguments of an a-structure by adding or suppressing the thematic roles within it. Alsina & Mchombo (1993) cited two examples of morpholexical operations which affect a-structures, they are *passive* and *applicative*. The morpholexical operation 'passive' affects an a-structure by *suppressing* the highest thematic role in the a-structure. For instance, after the morpholexical operation 'passive', the highest thematic role 'agent' within the a-structure for "John gave Mary a book." (i.e. 'give<agent recipient theme>') is suppressed:

(67)

give	<	agent	recipient	theme	>
		$\vert$			
		$\emptyset$			

and the passive sentence "Mary was given a book." is formed. This suppressed thematic role can appear in a passive sentence as an optional argument, as in "Mary was given a book *by John*.".

The morpholexical operation 'applicative', however, introduces an additional internal argument to the a-structure of a verb. This allows a role which normally appears in a sentence as an oblique to be expressed as a direct argument[12]. For instance, the Chinese verb '踢' (meaning

[11] Alsina & Mchombo (1993) used the term 'applied argument' to describe the argument introduced as a result of the morpholexical operation 'applicative' (cf. Section 4.3.3), e.g. the applicative verb '*gul-ir*' (meaning '*buy for*') in Chicheŵa takes an applied beneficiary.

[12] A direct argument can either be an external argument (i.e. subject) or an internal argument (i.e. direct and indirect objects). The direct argument here refers to internal arguments only.

'*kick*') takes two arguments: *agent* and *patient*, e.g.:

(68) 約翰 踢 了 一 個 球。
John kick ASPECT MARKER one QUANTIFIER ball.
John kicked a ball.

it can also subcategorise an optional recipient as an oblique, as in:

(69) 約翰 給 瑪莉 踢 了 一 個 球。
John *to* *Mary* kick ASPECT MARKER one QUANTIFIER ball.
John kicked a ball to Mary.

By introducing the character '給' (meaning ≈ '*give to*') to '踢' to form '踢給'[13]:

(70) 踢 < agent patient ∅ >
|
踢給 < agent patient *recipient* >

the optional recipient '瑪莉' in (69) would then appear in the sentence as an object$_\theta$[14]:

(71) 約翰 踢給 了 瑪莉 一 個 球。
John kick-to ASPECT MARKER Mary one QUANTIFIER ball.
John kicked a ball to Mary.

Note that the additional argument 'recipient' introduced to the a-structure for '踢給' in (70) appears after the argument 'patient'. This is because, according to the thematic hierarchy for Chinese (cf. (62)), the recipient is thematically lower than the patient. The introduction of the character '給' to a verb so as to allow an additional argument 'recipient' to be included in an a-structure is common to many Chinese verb, e.g. from '供應' to '供應給' and from '教'

[13]Unlike many European languages, Chinese usually being considered as having almost no inflectional morphology (cf. Hutchins & Somers (1992, page 15)). Although the derivational morphology in the Chinese language is not as obvious as the European languages like English and German, this does not mean that every word in Chinese forms a root form of its own. In fact, as pointed out by Lyons (1968, page 188):

> *"Chinese is often cited as a well-known example of the isolating type of language; but nowadays scholars appear to agree that many Chinese words are composed of more than one morpheme . . . "*

Though Chinese words are made up of one and often more characters, each character in fact has similar functions to a morpheme. In the Chinese language, as each Chinese character bears some meaning and can often be used as a single-character word, many words which bear more complicated meanings are formed by *compounding* the relevant single-character words. For instance, the Chinese noun '執筆者' (meaning '*writer*') is composed of three characters which means 'to hold' (執), 'pen' (筆) and 'person' (者) respectively. When putting these meanings together, the compounded meaning '*the person who hold the pen*' is formed. In an article, the one who hold the pen can be understood as the one who writes it. That is why '執筆者' mean '*writer*' in Chinese. Similarly, the verb '踢給' (meaning '*kick-to*') is formed from '踢' and '給' in Chinese.

[14]The object$_\theta$ in English and in Chinese appear in a different positions. Bresnan & Kanerva (1989) pointed out that the object$_\theta$ in an English ditransitive sentence is the direct object of the sentence; whereas, according to Huang (1993), the object$_\theta$ in a Chinese ditransitive sentence appears in the indirect object position of the sentence.

to '教給'. The morpholexical operation 'applicative' often applies to optional arguments like beneficiaries, recipients, instruments, etc.

The effects of the two morpholexical operations cited above have on a-structures are:

(72)

Passive	**Applicative**
$\hat{\theta}$ | $\emptyset$	$\emptyset$ | $< \cdots \theta_{appl}$[15] $\cdots >$

Default Role Classifications

The default role classifications apply after the thematic roles of an a-structure have received the corresponding intrinsic role classifications and the a-structure has already undergone all the relevant morpholexical operations. Alsina & Mchombo (1993) suggested that the default role classifications are sensitive to the hierarchical order of the thematic role in an a-structure. As mentioned in Section 4.3.1, the highest thematic role in an a-structure corresponds to the logical subject; whereas the lower roles often corresponds to non-subject. The default role classifications are designed to facilitate this correspondence (Bresnan & Kanerva 1989).

In the lexical mapping theory, the default role classifications assign the feature $[-r]$ or $[+r]$ to some of the thematic roles within an a-structure. Recall that thematic roles within an a-structure are hierarchically ordered (cf. Section 4.3.1). The default role classifications assign the feature $[-r]$ (i.e. unrestricted) to the highest thematic role ($\hat{\theta}$); whereas the feature $[+r]$ (i.e. restricted) is assigned to remaining roles in the a-structure:

(73)

$$\begin{array}{cc} \hat{\theta} & \theta \\ | & | \\ [-r] & [+r] \end{array}$$

However, there is one exception to this classification: the thematic role 'locative' can be optionally classified as $[-r]$ — though the hierarchical position of the thematic role 'locative' is not higher than the theme — when the theme is the highest expressed role:

(74)

$$\begin{array}{ccccc} & & \hat{\theta} & & \\ & & | & & \\ < & \text{theme} & \cdots & \text{locative} & > \\ & & & | & \\ & & & [-r] & \end{array}$$

The aim of this special classification is to facilitate *locative inversion*.

[15] 'θ_{appl}' stands for the thematic role of an applied argument.

According to Bresnan & Kanerva (1989), each lexical mapping principle can only be applied if it *adds* feature(s) to the thematic roles in an a-structure. If a lexical mapping principle would cause the existing features of the thematic roles in an a-structure to change or to be removed, it cannot be applied to the a-structure. Therefore, when applying the default role classifications, the thematic roles which have already been classified as $[-r]$ intrinsically are not subject to the default classification of the feature $[+r]$.

4.3.4 Well-formedness Conditions

After classifying the thematic roles in an a-structure according to the lexical mapping principles, the resulting lexical form must satisfy two well-formedness conditions (Bresnan & Kanerva 1989, Alsina & Mchombo 1993):

1. *Subject Condition:*
   Every verbal lexical form must have a subject.
2. *Function-argument Biuniqueness:*
   Every expressed thematic role in an a-structure must map to a unique syntactic function, and every syntactic function must map to a unique thematic role.

4.4 Lexical Mapping — A Demonstration

As discussed in Section 4.3, syntactic functions and the thematic roles within an a-structure are classified with the features $[\pm r]$ and $[\pm o]$. The mapping of thematic roles to the corresponding syntactic functions is mainly based on the matching of these features. In this section, we are going to look at how the lexical mapping theory can be practically used to guide the mapping between a-structure arguments and the corresponding syntactic functions within a sentence.

4.4.1 With the Verb 'give'

The verb 'give' can be expressed in two forms: ditransitive (as in "John gave Mary a book.") and transitive with oblique (as in "John gave a book to Mary.")[16]. The a-structure of the ditransitive form of 'give' is mapped to the corresponding syntactic functions in the sentence "John gave Mary a book." as follows:

[16]Note that the English verb 'give' is different from the Chinese verb '踢' because it can appear as a ditransitive verb or a transitive verb without the introduction of any additional root form, prefix or suffix; whereas the verb '踢', as exemplified in Section 4.3.3, can only behave as a ditransitive verb when it is compounding with '給'.

(75) Sentence : John gave Mary a book.

A-structure :	*give* <	agent	recipient	theme	>
Intrinsic :		$[-o]$	$[-r]$	$[+o]$	
Default :		$[-r]$		$[+r]$	
Syntactic Functions :		SUBJ	OBJ	OBJ_{th}	
NPs :		John	Mary	a book	

According to the intrinsic role classifications (66), the external argument 'agent', the internal argument 'recipient' and the internal argument 'theme' (which is hierarchically lower than the recipient) are classified as $[-o]$, $[-r]$ and $[+o]$ respectively. By default, the agent (which is the $\hat{\theta}$ of the a-structure) and the theme are classified as $[-r]$ and $[+r]$ respectively. Since the recipient has already been classified as $[-r]$ (i.e. thematically unrestricted) intrinsically, the default classification $[+r]$ does not apply to it.

With the classification $\left[\begin{smallmatrix}-r\\-o\end{smallmatrix}\right]$, the agent in (75) is mapped with the subject of the sentence. With the classification $[-r]$, the recipient in (75) can be mapped to either the subject or the object of the sentence[17]. However, the well-formedness condition 'function-argument biuniqueness' constrains the number of arguments to be mapped to the subject position to one. As the subject has already been mapped with the agent of the sentence which bears the classification $\left[\begin{smallmatrix}-r\\-o\end{smallmatrix}\right]$, the function-argument biuniqueness condition blocked the theme in (75) from mapping to the subject position. Therefore, the recipient in (75) is mapped to the object position instead.

The a-structure of another form of 'give' can be mapped with the following syntactic functions:

(76) Sentence : John gave a book to Mary.

A-structure :	*give* <	agent	theme	>	*to* <	recipient	>
Intrinsic :		$[-o]$	$[-r]$			$[-o]$	
Default :		$[-r]$				$[+r]$	
Syntactic Functions :		SUBJ	OBJ			OBL_{recip}	
NPs :		John	a book			Mary	

The assignment of a default feature is subject to the principle of preservation of information. Thus, even though the theme in (76) is not the highest thematic role ($\hat{\theta}$), it cannot be assigned with the default feature $[+r]$ as it has already been classified as $[-r]$ intrinsically.

After applying the intrinsic and default role classifications to the agent in (76), this thematic role bears the features $\left[\begin{smallmatrix}-r\\-o\end{smallmatrix}\right]$. It is therefore mapped to the syntactic function 'subject'. The theme is intrinsically classified as thematically unrestricted (i.e. $[-r]$). Without further specification of this thematic role, it can be mapped either to the subject or the object. Since the agent has already been mapped to the subject position, the theme is mapped to the object position. According to the intrinsic classification (66), the indirect argument 'recipient' is classified as

[17] The syntactic functions 'subject' and 'object' are classified as $\left[\begin{smallmatrix}-r\\-o\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}-r\\+o\end{smallmatrix}\right]$ respectively.

$[-o]$; and since it is not the $\hat{\theta}$ in the a-structure, it is classified as $[+r]$ by default. Bearing the features $\left[\begin{smallmatrix}-o\\+r\end{smallmatrix}\right]$, the recipient in (76) is mapped to the oblique function 'OBL_{recip}'.

4.4.2 With the Morpholexical Operation 'passive'

Consider the lexical mapping between the a-structure "kick<agent patient>" and the sentence "John kicked the dog.":

(77) Sentence : John kicked the dog.

A-structure :	*kick* <	agent	patient	>
Intrinsic :		$[-o]$	$[-r]$	
Default :		$[-r]$		
Syntactic Functions :		SUBJ	OBJ	
NPs :		John	the dog	

The agent and the patient in (77) are mapped to the subject and object respectively.

With the morpholexical operation 'passive', the highest thematic role ($\hat{\theta}$) specified in an a-structure is suppressed. The suppressed *agent* can appear in a passive sentence in the form of an oblique function governed by the preposition '*by*'. The effect of 'passive' on an a-structure is dealt with as follows:

(78) Sentence : The dog was kicked by John.

A-structure :	*kick* <	agent	patient	>	*by* <	agent	>
Intrinsic :		$[-o]$	$[-r]$			$[-o]$	
Passive :	*be-kicked*	$\emptyset$					
Default :						$[+r]$	
Syntactic Functions :			SUBJ			OBL_{ag}	
NPs :			The dog			John	

After being suppressed by the morpholexical operation 'passive', the $\hat{\theta}$ in (78) (i.e. the agent) can no longer be expressed as the subject of the sentence. Bearing the feature $[-r]$, the patient can be mapped to either the subject or the object position. The subject condition, which stated that every lexical form must have a subject, constrains the patient to be mapped to the subject position since there is no other thematic role in (78) can appear as a subject.

4.4.3 With the Morpholexical Operation 'applicative'

As illustrated in Section 4.3.3, the verb '踢' subcategorises an agent and a patient:

(79) 約翰 踢 了 那 個 球。
John kick ASPECT MARKER that QUANTIFIER ball.
John kicked that ball.

By introducing '給' to '踢', the resulting applicative verb '踢給' (meaning '*kick to*') subcategorises an additional internal argument 'recipient':

(80) 約翰 踢給 了 瑪莉 那 個 球。
John kick-to ASPECT MARKER Mary that QUANTIFIER ball.
John kicked that ball to Mary.

Unlike the syntactic structure of an English sentence, the NPs '瑪莉' and '那個球' in (80) are the object$_\theta$ and the object of the sentence (cf. Huang (1993)). The mapping between the a-structure for '踢給' and (80) is as follows:

(81) Sentence : 約翰踢給了瑪莉那個球。

A-structure :	踢 <	agent	patient	∅ >
Applicative :	踢給			recipient
Intrinsic :		$[-o]$	$[-r]$	$[+o]$
Default :		$[-r]$		$[+r]$
Syntactic Functions :		SUBJ	OBJ	OBJ$_{recip}$
NPs :		約翰	瑪莉	那個球

As shown in (81), the intrinsic and default role classifications assigned the features $\left[\begin{smallmatrix}-r\\-o\end{smallmatrix}\right]$ and $[-r]$ to the agent and the patient in the a-structure of '踢給' respectively. These thematic roles are mapped to the subject and the object respectively. The applied recipient in (81) is intrinsically classified as $[+o]$ (i.e. objective) since it is hierarchically lower than the patient in the thematic hierarchy for Chinese. With the feature $[+r]$ (i.e. thematically restricted) by default, this applied recipient is mapped to the object$_\theta$ of (80).

4.5 Is A-structure another variant of Case Grammar?

In LFG, a-structure is used to express the arguments subcategorised by a predicate in terms of thematic relations. A-structure and the case frame presented in the Case Grammar theory proposed by Fillmore (1968) share some similarities. Thus, to some researchers in NLP, the theory of a-structure in the LFG formalism seems to be a variant of Case Grammar. Owing to this confusion, this section gives a brief review on Fillmore's (1968) Case Grammar and compares it with the theory of a-structure and the lexical mapping theory in LFG.

4.5.1 Case Grammar

According to Anderson (1994*a*) and Bruce & Moser (1992), the term *'case'* is used to describe different syntactic roles of a noun as a result of its various morphological forms (e.g. a morphological form of the pronoun 'I' is 'my') in the Greco-Roman tradition. The meaning of case was later extended by other linguists to describe the representation of mainly the semantic roles of the NPs in a sentence with respect to the verbs. *Case Grammar* is a theory developed by Charles Fillmore in the late 1960s which defines the different meaningful combinations of NPs and verbs by the use of their semantic roles within a sentence. For instance, in the sentence:

(82) John ate his meal with a pair of chopsticks.

the three NPs 'John', 'his meal' and 'a pair of chopsticks' involved in the 'eat' event correspond to the agent (i.e. the participant who initiated the event and performed the action), the patient (i.e. the participant which received the action 'eat') and the instrument (i.e. the participant which was used to perform the action 'eat') of this event respectively. Under the definition of the verb 'eat', the agent should be an animate object who can initiate the 'eat' event, e.g. fish, insect, human being, etc.; the patient should be a physical object which is edible by the agent; and the instrument should be another physical object which serves as a tool for eating. Though the semantic properties of each participant identified in the Case Grammar theory aid the formation of a grammatical sentence, they do not serve as a constraint to the well-formedness of a sentence. For instance, in some cases, an 'eat' event consists of an inanimate agent (e.g. "Worry is eating John.") can also form a grammatical sentence.

The idea of representing the structure of a sentence in terms of the the semantic roles of the NPs in the sentence springs from the idea that *"meanings are relativized to scenes"* (Fillmore 1977). Fillmore observed that someone who understands the meaning of a verb would automatically realise the various participants relating to the event described by the verb; and the linguistic knowledge of the verb that he / she has would allow the production of a grammatical sentence with this verb (cf. (Fillmore 1977, pages 65–66)).

In Fillmore's (1968) case system, a universal structure of sentences is captured in a set of rules:

(83) $\text{Sentence} \longrightarrow \text{Modality} + \text{Proposition}$
$\text{Proposition} \longrightarrow \text{Verb} + \textit{Case}_1 + \textit{Case}_2 + \cdots + \textit{Case}_n$
$\textit{Case}_i \longrightarrow [\text{Preposition} \mid \text{Postposition} \mid \text{Case Affix}] + \text{Noun Phrase}$

According to the above structural rules, a sentence comprises two parts: a modality and a proposition. A proposition consists of a verb and one or more cases. A case is often referred to as a form of an NP, but it can also be an embedded sentence[18]. Fillmore identified a set of cases for his case system: agentive (A), instrumental (I), dative (D), factitive (F), locative (L)

[18]The rule for a case appearing as an embedded sentence is not shown here.

and objective[19] (O) (cf. Fillmore (1968, pages 24-25)). To aid ensuring that a proposition is semantically well-formed, features like $[\pm animate]$ are used to constrain the semantic properties of each case role.

4.5.2 A-structure and Case Grammar — A Comparison

In Case Grammar, the term *'case frame'* refers to the case environments provided by a sentence. For instance, the case frames of the sentences "John hit the dog with a stick." and "The key opened the window." are [— O + I + A] and [— O + I]. The structure of a case frame is fairly similar to the a-structure presented by Bresnan & Kanerva (1989). Both of them specified the participants involved in an event described by a verb, and these participants are ordered in a particular manner within the frame/structure. However, the cases and their hierarchical order identified by Fillmore are different from the thematic hierarchy suggested by Bresnan and Kanerva (cf. Section 4.3.1).

According to Fillmore, the function of a case frame is:

> "*to provide a bridge between descriptions of situations and underlying syntactic representations*"
>
> [Fillmore (1977, page 61)]

and this is accomplished by:

> "*assigning semantico-syntactic roles particular participants in the (real or imagined) situation represented by the sentence. This assignment determines or constrains the assignment of a perspective on the situation.*"
>
> [Fillmore (1977, page 61)]

A-structure also has a similar function to case frame, but, unlike a case frame, it focuses more on the semantics of a lexicon than the real-world situations. Instead of bridging the gap between descriptions of situations and underlying syntactic representations, a-structure acts as a link between lexical semantic and syntactic structures by abstracting the relationship between them. An example of this difference is shown in the use of a generalised case frame in Fillmore's Case Grammar. As illustrated in (55) of Section 4.2.1, a verb can be used to subcategorise different kinds of participants in an event. Fillmore suggested a more general case frame for describing the different sets of case environments provided by the sentences in (55): +[— O (I) (A)]; where the parentheses indicate an optional case. In the concept of a-structure, the thematic roles specified in an a-structure are the least required arguments to characterise the event described by the verb, thus no generalisation like that in case frames is made on a-structure.

[19]The term 'objective' refers to the case role whose meaning is identified by the semantic interpretation of the verb governing it (cf. Section 4.1.4). Fillmore did not use this term to describe the case when an NP appears as an object in a sentence.

A-structure describes the necessary participants for characterising an event. For instance, in a buy-event, the a-structure is: 'buy<agent theme>'. Any participant which does not involve in characterising this event (i.e. distinguishing the buy-event from other events) is not included in the angled-brackets. Case Grammar does not seem to have a restriction on the number of cases that can be included in a case frame. The inclusion of a case in a case frame is relatively flexible; so long as the cases are relevant to a particular case frame, they can be added to the case frame. For instance, with a commercial event described by the verb 'buy', the relevant cases are agent, theme, source, goal, instrument, etc., and they can all be included in the case frame.

As suggested by Fillmore, Case Grammar is NOT a complete grammar formalism:

> *"My proposal did not cohere into a model of grammar. Instead, they were suggestions about a level of organization of a clause that was relevant to both its meaning and its grammatical structure; and that offered convenient classifications of clause types."*
>
> [Fillmore (1977, page 62)]

> *"the deep case proposal was not intended as a complete model of grammar, but only as a set of arguments in favour of the recognition of a level of case structure organization of sentences."*
>
> [Fillmore (1977, page 68)]

A computer system built based on Case Grammar alone would not be adequate for Machine Translation. Although one might argue that Wilks (1976) proved that with the help of various semantic markers, a system developed based on the Case Grammar theory can capture the meaning of a sentence without the help of any syntactic analysis, and the ability to capture the meaning of sentences can facilitate the translation process; an MT system will still need to acquire some means to process the structure of sentences in both source and target language throughout the entire translation process, especially during the transfer and the sentence generation. Unlike the LFG formalism which defines the lexical mapping theory to map thematic information in an a-structure to the corresponding f-structure, Case Grammar does not have a well-defined means to link the syntactic and semantic information together. If Case Grammar is used in MT, an additional linguistic formalism is required to carry out syntactic analysis on the source and target sentences. However, there is still the lack of a practical way to link the resulting syntactic information with the corresponding case role. With the use of different structures to represent different levels of linguistic information and the use of structural correspondences to relate these structures, the LFG formalism, however, provides a complete framework to handle both the syntactic and semantic processing of sentences. Thus, the LFG formalism is more suitable for MT.

4.6 Conclusion

A-structure has two facets. In semantic terms, as thematic roles describe the different means of participating in an event, they show some semantic information about the characteristics of each participant of the event. For instance, the *agent* of an event is an animate object as it is the one responsible for initiating the event (Givón 1984). In syntactic terms, each a-structure is linked with the syntactic structure by mapping each thematic role to the corresponding syntactic function within a sentence. It is due to this dual function that a-structure can act as a link between lexical semantics and syntactic structures (Bresnan 1995). As we shall see in Chapter 6, although the representation method of traditional f-structure provides a suitable medium for carrying out transfer, the linguistic information captured in a traditional f-structure is relatively language-dependent and thus it is insufficient for transferring some kinds of sentences from one language to another, e.g. to transfer English passive sentences to Chinese. With the introduction of some semantic information, f-structure can provide more detailed information for improving the transfer. When compared with traditional f-structures, a-structure is relatively more language-independent and thus it provides a better medium for carrying out the transfer between sentences in different languages.

This chapter reviewed the theory of a-structure and the lexical mapping theory. The thematic information captured in a-structures is presented in terms of thematic roles. This chapter discussed the set of thematic roles and their meanings used in this study. Carlson (1984) suggested that verbs assigning different thematic roles should be considered as having different meanings. One of the well-known problems in Machine Translation is that a verb can have multiple meanings, depending on its use. The observation that a verb can possess different a-structures (cf. (55)) supports Carlson's suggestion. Owing to this observation, as it will be reported in Chapter 5, this study experiments with the effectiveness of using a-structure to improve the lexical selection process in MT over the use of the traditional lexical forms of verbs (cf. Section 3).

Owing to the similarities between a-structures and the theory of case frames in Case Grammar, a-structures are sometimes being mistaken as a variant of Case Grammar. However, as it was discussed in the previous section, although, to a certain extent, a-structures resemble case frames, this new extension of the LFG formalism (i.e. a-structure and the lexical mapping theory) differs from Case Grammar in many ways and thus cannot be viewed as a mere variant.

Chapter 5

Using A-structure and Lexical Mapping Theory for MT

The machine translation (MT) model adopted in this investigation employs a modular approach. The translation process is divided into three subtasks: source language parsing, source-to-target language transfer and target language generation. This chapter reports the findings on applying a-structure and lexical mapping theory to alleviate some of the problems in these subtasks.

5.1 Parsing Source Language Sentence

Throughout the history of NLP, researchers have been looking for methods to parse potentially ambiguous sentences correctly. In a text-based NLP system, source sentences are read in as a stream of words ordered according to their sequence of utterance. This order of words contains no explicit information about the structure of each sentence. A sequence of words often can be grouped in different manners which, as a result, convey different meanings. In an MT system, it is important to capture the original meaning conveyed by the author so as to produce an appropriate target language translation. It is therefore vital for the parser in an MT system to parse the sequence of words appropriately. Ideally, any potential alternative choice to parse a sentence which does not conform with the original meaning of the sentence should be pruned. However, there are sentences which are too ambiguous to be disambiguated by analysing their linguistic information alone. For instance, consider a highly ambiguous sentence like "*The man saw the girl in the park with a telescope.*": unless the parser is provided with some real world knowledge about the event (e.g. a picture of the scene which shows whether it was the girl who was carrying the telescope, or it was the man who used the telescope to view the girl) it is unlikely that the parser can produce a sentence structure which convey the original meaning. To find out how to parse this kind of sentences exceeds the scope of this study. After excluding

this kind of highly ambiguous sentences, there are sentences which are potentially ambiguous for a computer to parse, but they can be disambiguated through analysing the linguistic information carried by its constituents. For instance, consider the following English sentence:

(84) John played on words.

The sentence (84) described an event in which John was exploiting the meanings of words. This sentence normally would not be considered as ambiguous by average English speakers. However, when this sequence of English words is parsed by a pure syntax-based computer system, more than one combination of these words can be produced (cf. Figure 5.1): either

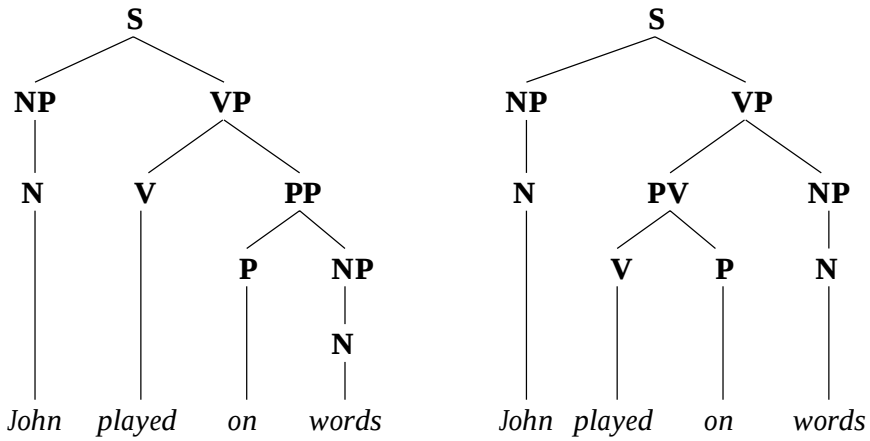


Figure 5.1: Two potential c-structures for the word sequence "*John played on words*"

the words 'played' and 'on' are combined to form one lexical unit, i.e. a phrasal verb (PV), or the words 'on' and 'words' are combined to form a prepositional phrase (PP). As we will discuss later in this chapter, each of these combinations of this sequence of words would result in different translations of these words.

To process natural language sentences is a tricky task. As we have seen from the above example, a sentence which is not considered as potentially ambiguous by human beings can become ambiguous when it is processed by average NLP systems. This is because programming a computer system to understand a sentence is a difficult task. Amongst the various kinds of information encoded in a sentence, it is relatively easy to program a computer system to process the syntactic information of the sentence. However, as we have discussed in the previous chapters, the use of syntax-oriented information alone is insufficient to ensure a computer system to parse natural language sentences appropriately (cf. Section 3.2.3). A means to capture and process the meaning of these sentences is required so as to increase the ability of a computer system to produce a translation to a source language sentence appropriately.

As illustrated in Chapter 4, thematic information of a sentence represented in an a-structure acts as a link between lexical semantic and syntactic structures. This information not only

enables a computer system to know what kind of participants are involved in an event, it also allows the computer system to have some knowledge about the meaning of words within a sentence. Therefore, a-structure can be used to alleviate the problem of ambiguity in NLP which is caused by insufficient linguistic information in the system. However, using thematic information alone is insufficient to aid NLP effectively. As discussed in Chapter 3, other kinds of linguistic information (e.g. syntactic and functional information of sentences) also play an important role in helping an NLP system to understand the sentences. Therefore, in order to facilitate the analysis and processing of sentences, different kinds of linguistic information are required to work co-operatively. Thematic information within an a-structure, as exemplified in Section 4.3, can be mapped with the corresponding syntactic functions in a sentence according to the lexical mapping theory. This allows thematic information of a sentence to work co-operatively with other linguistic information of the sentence to perform NLP.

This section discusses the ability of a-structure to alleviate some of the problems in structural disambiguation during sentence parsing. It also illustrates how lexical mapping theory aids a-structure to perform this task.

5.1.1 Differentiating V + PP from Phrasal Verb + NP

Knowing the syntactic category of a word, as exemplified in Section 3.1.1, is sufficient to determine the meaning of a homograph in some cases. Therefore, it is important that a parser can produce the appropriate structure of an input sentence. As mentioned earlier, the raw input of a parser is a sequence of words. Though this sequence of words often contains punctuation marks in between for signifying how these words are broken up into smaller chunks and where a sentence ends, this is not sufficient to determine how words are grouped to form phrases and what syntactic role (in terms of syntactic category, i.e. noun, verb, preposition, etc.) a word is functioning as. In order to work out the structure of a sentence from an input sequence of words, many conventional parsers carry out analysis on the sequence of words based on the information about the various morphological forms of words and phrase structure rules. This analysis is often carried out on a trial-and-error basis. When a sequence of words is potentially ambiguous (i.e. more than one sequence of phrase structure rules generates this sequence of words) the use of pure syntactic information becomes insufficient to resolve this kind of structural ambiguity. One example examined is the different combinations of verbs and prepositions within sentences.

The Problem

A prepositional phrase is made up of a preposition and an NP and it can appear after a verb or a noun. Prepositional phrases are often used to express additional information about when, how much and where an event or a situation occurs. They are also often used to express other

necessary information about some participants in the event, e.g. beneficiary (as in "John cooked a meal *for Mary*."), recipient (as in "John gave a book *to Mary*.") and instrument (as in "John killed the snake *with a stick*."). However, in an English sentence, not every NP that is preceded by a preposition functions as part of a prepositional phrase. In some sentences, a preposition is grouped with the preceding verb to form a new meaning. This kind of verb and preposition combination is called *phrasal verb*.

According to Collins COBUILD English Grammar (Sinclair 1990):

> *"Phrasal verbs are a special group of verbs which are made up of a verb and an adverb and/or a preposition which are used to extend or change the meaning of a verb."*

The meaning of phrasal verbs is typically not derived from the meaning of its two parts. If the translation is done on a direct word-to-word basis (cf. Section 1.1.2), the original meaning of the source sentence will be lost. For instance, the phrasal verb '*look for*' means *to seek*. The more commonly used meaning of the verb 'look' is *try to see*[1] and the preposition 'for' is normally used as an indication for destination, direction or purpose. When the words 'look' and 'for' are translated independently, the phrasal verb 'look for' will have a meaning that is completely different from its original meaning 'to seek'. For instance, with the sentence "John looked for Mary.", its word-for-word literal translation would mean "*John tried to see because of Mary*", which is different from its original meaning "*John sought Mary*".

Though phrasal verbs possess a different meaning from the literal meaning of their components, the literal meaning of the components of some phrasal verbs can be used to form a meaningful sentence. For example, the phrasal verb '*play on*' means *to exploit*, however, if the words 'play' and 'on' are treated as a normal pair of verb+preposition, literally it means 'amuse oneself', with the preposition 'on' and its following NP indicating where this play event happens. While parsing a sentence with the words 'play' and 'on', a parser will have to decide if these words should be treated as a phrasal verb or a normal occurrence of a verb and a preposition. Looking at the syntactic categories of the words 'play' and 'on' is insufficient to perform an appropriate judgment because in these two cases, they are considered as a verb and a preposition respectively. More examples of this kinds of verbs and prepositions are shown in Table 5.1[2]. Though it is believed that considering the context of the sentence helps in solving this problem, to represent word meaning for cross-referencing and verification process is a difficult task.

A Solution

Carlson (1984) suggested:

[1]Though in a different context, the verb 'look' can mean 'appear to be' or 'appear to other people as', as in "John looked well.". This meaning is not of interest to this example.

[2]The data are obtained from Sinclair (1987) and Sinclair (1989)

Verb + Preposition	**Examples**
come by	• John **came** *by car*.
	• John **came by** a fortune.
drink in	• John **drank** *in the school*.
	• John **drank in** every word that Mary said.
fish for	• John **fished** *for Mary*.
	• John **fished for** invitations.
go by	• John **went** *by bus*.
	• John **went by** the lake.
jump on	• John **jumped** *on the sofa*.
	• John **jumped on** Mary (for lying).
live by	• John **lived** *by a lamp-post*.
	• John **lived by** the rules.
play on	• John **played** *on the table*.
	• John **played on** words.
stand by	• John **stood** *by the lamp-post*.
	• John **stood by** Mary.
wait on	• John **waited** *on a table*.
	• John **waited on** a big event.

Table 5.1: Some examples of different combinations of verbs and prepositions

> *"... verbs assigning different thematic roles should be considered as meaning somewhat different things."*

This means that a verb which subcategorises more than one set of participants should have more than one meaning. Inspired by this suggestion, the a-structures of different combinations of verbs and prepositions were studied. It was found that different combinations of verbs and prepositions have different a-structures. The difference in these a-structures helps to disambiguate phrasal verbs from verb-and-prepositional-phrase pairs. Consider the following sentences and their corresponding f-structures shown in Figures 5.2 and 5.3:

(85) John *played on* words.

$$\begin{bmatrix} \text{PRED} & \text{'PLAY ON}\langle(\uparrow \text{SUBJ})\ \ (\uparrow \text{OBJ})\rangle\text{'} \\ \text{TENSE} & \text{PAST} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'JOHN'} \\ \text{SPEC} & - \\ \text{NUMB} & \text{SG} \\ \text{PERSON} & 3\text{RD} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'WORD'} \\ \text{SPEC} & - \\ \text{NUMB} & \text{PL} \\ \text{PERSON} & 3\text{RD} \end{bmatrix} \end{bmatrix}$$

Figure 5.2: F-structure for "John played on words."

(86) John *played* on the table.

$$\begin{bmatrix} \text{PRED} & \text{'PLAY}\langle(\uparrow \text{SUBJ})\rangle\text{'}(\uparrow \text{OBL}_\theta) \\ \text{TENSE} & \text{PAST} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'JOHN'} \\ \text{SPEC} & - \\ \text{NUMB} & \text{SG} \\ \text{PERSON} & 3\text{RD} \end{bmatrix} \\ \text{OBL}_\theta & \begin{bmatrix} \text{PRED} & \text{'ON}\langle(\uparrow \text{OBJ})\rangle\text{'} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'TABLE'} \\ \text{SPEC} & \text{'THE'} \\ \text{NUMB} & \text{SG} \\ \text{PERSON} & 3\text{RD} \\ \text{PCASE} & \text{'ON'} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 5.3: F-structure for "John played on the table."

The phrasal verb '*play on*' in (85) subcategorises the syntactic functions 'SUBJ' and 'OBJ' whereas

the verb-and-prepositional-phrase pair in (86) subcategorises a SUBJ and an OBL_θ. The events described by the sentences (85) and (86) involve different kinds of participants and thus result in different a-structures for each case:

(87) 'play on'<agent theme>

(88) play<agent> on<locative>

The lexical mapping between a-structure arguments and the syntactic functions restricts which a-structure can be assigned to the appropriate f-structure of a given sentence. The *agent* and the *theme* in the a-structure (87) are an external argument and an internal argument of this a-structure. According to the intrinsic role classifications (cf. Section 4.3.3), these two arguments are classified as $[-o]$ and $[-r]$ respectively. The default role classification (cf. Section 4.3.3) assigns an additional feature $[-r]$ to the *highest thematic role* (i.e. agent) of (87). Thus the thematic roles in (87) bear the features $\left[\begin{smallmatrix}-r\\-o\end{smallmatrix}\right]$ and $[-r]$ respectively. As shown in (89), they are mapped with the syntactic functions SUBJ (which bears the features $\left[\begin{smallmatrix}-r\\-o\end{smallmatrix}\right]$) and OBJ (which bears the features $\left[\begin{smallmatrix}-r\\+o\end{smallmatrix}\right]$) respectively, but not SUBJ and OBL_θ because the syntactic function 'OBL_θ' bears the feature $\left[\begin{smallmatrix}+r\\-o\end{smallmatrix}\right]$.

(89) Sentence : John played on words.

A-structure:	*'play on'*	<agent	theme>
Intrinsic:		$[-o]$	$[-r]$
Default:		$[-r]$	
Syntactic Functions:		SUBJ	OBJ
Noun Phrases:		John	words

Similarly, as shown in (90):

(90) Sentence : John played on the table.

A-structure:	*play*	<agent>	*on*	<locative>
Intrinsic:		$[-o]$		$[-o]$
Default:		$[-r]$		$[+r]$
Syntactic Functions:		SUBJ		OBL_{loc}
Noun Phrases:		John		the table

the *agent* and the *locative* in (88) which bear the features $\left[\begin{smallmatrix}-r\\-o\end{smallmatrix}\right]$ and $[-o]$ respectively (cf. Section 4.3.3) are assigned to the syntactic functions SUBJ and OBL_θ, but not SUBJ and OBJ. Since the thematic roles in (87) are mapped to the syntactic functions in Figure 5.2, and the thematic roles in (88) are mapped to the f-structure in Figure 5.3, but *not* vice versa, thus by looking at the correspondence between the a-structures and the f-structures of the sentences (85) and (86), a parser can identify the appropriate syntactic structure for the participants of an event described in an a-structure.

Note that the appropriate lexical mapping between a given f-structure and the corresponding a-structure is not always sufficient to resolve the structural ambiguity of sentences like (85) and (86). Given the sequence of words (85), an f-structure similar to that in Figure 5.2 or Figure 5.3 can be generated by a syntax-oriented parser. If an inappropriate f-structure is generated for describing the syntactic structure of the sentence, the above lexical mapping process will assign a wrong set of thematic role(s) to the input sentence and allow the inappropriate syntactic structure to be assigned to the sentence.

A-structure and lexical mapping theory will help the selection of an appropriate syntactic structure for a sentence effectively *only* when the semantic properties of lexical items and thematic roles are known to the parser. Recall that, as mentioned in Section 4.1, each thematic role bears some semantic properties. These semantic properties can be represented in a parser simply in the form of semantic markers like $[\pm animate]$ and $[\pm physical]$. Likewise, the semantic information of some words (e.g. nouns and adjectives) can be represented in a similar manner. For instance, a *table* is an inanimate physical object whereas *words* tend to be considered as something abstract. When this kind of semantic information is made available to a parser, during the lexical mapping, the mis-matching semantic properties between each pair of thematic role and lexical item will force the parser to reject the syntactic structure which has been inappropriately generated for a sentence.

The above disambiguation method is also applicable to resolve the potential structural ambiguity in other sentences, e.g. those shown in Table 5.1. The lexical mapping between the a-structure arguments of some of these sentences and their corresponding syntactic functions is shown in Figure 5.4.

A-structure shows the necessary participants of an event in terms of thematic roles. The mapping between these thematic roles and their corresponding syntactic functions is defined in lexical mapping theory. By carrying out the lexical mapping between a-structure arguments and the arguments within a syntactic structure (e.g. SUBJ and OBJ), inappropriate phrase structure groupings can be eliminated during the parsing. As a result, the chance of producing a target sentence with distorted meaning due to an inappropriate sentence structure produced by the parser would be reduced. This is made possible through matching the semantic properties of thematic roles and the corresponding lexical items during lexical mapping. Note that, without the aid of a-structure and the lexical mapping theory, using semantic markers alone cannot effectively aid the selection of an appropriate syntactic structure unless some semantic properties are hard-coded in each syntactic function within a semantic form of a verb. These semantic properties would then restrict the selection of appropriate lexical entries for describing the linguistic information about an input sentence during the parsing process. The use of a-structure and the lexical mapping theory allows a more natural and effective way to incorporate lexical semantic information in a parser to aid structural disambiguation.

Sentence :	**John came by a fortune.**	
A-structure: '*come by*'	<recipient	theme>
Intrinsic:	$[-o]$	$[-r]$
Default:	$[-r]$	
Syntactic Functions:	SUBJ	OBJ
Noun Phrases:	John	a fortune

Sentence :	**John came by car.**		
A-structure: *come*	<agent>	*by*<instrument>	
Intrinsic:	$[-o]$	$[-o]$	
Default:	$[-r]$	$[+r]$	
Syntactic Functions:	SUBJ	OBL_{instr}	
Noun Phrases:	John	car	

Sentence :	**John fished for invitations.**	
A-structure: '*fish for*'	<agent	theme>
Intrinsic:	$[-o]$	$[-r]$
Default:	$[-r]$	
Syntactic Functions:	SUBJ	OBJ
Noun Phrases:	John	invitations

Sentence :	**John fished for Mary.**		
A-structure: *fish*	<agent>	*for*<beneficiary>	
Intrinsic:	$[-o]$	$[-o]$	
Default:	$[-r]$	$[+r]$	
Syntactic Functions:	SUBJ	OBL_{ben}	
Noun Phrases:	John	Mary	

Sentence :	**John lived by the rules.**	
A-structure: '*live by*'	<agent	theme>
Intrinsic:	$[-o]$	$[-r]$
Default:	$[-r]$	
Syntactic Functions:	SUBJ	OBJ
Noun Phrases:	John	the rules

Sentence :	**John lived by a lamp-post.**		
A-structure: *live*	<agent>	*by*	<locative>
Intrinsic:	$[-o]$		$[-o]$
Default:	$[-r]$		$[+r]$
Syntactic Functions:	SUBJ		OBL_{loc}
Noun Phrases:	John		a lamp-post

Sentence :	**John waited on a big event.**	
A-structure: '*wait on*'	<agent	theme>
Intrinsic:	$[-o]$	$[-r]$
Default:	$[-r]$	
Syntactic Functions:	SUBJ	OBJ
Noun Phrases:	John	a big event

Sentence :	**John waited on a table.**		
A-structure: *wait*	<agent>	*on*	<locative>
Intrinsic:	$[-o]$		$[-o]$
Default:	$[-r]$		$[+r]$
Syntactic Functions:	SUBJ		OBL_{loc}
Noun Phrases:	John		a table

Figure 5.4: The lexical mapping between a-structure arguments and their corresponding syntactic functions for the sentences in Table 5.1

5.1.2 Differentiating NP with N and PP from NP + PP

Another structural confusion that often happens during parsing English sentences is whether to treat a sequence of Noun (N), Preposition (P) and Noun Phrase (NP) as one NP with a Prepositional Phrase (PP) modifying the N or to treat it as an NP + PP, i.e. with the PP modifying the verb of the sentence. For instance, consider the sentence:

(91) John bought a book in a bookshop in Prague.

Both PPs in the above sentence can be considered to be adjuncts of the sentence. Unlike the PPs which are used for expressing mandatory arguments of a predicator (e.g. the PP 'on the table' in "John put a book on the table."), the number of PPs that can appear in a sentence is unlimited

(e.g. "John gave a speech through the microphone with his arm bound in the hall at four o'clock on Monday in front of many people . . . "). If both PPs in (91) are considered as separate adjuncts of the sentence, the resulting c-structure would be like the one shown in Figure 5.5. If the PP

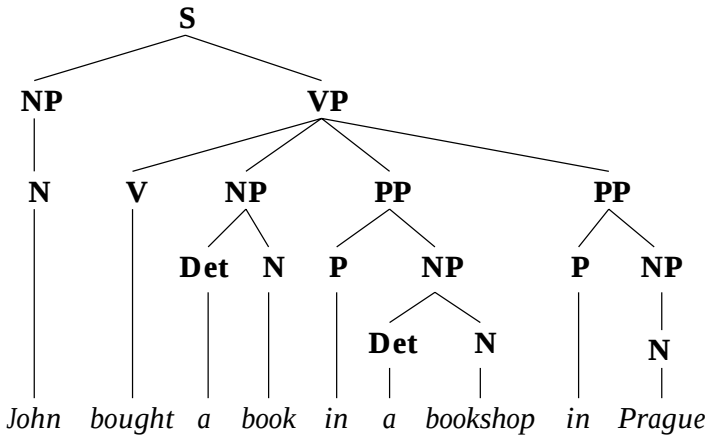


Figure 5.5: A possible c-structure for "*John bought a book in a bookshop in Prague.*" produced by a syntax-based parser.

'in Prague' is considered as the PP modifying the NP 'a bookshop', the resulting c-structure would be Figure 5.6. Although in terms of the actual meaning of the sentence, unlike the cases we considered in the previous section, either of the c-structures in Figures 5.5 and 5.6 would be understood as "John bought the book in Prague". When translating the sentence (91) to Chinese, using the c-structure in Figure 5.5 would produce a sentence which sounds odd to many native Chinese speakers, i.e.:

(92)

?	約翰	在	布拉格	在	一	間	書店	內	買	了	一	本	書。
	John	at	Prague	at	one	QUAN-TIFIER	book-shop	inside	buy	ASPECT MARKER	one	QUAN-TIFIER	book.

John, in Prague, in a bookshop, bought a book.

An appropriate Chinese translation to the sentence (91) should be:

(93)

約翰	在	布拉格的	一	間	書店	內	買	了	一	本	書。
John	at	Prague's	one	QUAN-TIFIER	book-shop	inside	buy	ASPECT MARKER	one	QUAN-TIFIER	book.

John, in a bookshop in Prague, bought a book.

In this sentence, unlike the sentence (92), instead of treating the PP 'in Prague' as a PP modifying the verb of the sentence, this PP is treated as a part of the NP 'a bookshop' for modifying this NP. This information is reflected by the c-structure shown in Figure 5.6.

When parsing the English sentence (91), if the parser generates the c-structure in Figure 5.5

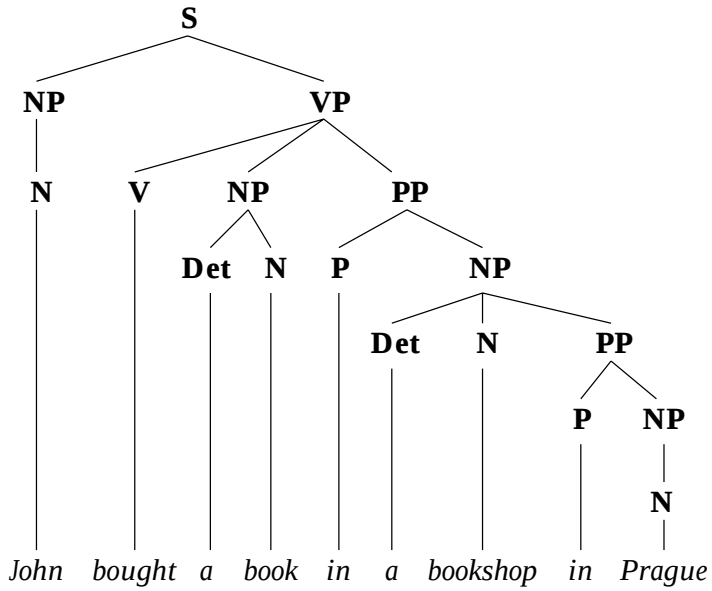


Figure 5.6: Another possible c-structure for "*John bought a book in a bookshop in Prague.*" produced by a parser.

instead of the one shown in Figure 5.6, the resulting Chinese translation would cause confusion to native Chinese speakers. From this translation example, we can see that the production of an appropriate c-structure for a source sentence helps to generate an appropriate translation for the sentence. However, as it was mentioned earlier, there can be unlimited number of PPs which act as adjuncts of a sentence to indicate time, manner, place, etc. of the event described, we cannot determine which PP is attached to which phrase (i.e. VP or NP) by simply counting the number of PP in a sentence. Using syntactic information alone, again, is insufficient to solve this problem. However, it is observed that one element in the lexical mapping theory offers a potential solution to this problem.

According to one of the well-formedness conditions of the lexical mapping theory — *functional biuniqueness*, each thematic role in an a-structure must be unique. This means that each thematic role can appear in an a-structure only once. This well-formedness condition helps a parser to differentiate an NP with N and PP from an NP + PP pairs. Take the sentence (91) "John bought a book in a bookshop in Prague." as an example. While parsing this sentence, if the parser considers the NPs 'in a bookshop' and 'in Prague' to be separate adjuncts for the verb 'buy', it would produce the following a-structure:

(94) buy<agent theme> in<locative> in<locative>

because the preposition 'in' subcategorises a locative argument when it is used in conjunction with the verb 'buy'. However, when this a-structure is checked against the well-formedness conditions defined in the lexical mapping theory, the functional biuniqueness would fail this

a-structure because the same thematic role, i.e. locative, appears twice in this a-structure. This error forces the parser to try another c-structure for this sentence, i.e. the c-structure in Figure 5.6. This c-structure treats the PP 'in Prague' as a part of the NP 'a bookshop'. With this c-structure, the parser would produce the a-structure:

(95) buy<agent theme> in<locative>

in which the locative argument would be mapped to the NP 'a bookshop in Prague'. The a-structure (95) does not have repetitive thematic roles, so it would not fail the check against the functional biuniqueness condition. The resulting c-structure and a-structure can then be used for aiding the remaining translation process to produce an appropriate Chinese translation (i.e. (93)).

Similarly, if the word sequence:

(96) "John saw a girl with a dog with a telescope."

is input to a parser, the PPs 'with a dog' and 'with a telescope' would not be considered as two separate adjuncts to the sentence because if both of these PPs are adjuncts to the sentence, the resulting a-structure would be:

(97) see<agent theme> with<instrument> with<instrument>

which is ill-formed. Thus at least one of these PPs (i.e. 'with a dog' or 'with a telescope') must be an NP modifier. However, the parser would still need to decide which PP is modifying which NP so as to work out an appropriate structure for the sentence (96). For a human parser, this is a simple task because there is no doubt that the first PP 'with a dog' is for modifying the NP 'a girl' and the second PP 'with a telescope' is describing the instrument that John used to view the girl. However, without knowing the meaning of the words in the sentence, a computer parser cannot determine the function of each PP successfully. Therefore, performing the well-formedness checks on a-structures of sentences does not always disambiguate the structure of sentences successfully.

One way to alleviate the above disambiguation problem is by looking at the semantic properties of the nouns 'dog' and 'telescope'. As mentioned in Chapter 4, each thematic role bears some semantic properties. An instrument tends to be an inanimate object (cf. Section 4.1.3). When the usual semantic property for the noun 'dog' (i.e. [+animate]) is checked against the semantic property of an instrument (i.e. [-animate], it would be clear that the PP 'with a dog' is not expressing the instrument of the event described by the sentence (96). Thus, the parser can then treat this PP as part of the NP 'a girl' and the remaining PP 'with a telescope' as an adjunct of the sentence, yielding the c-structure shown in Figure 5.7.

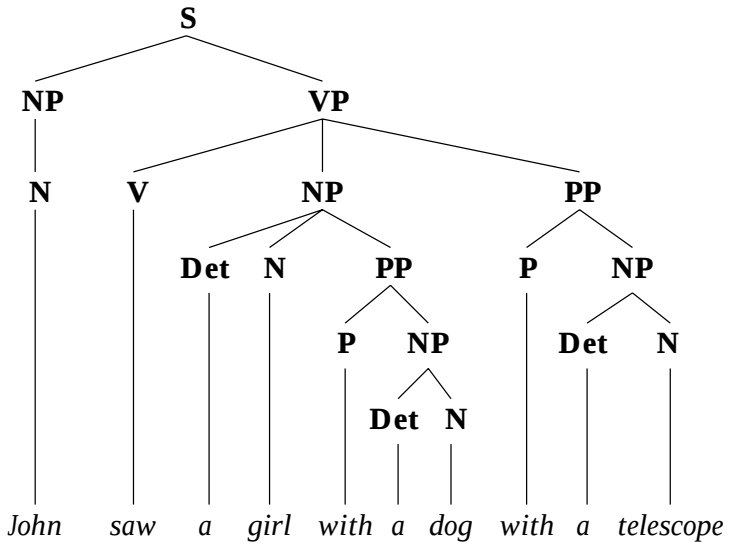


Figure 5.7: The c-structure for "*John saw a girl with a dog with a telescope.*"

5.2 Lexical Selection

As illustrated in Section 3.1.1, a word can have more than one meaning depending on its part of speech. This kind of words, i.e. words with same spelling but different meanings, are called homographs. There are three kinds of homographs:-

1. Different meanings occur *only* in different word categories, e.g. the word '*yank*' when used as a noun, it means a person coming from America; when it is used as a verb, it means pulling someone or something suddenly and with a lot of force.
2. Different meanings occur *only* in the same word category, e.g. the word '*ball*' when used as a noun, it can mean either an object for sports with shape normally like a sphere (e.g. foot*ball*) or a social gathering for dancing.
3. Different meanings occur in *both* different and same word category (i.e. a hybrid of the first two kinds), e.g. the word '*program*' means composing a set of instructions for computer system when it is used as a verb; when it is used as a noun, it can mean a set of computer instructions or a list of performance.

The ambiguity caused by the first kind of homographs can be resolved easily with the aid of syntactic analysis, i.e. to look at what syntactic role a word plays within a sentence. However, the other two, which involve representation of different meanings within the same word category, are more difficult to resolve. For instance, consider the following sentences:-

(98) John *lit* a cigarette.

(99) The fire *lit* the road.

The verb '*light*' in these sentences mean '*to ignite*' and '*to illuminate*' respectively. Though both meanings are related to the use of some sort of heat energy, they have different translations in Chinese (i.e. '點燃' and '照亮' respectively):

(100) 約翰 點燃 了 一 根 香煙。
John to ignite ASPECT MARKER a QUANTIFIER cigarette.
John lit a cigarette.

(101) (這 堆) 火 把 (這 條) 道路 照亮 了 。
this QUANTIFIER fire ≈ make this QUANTIFIER road illuminate ASPECT MARKER .
The fire made the road become bright.

Looking at the meaning of these two Chinese translations, i.e. 'to ignite' and 'to illuminate', we can see that these translations are *not* inter-changeable. If '照亮' is used in (100), a native Chinese speaker would perceive that the cigarette was illuminated instead of being ignited. Likewise, if '點燃' is used in (101), the reader would understand the road to be ignited instead of being illuminated. While translating the sentences (98) and (99), if a wrong translation is chosen by the MT system, not only can the meaning of the original sentence not be preserved, it might even cause misunderstanding and worry to the reader. Therefore, it is important that an MT system is able to choose the most appropriate translation for the source language words. This selection process is called *Lexical Selection*.

As discussed in Section 3.2.3, some researchers attempted to disambiguate the meaning of verbs by using the semantic forms of verbs. However, this method, as exemplified in Section 3.2.3, is inadequate to aid the lexical selection process. For instance, both of the verb 'light' in (98) and (99) have the same semantic form 'LIGHT<(↑SUBJ) (↑OBJ)>'. Knowing the syntactic category and the semantic form of the word '*lit*' in these sentences is insufficient to distinguish its meaning in each case. There is the need to capture a higher level of linguistic information of (98) and (99) so as to resolve this ambiguity.

In order to understand the meaning of each lexical unit in a sentence, the most thorough way perhaps is to use a semantic network to capture the meaning of words (e.g. Palmer & Wu 1995). A semantic network defines the detailed semantic properties of each word and how meanings of words are related to each other through their semantic properties. The use of a semantic network to aid sentence analysis allows an MT system to have a more thorough understanding on the meaning of input sentences. This, in turn, improves the lexical disambiguation process. However, to construct a semantic network is not a trivial task. One vital problem is that it is difficult to define a set of semantic properties which is adequate to capture the meaning of each word in the lexicon. In addition, even for the construction of a semantic network for a very small toy lexicon, a lot of time and effort is required. Furthermore, the detailed semantic properties of words often do not form part of the linguistic information captured in contemporary

linguistic formalisms. As was discussed in Chapter 3, different levels of linguistic information help an MT system to analyse source language sentences and to generate target language sentences. This means that some kind of additional mechanism is required for an MT system to relate the various kinds of linguistic information (e.g. syntactic and functional information) of sentences and the semantic information of words offered by a semantic network. These problems seem to out-weigh the benefits that a semantic network offers to an MT system.

Ideally, an effective method for lexical selection should be sufficient to disambiguate the meanings of words and at the same time be relatively easy to implement and be able to be used in conjunction with other kinds of linguistic information. When considering the characteristics of a-structure and lexical mapping theory, it is found that a-structure and lexical mapping theory is a good candidate to be chosen as an effective method for lexical selection. The following of this section discusses the ability of a-structure and lexical mapping theory in aiding the lexical selection process.

5.2.1 Lexical Selection for Ergative Verbs

Ergative verbs are verbs which can be used:

> "*to describe an action from the point of view of the performer of the action or from the point of view of something which is affected by the action, i.e. the same verb can be used transitively, followed by the object, or intransitively, without the original performer being mentioned.*"
>
> [Sinclair (1990, Section 3.60, page 155)]

In English, ergative verbs are fairly common. Some ergative verbs in English have an exact matching counterpart in Chinese, i.e. their Chinese translations can also be used both transitively and intransitively. Two examples of this kind of ergative verbs are the verbs 'sound' and 'open' (cf. Figure 5.8). To translate this kind of English ergative verbs is relatively easy because despite whether they are used in their transitive form or intransitive form, their translations to Chinese would be the same.

When compared with English, ergative verbs are not as common in Chinese. There are a larger collection of ergative verbs in English which do not have an ergative counterpart in Chinese (cf. Figure 5.9). When translating this kind of English ergative verbs to Chinese, there is the need to differentiate whether they are used in their transitive form or intransitive form so as to produce an appropriate translation for each case. Therefore, the question is: *how to program an MT system to decide which translation to be used during processing time?*

Consider the difference in the syntactic structure between the sample sentences in Figure 5.9: the transitive examples subcategorise a SUBJ and an OBJ whereas the intransitive examples subcategorise a SUBJ only. Therefore, it can be concluded that one possible way to perform lexical

1. 'sound' *versus* '響起'

Transitive 'sound':	John *sounded* the alarm.					
Transitive '響起':	約翰	**響起**	了	(這	個)	警報。
	John	sound	ASPECT MARKER	(this	QUANTIFIER)	alarm.
Intransitive 'sound':	The alarm *sounded*.					
Intransitive '響起':	(這	個)	警報	**響起**	了。	
	(this	QUANTIFIER)	alarm	sound	ASPECT MARKER.	

2. open *versus* '開'

Transitive 'open':	John *opened* the door.					
Transitive '開':	約翰	開	了	這	扇	門。
	John	open	ASPECT MARKER	this	QUANTIFIER	door.
Intransitive 'open':	The door *opened*.					
Intransitive '開':	(這	扇)	門	開	了。	
	(this	QUANTIFIER)	door	open	ASPECT MARKER.	

Figure 5.8: Examples of English ergative verbs with matching Chinese counterpart

selection on this kind of verbs is, as the method suggested by Her et al. (1994), to consider their semantic form: if an English ergative verb bears the semantic form '<(↑SUBJ) (↑OBJ)>', the corresponding transitive Chinese translation is selected; if the English ergative verb bears the semantic form '<(↑SUBJ)>', the corresponding intransitive Chinese translation is used. However, this method does not exercise any checking on the semantic properties of the arguments in the semantic form. If an input sentence is incomplete or ill-formed (e.g. the sentence "John peeled off."), an MT system would not be able to track down this error until the MT system tried to generate the target sentence by looking at the semantic form of the selected Chinese translation.

Apart from semantic forms of verbs, a-structures also capture the difference between the transitive and intransitive sentences in Figure 5.9. The transitive version of the ergative verbs in Figure 5.9 have an a-structure which takes two arguments: <agent theme>; whereas the intransitive version of these verbs have an a-structure which takes only one argument: <theme>. Therefore, an MT system can also perform lexical selection for the kind of verbs in Figure 5.9 based on the information given by the a-structures of sentences. Given an English ergative verb which can be translated into two different Chinese verbs, e.g. 'dry'. If the a-structure of the sentence is 'DRY<agent theme>', the Chinese verb '弄乾' would be selected during the lexical selection process; otherwise (i.e. if the sentence has the a-structure 'DRY<theme>') the Chinese verb '乾' would be selected. This method is better than the method mentioned previously because it can track down the input sentences which are incomplete or ill-formed. Recall that Section 4.1 illustrated that each thematic role carries certain semantic properties.

1. 'stop' *versus* '弄停' and '停'

Transitive 'stop': John *stopped* the car.
'弄停': 約翰 弄停 了 這 輛 汽車。
John make-stop ASPECT MARKER this QUANTIFIER car.
Intransitive 'stop': The car *stopped.*
'停': 這 輛 汽車 停 了。
this QUANTIFIER car stop ASPECT MARKER.

2. 'break' *versus* '打破' and '破

Transitive 'break': John *broke* the window.
'打破': 約翰 打破 了 (這 隻) 窗子。
John hit-break ASPECT MARKER (this QUANTIFIER) window.
Intransitive 'break': The window *broke.*
'破': (這 隻) 窗子 破 了。
(this QUANTIFIER) window break ASPECT MARKER.

3. 'sink' *versus* '弄沉' and '沉'

Transitive 'sink': John *sank* the boat.
'弄沉': 約翰 弄沉 了 (這 艘) 小船。
John make-sink ASPECT MARKER (this QUANTIFIER) boat.
Intransitive 'sink': The boat *sank.*
'沉': 這 艘 小船 沉 了。
this QUANTIFIER boat sink ASPECT MARKER.

4. 'dry' *versus* '弄乾' and '乾'

Transitive 'dry': John *dried* the dishes.
'弄乾' 約翰 弄乾 了 (這些) 盤子。
John make-dry ASPECT MARKER (these) dish.
Intransitive 'dry': The dishes *dried.*
'乾': 這些 盤子 乾 了。
these dish dry ASPECT MARKER.

5. 'peel off' *versus* '剝' and '脫落'

Transitive 'peel off': John *peeled off* the wrapper.
'剝': 約翰 剝 了 這 張 包裝紙。
John peel-off ASPECT MARKER this QUANTIFIER wrapper.
Intransitive 'peel off': His skin *peeled off*.
'脫落': 他的 皮 脫落 了。
his skin strip down ASPECT MARKER.

6. 'quieten down' *versus* '使…平靜下來' and '平靜下來'

Transitive 'quieten down': John *quietened* Mary *down*.
'使…平靜下來': 約翰 使 瑪莉 平靜 下來 了
John make Mary quieten down ASPECT MARKER
Intransitive 'quieten down': Mary *quietened down*.
'平靜下來': 瑪莉 平靜 下來 了。
Mary quieten down ASPECT MARKER.

Figure 5.9: Examples of English ergative verbs with different Chinese translation in transitive and intransitive cases

While processing the sentence "John dried.", after looking at the semantic property of 'John' (i.e. [+animate]) and the semantic property of a theme (i.e. [-animate]), an MT system can rule out the possibility to use the Chinese translation '乾' because the semantic properties of 'John' and theme contradict each other.

This method also shows the relationship between the participants of transitive and intransitive sentences. For instance, consider the sentence pairs "John stopped a car. The car stopped.". The car mentioned in the second sentence is the one that John stopped in the first sentence. This relationship is not shown in the semantic forms of the verb 'stop' in these sentence (i.e. 'STOP <(↑SUBJ) (↑OBJ)>' and 'STOP<(↑SUBJ)>') as the SUBJ of the second sentence (i.e. "The car stopped.") is *not* the SUBJ of the first sentence (i.e. "John stopped the car."). The a-structures of these sentences are 'STOP<agent theme>' and 'STOP<theme>' respectively. From these a-structures, we can clearly see that the second argument of the sentence "John stopped a car." is the first argument of the sentence "The car stopped." because they both play the same role in the event structures described by these sentences.

5.2.2 Lexical Selection for Verbs

Most English verbs, when used in different situations (e.g. when used in conjunction with different nouns and / or prepositional phrases), possess different meanings. Though some of these meaning differences are insignificant in one language (e.g. in English), when these verbs are translated to Chinese, these minute differences can affect the readability of the output translation or even cause confusion and misunderstanding to the reader[3]. The use of semantic forms of verbs for lexical selection, as discussed in Section 3.2.3 and at the beginning of Section 5.2, is syntax-oriented and thus it is insufficient to capture these relatively insignificant meaning differences. A more language-independent structure which is capable of capturing the meanings of verbs is required to alleviate this problem. Recall that Carlson (1984) pointed out that:

> "*... verbs assigning different thematic roles should be considered as meaning somewhat different things.*"

Thematic information of sentences is represented in a-structures in LFG (cf. Chapter 4). If the difference in thematic roles characterised by a verb helps to distinguish the meaning difference of a verb, a-structure would be capable of aiding the selection of appropriate translations for verbs in an MT system.

Consider the a-structures and sample sentences for the English verb 'tell' and its Chinese counterparts '說' and '告訴' shown in Figure 5.10. The meaning difference between the Chinese verbs '說' (i.e. *to deliver information*) and '告訴' (i.e. *to give information to someone*) is distinguished by the absence and the presence of the thematic role *recipient* respectively. The inadequacy of

[3] cf. the Chinese translations of (98) and (99) in Section 5.2

1. **'tell<agent recipient theme>'** *versus* **'告訴<agent theme recipient>'**

English sentence :	John told Mary a matter.						
Chinese translation :	約翰	告訴	了	瑪莉	一	件	事情。
	John	tell	ASPECT MARKER	Mary	one	QUAN-TIFIER	matter.

2. **'tell<agent theme> to<recipient>'** *versus* **'說<agent theme> 給<recipient>'**

English sentence :	John told a story to Mary.							
Chinese translation :	約翰	給	瑪莉	說	了	一	個	故事。
	John	to	Mary	say	ASPECT MARKER	one	QUAN-TIFIER	story.

3. **'tell<agent theme>'** *versus* **'說<agent theme>'**

English sentence :	John told a lie.					
Chinese translation :	約翰	說	了	一	個	謊話。
	John	say	ASPECT MARKER	one	QUANTIFIER	lie.

4. **'tell<agent recipient>'** *versus* **'告訴<agent recipient>'**

English sentence :	I told you!			
Chinese translation :	我	告訴	了	你!
	I	tell	ASPECT MARKER	you!

Figure 5.10: A-structures and sample sentences for the English verb 'tell' and its Chinese counterparts

the method for lexical selection suggested by Her et al. (1994), as discussed in Section 3.2.3, can therefore be overcome by using the thematic information encoded in a-structures to perform lexical selection.

Consider the semantic forms of the verb 'told' in the sentences:

1. *John told a story.*
2. *I told you.*

Though both of them govern the same pair of syntactic functions: <SUBJ OBJ>, the corresponding thematic roles assigned to OBJ in each case are different: *theme* and *recipient* respectively. Recall that syntactic functions are assigned with the features $[\pm r]$ and $[\pm o]$:

$$\text{SUBJ} \begin{bmatrix} -r \\ -o \end{bmatrix} \qquad \text{OBJ} \begin{bmatrix} -r \\ +o \end{bmatrix} \qquad \text{OBL}_\theta \begin{bmatrix} +r \\ -o \end{bmatrix} \qquad \text{OBJ}_\theta \begin{bmatrix} +r \\ +o \end{bmatrix}$$

The real factor governing the different usages of the verb 'tell' in the above sentences appears to be what thematic role the syntactic function 'OBJ' is assigned with:

(102) Sentence : John told a lie.

A-structure :	*tell* <	agent	theme	>
Intrinsic :		$[-o]$	$[-r]$	
Default :		$[-r]$		
Semantic form :	TELL <	SUBJ	OBJ	>
NPs :		John	a lie	

(103) Sentence : I told you.

A-structure :	*tell* <	agent	recipient	>
Intrinsic :		$[-o]$	$[-r]$	
Default :		$[-r]$		
Semantic form :	TELL <	SUBJ	OBJ	>
NPs :		I	you	

As a-structure captures thematic information of sentences, it can be used as a replacement for the semantic forms of verbs for improving lexical selection. As shown in Figure 5.11, the appropriate source-to-target language equivalent can be obtained by matching the a-structure arguments of the source and target language verbs.

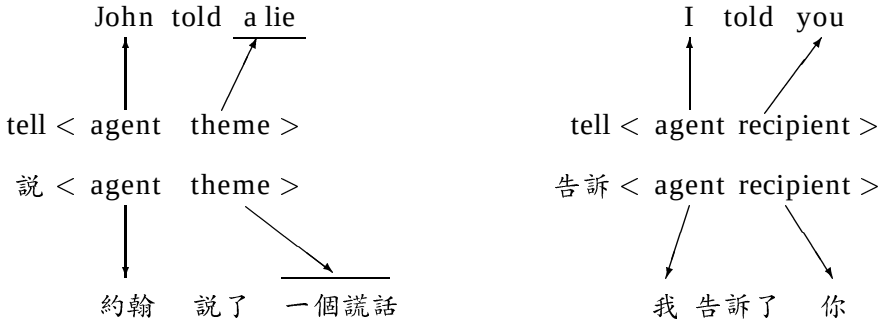


Figure 5.11: The use of a-structures for lexical selection

Though the thematic hierarchy (61) shown in Section 4.3.1 was established with the aim of generalising the universal order of thematic roles across languages, as illustrated in Section 4.3.1, the hierarchical order of thematic roles in Chinese differs from this hierarchy. As the order of thematic roles in an a-structure corresponds to a thematic hierarchy, different thematic hierarchies result in a slight difference in a-structures between different languages. For instance, as shown in Figure 5.10, the a-structure of the ditransitive Chinese verb '告訴' is:

(104) 告訴<agent *theme recipient*>

whereas the a-structure of its English counterpart 'tell' is:

(105) tell<agent *recipient theme*>

These two a-structures differ in their order of the thematic roles *recipient* and *theme*. Due to this difference, when performing lexical selection based on matching source and target language a-structures, the matching process is based only on the types and the number of thematic roles governed by the source and target a-structures; the orders of thematic roles within these a-structures are not taken into account.

Consider the lexical mapping in (102) and (103): both of the thematic roles in the a-structures 'tell<agent theme>' and 'tell<agent recipient>' have the same set of intrinsic and default features, i.e. $[-r]$ for the theme and the recipient, and $\begin{bmatrix}-o\\-r\end{bmatrix}$ for both agents. One might argue that looking at these features alone would not be able to differentiate which thematic role should be mapped to which NP. That is to say, the a-structure 'tell<agent recipient>' can be assigned to the sentence in (102); whereas the a-structure 'tell<agent theme>' can be assigned to the sentence in (103).

In some cases, looking at the $[\pm r]$ and $[\pm o]$ features alone in performing the lexical mapping can be insufficient to map the appropriate thematic roles to the syntactic functions of a sentence. However, as exemplified in Section 4.1, each thematic role carries some semantic properties. These semantic properties help to restrict which NP can be assigned with which thematic role. For instance, the NP 'a lie' in (102) is an inanimate, abstract object (i.e. bearing the semantic features [-physical] and [-animate]), but the thematic role 'recipient' tends to refer to an animate and physical object (i.e. bearing the semantic features [+physical] and [+animate]). When these semantic properties are checked against each other through simple matching, it is clear that the NP 'a lie' *cannot* be mapped to the thematic role 'recipient'. The semantic properties of the NP 'a lie' do not contradict with those of the thematic role 'theme'. Thus, the a-structure of the sentence "John told a lie." is 'tell<agent theme>', but *not* 'tell<agent recipient>'.

Similarly, the difficulty in selecting the appropriate translation for the English verb 'light' in (98) and (99) can be overcome by matching source and target language a-structures:

- **light<agent theme>** $\Longleftrightarrow$ **點燃<agent theme>**

John lit a cigarette.	$\Longleftrightarrow$	約翰	點燃	了	一	根	香煙
		John	to start burning	ASPECT MARKER	a	QUANTIFIER	cigarette

- **light<instrument theme>** $\Longleftrightarrow$ **照亮 <instrument theme>**

The fire lit the road.	$\Longleftrightarrow$	火	把	道路	照亮	了
		fire	make	road	illuminate	ASPECT MARKER

More examples of lexical selection for verbs by using a-structures are shown in Figure 5.12.

- **cook**

1. **cook**<**agent**> ⟺ **下廚**<**agent**>

   John cooked yesterday. ⟺ 約翰 昨天 下廚 了
   John yesterday be in kitchen ASPECT MARKER

2. **cook**<**theme**> ⟺ **烹煮**<**theme**>

   The apples cook well. ⟺ 這些 蘋果 適於 烹煮
   these apple suitable cook

3. **cook**<**agent theme**> ⟺ **烹煮**<**agent theme**>

   John cooked a chicken. ⟺ 約翰 烹煮 了 一 隻 雞
   John cook ASPECT MARKER one QUANTIFIER chicken

4. **cook**<**agent theme**> **for**<**beneficiary**> ⟺ **烹煮**<**agent theme**> **爲**<**beneficiary**>

   John cooked a chicken ⟺ 約翰 爲 瑪莉 烹煮 了 一 隻 雞
   for Mary. John for Mary cook ASPECT MARKER one QUANTIFIER chicken

- **escape**

1. **escape**<**agent**> ⟺ **逃走**<**agent**>

   John escaped from the prison. ⟺ 約翰 從 監獄 中 逃走 了
   John from prison middle escape ASPECT MARKER

2. **escape**<**theme**> ⟺ **漏出**<**theme**>

   Water escaped from the container. ⟺ 水 從 容器 中 漏出 了
   water from container middle leak out ASPECT MARKER

- **leave**

1. **leave**<**agent**> ⟺ **離開** <**agent**>

   John left yesterday. ⟺ 約翰 昨天 離開 了
   John yesterday leave ASPECT MARKER

2. **'leave for'**<**agent locative**> ⟺ **到…去**<**agent locative**>

   John left for Rome yesterday. ⟺ 約翰 昨天 到 羅馬 去 了
   John yesterday to Rome go ASPECT MARKER

3. **leave**<**agent theme**> ⟺ **留下**<**agent theme**>

   John left a pen ⟺ 約翰 在 桌子 上 留下 了 一 支 筆
   on the table. John at table top leave ASPECT MARKER one QUANTIFIER pen

- **lie**

1. **lie**<**agent**> ⟺ **說謊**<**agent**>

   John lies to Mary. ⟺ 約翰 對 瑪莉 說謊
   John to Mary lie

2. **lie**<**agent 'on**<**locative**>'> ⟺ **躺** <**agent '在…上**<**locative**>'>

   John lies on the sofa. ⟺ 約翰 躺 在 沙發 上
   John lie at sofa top

Figure 5.12: Some examples on lexical selection for verbs by using a-structures

5.2.3 Lexical Selection for Phrasal Verbs

In Section 5.1.1, we looked at why it is important to identify the presence of phrasal verbs in sentences and how to differentiate a phrasal verb from a verb-and-prepositional-phrase pair. However, even if a phrasal verb can be identified by a parser adequately, it does not mean that an appropriate translation for the phrasal verb can be produced easily. Despite the literal meaning of phrasal verbs, there are many phrasal verbs which have more than one modified meaning. Therefore, after differentiating a phrasal verb (i.e. a verb-and-preposition pair with modified meaning) from the normal utterance of a verb followed by a preposition, there is the need for an MT system to select the most appropriate translation for phrasal verbs. Owing to the fact that a phrasal verb can be made up of a verb and an adverb and/or a preposition (Sinclair 1990), apart from the combination of a verb and a preposition, there is more room for ambiguity to occur. For instance, the words 'look' and 'up' can occur in a sentence in three different manners:

1. John *looked up*.
2. John *looked up* a word from the dictionary.
3. John *looked up* to Mary.

with the first 'looked up' is a verb with an adverb modifying the action 'look', the second is a phrasal verb with a verb and an adverb and the third is a phrasal verb with a verb, an adverb and a preposition (i.e. 'look up to'). Each of the above occurrences of the words 'look up' in a sentence conveys a different meaning, thus resulting in a different Chinese translation for each case. With many different ways to combine a verb, an adverb and/or a preposition and the fact that each combination can have more than one meaning, syntactic processing alone cannot provide adequate information for the lexical disambiguation and selection processes.

According to Bresnan (1995), thematic information helps to characterise different event structures. Event structure, unlike syntactic structure, is relatively universal across languages. It captures some information about the real-world by describing the kind of participants who take part in an event. This means that similar event structure in different languages is very likely to have similar a-structures. As a result, a-structure can be used to aid word sense disambiguation during the lexical selection process.

Based on the over 3000 phrasal verbs listed in Sinclair (1989), an investigation into the ability of a-structure to aid the selection of the appropriate Chinese translation for each English phrasal verb was carried out. Amongst the listed phrasal verbs, over 60% of them do not have multiple meanings. For those phrasal verbs which have multiple meanings, it is found that the different meanings of over 100 of them can be distinguished by considering their corresponding a-structures. For instance, consider the following sentences with the occurrences of the words '*look up*':

	Meaning of '*look up*'	**Chinese Translation**
1.	to raise one's eyes	向上看
2.	to become better	向好發展
3.	to find information from a dictionary or a reference book	查
4.	to visit	探訪

Table 5.2: Different Meanings of '*look up*'

(106)
1. John is *looking up*.
2. The business is *looking up*.
3. John is *looking up* a vocabulary.
4. John is *looking* Mary *up*.

Each occurrence of 'look up' above refers to a different event. Thus, the words 'look' and 'up' in each of the above cases have different meaning and thus resulting in different Chinese translation (cf. Table 5.2 and (107)).

(107)
1. 約翰 正在 向上 看。
   John at the moment up look.
2. 生意 正在 向好 發展。
   business at the moment towards-good develop.
3. 約翰 正在 查 一 個 詞彙。
   John at the moment look up one QUANTIFIER vocabulary.
4. 約翰 正在 探訪 瑪莉。
   John at the moment visit Mary.

Recall that the definitions of the thematic roles 'theme' and 'patient' adopted in this study are: the argument "*of which location or state is predicated*", or the argument of "*change of location or state*" is the theme; and the argument which displays the locus of the effect is the patient (cf. Section 4.1.4). When considering the nature of the participants involved in each of the events described in (106), it is found that each of the sentences in (106) is described by a different a-structure. Table 5.3 shows the a-structure argument(s) of the sentences in (106) and (107).

According to the lexical mapping theory, both theme and patient possess the feature $[-r]$ intrinsically[4]. Therefore, while carrying out the lexical mapping, the NPs 'a vocabulary' and 'Mary' can be mapped with either theme or patient. This causes a problem while assigning an appropriate a-structure for each case. The main distinction between the 'look up' events in each case is that the theme in the case of 'to find information' is an *inanimate* object; whereas the patient

[4]cf. Section 4.3.3

	English	A-structure Argument(s)	Chinese
1.	look	agent	向上看
2.	look up	theme	向好發展
3.	look up	agent, theme	查
4.	look up	agent, patient	探訪

Table 5.3: The a-structure arguments for 'look up' and its Chinese equivalents

in the 'to visit' case is an *animate* object (i.e. to look *someone* up). Therefore, by introducing a simple semantic marker $[\pm animate]$, the problem in the lexical mapping can be resolved.

In addition to aiding lexical mapping, the use of some simple semantic markers can also help to carry out the lexical selection process. For instance, one can say:

(108) John falls for Mary.

(109) John falls for a trick.

The phrasal verb 'fall for' has different Chinese translation in each case (i.e. 爲…而傾倒 and 被…騙 respectively):

(110) 約翰 爲 瑪莉 而 傾倒。
John because / for Mary fall down (emotionally).

(111) 約翰 被 一 個 詭計 騙 了。
John *bei* one QUANTIFIER trick (noun) trick (verb) ASPECT MARKER.

Though both of these sentences take an experiencer and a theme as their a-structure arguments, the theme in (108) is a person, which is an animate object; whereas in (109), the theme is an inanimate object. This difference in the semantic property of the theme characterises the difference in the resulting Chinese translation. Thus, the use of the semantic marker $[\pm animate]$ is capable of helping an MT system to choose the appropriate translation for the phrasal verb 'fall for'.

5.3 Aiding Sentence Generation

In the target sentence generation process, target language words obtained after lexical selection are reorganised according to the target language structure. Due to the differences in the use of words, grammar, way of expressing ideas, etc. between different languages, in addition to reorganising the selected target language words according to the target language grammar, additional words often need to be inserted and / or existing words need to be deleted or modified so as to produce a meaningful and grammatical target language sentence. Syntactic information of sentences is often inadequate to support more sophisticated sentence generation

process. When considering English and Chinese as the source and target languages in an MT process, two of the problems in target sentence generation observed are handling verb copying in Chinese and distinguishing the position where a prepositional phrase (PP) should appear in a Chinese sentence.

5.3.1 Verb Copying in Chinese

Chang (1991) observed that in some Chinese sentences which involve the expression of the duration or frequency of the event described, the verb of the sentence is copied. For instance:

(112) 約翰 烹煮 一 隻 雞 烹煮 了 兩 小時。
John cook a QUANTIFIER chicken cook ASPECT MARKER two hour
John cooked a chicken for two hours.

(113) 約翰 吃 一 個 蘋果 吃 了 一 小時。
John eat one QUANTIFIER apple eat ASPECT MARKER one hour.
John ate an apple for one hour.

(114) 約翰 看 這 本 書 看 了 三 次。
John read this QUANTIFIER book read ASPECT MARKER three times
John read the book three times.

Though in English, information about the duration of an event tends to be expressed in a PP with the preposition 'for', we cannot conclude that whenever there is an occurrence of a PP with the preposition 'for' in an English sentence, the resulting Chinese translation would require verb copying. It is because, as we have seen in the sentence "John cooked a chicken for Mary." in Figure 5.12, not every PP with the preposition 'for' expresses the duration of an event[5]. In fact, as observed by Chang (1991), not every Chinese sentence with information about duration or frequency of an event requires verb copying. Verb copying in some of these sentences can make them become ungrammatical. For instance, one can say (115), but not (116).

(115) 約翰 打碎 花瓶 三 次 了
John shatter vase three times ASPECT MARKER
John shattered a vase three times.

(116) ? 約翰 打碎 花瓶 打碎 了 三 次 了
John shatter vase shatter ASPECT MARKER three times ASPECT MARKER

Chang (1991) observed that if a Chinese sentence expresses frequency or duration of an event and is described by one of the following a-structure constructs:

[5]The PP 'for Mary' in "John cooked a chicken for Mary." denotes the beneficiary of the event.

(117)
- <agent locative>
- <theme locative>
- <agent theme recipient>

no verb copying is required[6]. Syntactic and functional information is insufficient to indicate this kind of information. If an MT system performs sentence generation by using information encoded in traditional c-structure and f-structure (cf. Chapter 3), it will have difficulty in determining when an additional verb should be generated while translating a sentence like "John ate a chicken for three hours.". Thematic information, however, which shows the role that the governing NP plays, provides sufficient information to indicate the need for verb copying in a Chinese sentence.

Consider the following sentences and their corresponding a-structures:

(118)
1. "John cooked a chicken for two hours."
   A-structure: cook<agent theme> for<duration>
2. "John ate an apple for three hours."
   A-structure: eat<agent theme> for<duration>
3. "John read the book for a long time."
   A-structure: read<agent theme> for<duration>
4. "John played the piano for five hours."
   A-structure: play<agent theme> for<duration>

The Chinese translations of the sentences in (118) require verb copying (e.g. (112) and (113)). The a-structures of these sentences all govern very similar participants: agent and theme, and they all have the sub-structure 'for<duration>'. When generating Chinese sentences which involve an expression of duration of the events, an MT system can determine whether to perform verb copying by using thematic information encoded in a-structures. For instance, if an English sentence has similar a-structure as those in (118), during the lexical selection process, instead of translating the preposition 'for' to '爲' in Chinese (cf. the Chinese translation for the sentence "John cooked a chicken for Mary." in Figure 5.12), the head of the a-structure (e.g. the verb 'cook' in "John cooked a chicken for two hours.") is 'copied' to the position of this preposition (i.e. 'for'). Thus, the a-structures of the examples in (118) become:

[6] cf. Chang (1991) for a full discussion and illustration of sentences which do not required verb copying.

(119)
1. 烹煮<agent theme> 烹煮<duration>
2. 吃<agent theme> 吃<duration>
3. 看<agent theme> 看<duration>
4. 彈<agent theme> 彈<duration>

if they are to be translated to Chinese. While generating the target Chinese sentence, instead of inserting the aspect marker '了' to the position after the main verb (e.g. the Chinese translation for "John cooked yesterday." in Figure 5.12, this aspect marker is inserted after the 'copied' verb, i.e. in front of the duration NP (cf. the sentences (112) and (113)).

Likewise, thematic information encoded in an a-structure can also be used to determine if there is the need for verb copying. As mentioned earlier, Chang (1991) observed that there are some Chinese verbs which do not require verb copying. These Chinese verbs tend to be described by the a-structures with the arguments similar to those in (117). Therefore, while generating Chinese sentences which express the duration of the events concerned, if their a-structures involve one of the constructs in (117), there will not be a second occurrence of the main verb within the resulting sentences.

One might argue that the term 'duration' does not appear in either of the thematic hierarchies (i.e. (61) and (62)) cited in Chapter 4. Therefore, it should not appear as an argument in an a-structure. However, recall that[7]:

> "*The term* LOCATIVE *will be used to subsume a broad range of spatial locations, paths, or directions, and their extensions to some temporal and abstract locative domains . . .*"
>
> [Bresnan (1994, page 75)]

Bresnan (1994) suggested that the thematic role '*locative*' can be used to describe temporal information. The term 'duration', which expresses for how long an event takes place, is a kind of temporal argument. Therefore, it can be viewed as a subcategorisation of 'locative' which denotes temporal information of an event. The thematic role 'locative' is classified as $[-o]$ intrinsically[8], thus the argument 'duration' also bears the feature $[-o]$ intrinsically.

5.3.2 Positioning PPs within a Chinese Sentence

In English, all prepositional phrases (PPs) can appear after the verb which they modify, e.g.:

- John prepared a meal for Mary in the barn.
- John played a duet on the piano with Mary.
- John knocked on the door.

[7] cf. Section 4.1.5
[8] cf. Section 4.3.3

- John delivered a letter to Mary.
- John waited for Mary at the cinema for three hours.

Sometimes a PP can also be placed at the beginning of an English sentence (e.g. in the case of locative inversion). For instance:

- In the barn, John prepared a meal for Mary.
- At the door, the security guards checked the identity of every participant.
- For Mary, John sacrificed his life.

However, in Chinese, some PPs can only appear before the verb, e.g.:

(120) 約翰 爲 瑪莉 烹煮 一 隻 雞 烹煮 了 兩 小時。
John for Mary cook a QUANTIFIER chicken cook ASPECT MARKER two hour
John cooked a chicken for Mary for two hours.

(121) 約翰 從 倫敦 出發。
John from London set off.
John set off from London.

(122) 約翰 往 倫敦 去 了。
John to London go ASPECT MARKER.
John went to London.

(123) 約翰 在 沙發 上 睡覺。
John at sofa top sleep.
John sleeps on the sofa.

an some can only appear after the verb, e.g.:

(124) 約翰 駕車 往 倫敦。
John drive-vehicle to London.
John drives to London.

Therefore, while generating a Chinese sentence, an MT system cannot simply use the phrase structure rule (125) or (126):

(125) VP $\longrightarrow$ V NP PP

(126) VP $\longrightarrow$ PP V NP

to determine where a PP should appear in the sentence. Considering the preposition in the corresponding source English sentence, again, cannot solve this problem because, as shown in

(120), though the oblique PPs of the English translation "John cooked a chicken for Mary for two hours." are both governed by the preposition 'for', when they are translated to Chinese, they appear before (i.e. '爲 了') and after the main verb (i.e. the copied verb '烹煮' appeared after the NP '一隻雞') respectively. In addition, there is the need for an MT system to decide what the preposition 'for' should be translated in each case.

In order to translate each of the occurrences of the preposition 'for' in the sentence "John cooked a chicken *for* Mary *for* two hours." appropriately, similar to what we have discussed in Sections 5.2 and 5.3.1, we can consider the type of argument that this preposition governs in each case, i.e.:

(127) for<beneficiary>

(128) for<duration>

The sub-structures 127 and 128 govern different arguments. With the argument 'beneficiary', the preposition 'for' should be translated to '爲' in Chinese (e.g. from 'for Mary' to '爲瑪莉'); whereas the preposition 'for' in the second case, i.e. with the argument 'duration', should be subjected to verb copying in Chinese (e.g. from 'for two hours' to '烹煮了兩小時'). Similar to the problem described earlier in this chapter, these 'for'-PPs can be mapped with either (127) or (128) if we consider their $[\pm r]$ and/or $[\pm o]$ features only. Therefore, the lexical mapping for these PPs and their corresponding a-structures requires the aid of the semantic features $[\pm animate]$ and $[\pm physical]$ also.

Similar to the selection of appropriate translation for the preposition 'for', thematic information helps an MT system to determine whether the resulting Chinese PP should be placed in front of or after the main verb of a sentence. With the argument 'beneficiary', a Chinese PP is placed *in front of* the main verb; with the argument 'duration', a Chinese PP is placed *after*.

A similar method can be applied to determine the positioning of a PP with the Chinese preposition '往' in a sentence. Consider the a-structures for the sentences (122) and (124):

(129) 去<agent> 往<locative>

(130) 駕車<agent> 往<locative>

Bearing the a-structure (129), the '往'-PP in a Chinese sentence is positioned *before* the main verb of the sentence (cf. (122)). With the a-structure (130), the '往'-PP is positioned *after* the main verb of the sentence (cf. (124)). Similarly, the positioning of the PP in sentences (121) and (123) can also be determined by considering their corresponding a-structures, i.e. (131) and (132).

(131) 出發<agent> 從<locative>

(132) 睡覺<agent> 在…上<locative>

As discussed above, the meaning of PPs affects the generation of grammatical Chinese sentence. To some extent, thematic information within a-structures aids characterising the meaning of different prepositional phrases. As a result, by disambiguating the meaning of each PP through the use of corresponding a-structure, some of the problems in the generation of Chinese sentences can be resolved.

A-structures represent the necessary participants in an event structure systematically. This information, as shown above, can help an MT system to determine the positioning of PPs within a Chinese sentence. Therefore, a-structures can be used to indicate where the Chinese translation of PPs should appear within the target sentence.

5.4 Discussion

Traditional c-structures and f-structures which display the syntactic and functional information of sentences do not provide sufficient information for handling the disambiguation of source language words during the source-to-target language transfer. In this study, it is found that a-structure, which shows the thematic information of sentences, is capable of providing linguistic information which is more adequate for handling disambiguation than c-structure and f-structure. However, the use of a-structure and lexical mapping theory for MT is inadequate to handle the disambiguation of all verbs. For instance, in Section 5.1.1, we looked at how a-structure and lexical mapping theory can be used to differentiate the structure 'V + PP' from 'Phrasal Verb + NP' in English sentences, especially those which contain a locative adjunct. However, one potential problem in this method is that most physical objects, ranging from a tiny cigarette stub to a mountain, can serve as a locative. For instance, one can say "John drew on a cigarette." to mean drawing some kind of a picture on a cigarette for decoration. However, in most cases, this sentence means to breath in through a cigarette and inhale the smoke deeply (Sinclair 1989, page 88). It is therefore impossible to eliminate many potential ambiguities based only on whether or not something is serving as a locative or a theme in an event. However, with this kind of sentence, where both literal and modified meaning of a verb-and-preposition pair are valid, it is not desirable to discard one of the two possible meanings.

Though, as discussed in Section 4.1, each thematic role possesses certain semantic properties which enable it to facilitate a certain level of semantic disambiguation in an MT system, the relatively brief semantic information present in an a-structure is still insufficient to handle the transfer of all English verbs to Chinese appropriately. An example of the kind of verbs which cannot be disambiguated by this method are those that can be used extensively in various situations to express a subtle scale of similar meanings in the source language, but have a distinctive translation for each case in the target language. For instance, the English verb '**break**' for denoting the change-of-state of an object has numerous translations in Chinese (Palmer & Wu 1995).

In English, one may use the verb 'break' to describe various kinds of breaking events:

1. John *broke* a window.
2. John *broke* a rope.
3. John *broke* a vase with a hammer.

Each occurrence of the English verb 'break' in the above sentences has a distinctive translation in Chinese:

1. John broke a window.
   Chinese translation: 約翰 **打破** 了 窗子。
   John break ASPECT MARKER window.
2. John *broke* a rope.
   Chinese translation: 約翰 **弄斷** 了 繩子。
   John snap ASPECT MARKER rope.
3. John *broke* a vase with a hammer.
   Chinese translation: 約翰 用 鎚子 **打碎** 了 花瓶。
   John use hammer hit-to-pieces ASPECT MARKER vase.

All of the above sentences have very similar a-structure arguments (i.e. agent and theme), but each of the occurrences of 'break' is translated differently. Thus, the lexical selection of the verb 'break' for the above sentences cannot rely on their a-structures. To transfer these kinds of verbs successfully, a much higher level of semantic information is required.

Recent work carried out by Palmer & Wu (1995) on handling the disambiguation of words with one-to-many translations in the target language used selectional restrictions and conceptual primitives. An interlingual conceptual lattice is built by merging the hierarchies of conceptual primitives for verb senses in English and Chinese. The lexical selection was performed by calculating the meaning similarity between words within the conceptual lattice and selecting the best translation based on the calculated meaning similarity. The result in Palmer & Wu (1995) showed that this method is capable of disambiguating different verb senses of the English verb 'break' when translating it to Chinese. It can also approximate the most appropriate source-to-target language translation from the existing lexicon if the appropriate verb sense is not defined in the conceptual lattice. This is particularly useful when the required MT system is not confined to processing a sublanguage only, but a broader coverage of a natural language, as it is impossible to specify a complete collection of verb senses for any verb. The drawback of this method is that in order to ensure its effectiveness, a fairly broad range of verb senses has to be incorporated. As a result, a complicated conceptual lattice is required to be built. The larger the lexicon, the bigger and more complicated the conceptual lattice will become. In addition to the difficulty in handling a large and complicated conceptual lattice, a lot of time and human effort will also be required to build this lattice for the resulting MT system, thus making this method relatively costly and difficult to implement for real-life MT tasks.

Although thematic information is inadequate to support the kind of high-level semantic disambiguation in Palmer & Wu (1995), it allows the disambiguation of a wide range of words

(e.g. verbs and prepositions) whose translations are dictated by their governing thematic roles to be performed in a relatively less costly and simple way. In addition, as was illustrated earlier in this section, thematic information encoded in a-structure is capable of alleviating various problems in different stages of MT processing, i.e. lexical and structural disambiguation, lexical selection and sentence generation, which a lot of methods for lexical selection fail to do. It can also be used in conjunction with other kinds of linguistic information, e.g. the syntactic and functional information encoded in c-structure and f-structure. Unlike the proposed method, other existing methods for lexical selection which involve the use of complicated semantic networks, e.g. Palmer & Wu (1995), require some kind of additional mechanism to integrate it with other parts (e.g. parsing of source sentences and generation of target sentences) of an MT system.

This approach to MT uses a-structure and lexical mapping theory to help to solve some of the common problems in lexical and structural disambiguation, e.g. to disambiguate source language verbs and prepositions which have one-to-many translations in a target language. This approach relies on the definition of a-structure for each event to disambiguate those words which possess multiple translations. As an event structure is characterised by the participants in the event, but not by a detailed description of what happened, an event structure in one language sometimes, though fairly rarely, may not have an equivalent in another language. One example observed is the a-structure 'notify<agent theme> of<instrument>' for describing the event presented in the sentence "*John notified the police of the robbery.*". This a-structure has no equivalent a-structure in Chinese, though '*someone* notified *somebody* of *something*' can be translated to '*someone* 通知 *somebody something*' in Chinese. For this kind of non-matching event, it is more difficult, though not impossible, for thematic information within a-structure to aid the lexical selection. However, with the verbs 'notify' and '通知', as there is a one-to-one correspondence between these verbs, the lexical selection can be carried out based on the translation correspondence. This means that, if an MT system encounters the a-structure 'notify<agent theme> of<instrument>', it will select the a-structure '通知 <agent theme recipient>' during the lexical selection. This Chinese a-structure can then be used to aid the target sentence generation.

In the earlier sections of this chapter, we have looked at the inadequacy of using the $[\pm r]$ and $[\pm o]$ features alone in an MT system to perform lexical mapping. Though thematic roles, as described in Section 4.1, are defined according to the role they played in an event and thus each of them contains some kind of semantic properties, this does not mean that a computer system automatically realises the semantic properties of, say, the thematic role 'agent' whenever the term 'agent' is encountered. Therefore, some kind of mechanism is required to enable an MT system to 'understand' the semantic properties possessed by a thematic role. This can be achieved by representing these semantic properties as semantic markers (e.g. $[\pm animate]$ and $[\pm physical]$) as additional features to the thematic roles within the a-structure. The introduction of some semantic markers to enhance the understanding of the underlying semantic properties of each

thematic role alleviates the inadequacy of an MT system to perform lexical mapping.

Although this approach cannot support high level semantic disambiguation, it provides a readily applicable (i.e. no need for pursuing extensive knowledge of formal linguistics) approach to MT that can be implemented through simple feature matching. As a computer is suited to symbolic processing and matching, this approach is computationally viable.

5.5 Conclusion

Inspired by the problems in MT discussed in Chapters 1 and 2, an investigation into the effectiveness of a-structure and lexical mapping theory to alleviate some of these problems was carried out. This chapter reported the results of this investigation. It was found that, in many cases, a-structure provides sufficient information for disambiguating source language words, for selecting the appropriate target language translation and for reorganising the target language words to form the desirable target language sentences. This chapter exemplified how a-structure and lexical mapping theory can alleviate some of the problems caused by lexical and structural ambiguities involving different combinations of verbs and prepositions and adverbs.

The different disambiguation processes described in this chapter were made possible by the thematic information represented in a-structures. As discussed in Chapter 4, a-structure shows the necessary participants involved in an event structure. When compared with the semantic form of the verb, a-structure is less language dependent because the information it captures is not purely syntactic-based. It also describes some level of lexical semantic information by means of semantic properties encoded in each thematic role, e.g. an agent must be an animate object as it is the conscious initiator of an event. Bearing this kind of information, a-structure reveals some kind of real-world knowledge about the event described by a sentence.

When compared with the s-structure in LFG proposed by Halvorsen & Kaplan (1988), a-structure carries more substantial semantic information about what happened in the real-world, instead of merely showing the relationship between each syntactic argument in a sentence. This kind of semantic information is very useful to helping a computer system to ‘understand’ the meaning of natural language sentences. Evidence of this capability of a-structure is shown throughout this chapter, e.g. the disambiguation of different sentential structures, the disambiguation of word senses, the selection of the most appropriate target language translation, etc. Each of the illustrated functions of a-structure and lexical mapping theory helps an MT system to reduce the chance to ‘misunderstand’ a source language sentence and thus increases its chance to produce an appropriate and grammatical target language sentence.

This chapter also discussed some of the weaknesses observed in applying a-structure and lexical mapping to MT. As illustrated in Section 5.2, a-structure is capable of disambiguating the different possible uses of some verbs. However, as thematic information only shows the types of participants involved in an event, but not the context of the sentence in detail, it does not

provide sufficient information for disambiguating all verbs. In some cases, the same a-structure can be used to describe different event structures which bear similar meanings with very subtle differences. These a-structures will be inadequate to support the disambiguation required. As a-structures are not designed to represent the complete underlying meaning of sentences, they cannot support high level semantic disambiguation.

If an MT system is required to produce high quality translation of highly ambiguous sentences without human intervention, a sophisticated method (e.g. the use of semantic network to represent meaning of words) in word sense disambiguation would be needed. However, this kind of disambiguation method often is complicated to implement and a separate method (e.g. the use of a linguistic formalism) is required to deal with the linguistics of sentences so as to facilitate the analysis and generation of sentences. Thus, the implementation of the resulting MT system would not be very straight-forward. The use of a-structure and lexical mapping theory in MT, however, is fairly straight-forward and no additional mechanism is required to bind different kinds of linguistic analysis together because they are in fact a part of a well-developed and frequently-used linguistic formalism for NLP — the LFG formalism.

Chapter 6

Dealing with the Transfer of Passive Sentences

In many languages, passivization allows the *focus* of a sentence to be changed from the *agent* (i.e. the "do-er") of the event described by the sentence to the *theme* (i.e. the participant of the event which undergoes the action). Instead of appearing as the object (OBJ) of a sentence, the theme becomes the subject (SUBJ) of the resulting passive sentence. The agent of the event can either appear as an oblique (OBL) or be omitted from the resulting passive sentence. This characteristic of passivization, as stated by Bresnan (1982*a*), is universal across languages. Though passive sentences in different languages share the same characteristic, they are dissimilar in some ways. While transferring a passive sentence from one language to another, a way to bridge this dissimilarity is required. This chapter will illustrate a method to deal with the difference between English passive sentences and their Chinese counterparts in the transfer.

6.1 Using F-structure as a medium for Source-to-Target Language Transfer

Her et al. (1994) suggested that f-structure provides a suitable medium for transfer as it is order-free and relatively language-independent. While the attribute-value bundles representation of f-structures provides a flexible medium to express syntactic and functional information of sentences in different languages, it does not mean that the f-structures of any sentence in one language and its equivalent in another language have identical structures[1]. If these f-structures are the same, the source-to-target language transfer would be straight-forward. For instance, the transfer of the f-structure of the sentence "John likes Mary." to Chinese can simply be done by transferring the predicators of the English sentence to Chinese:

[1] An f-structure of a sentence in one language is said to be identical with its equivalent in another language if they have similar attributes which form the same overall structure, but different values for each attribute.

English: John likes Mary. $\Longrightarrow$ **Chinese: 約翰喜歡瑪莉。**

$$\begin{bmatrix} \text{PRED} & \text{`LIKE}\langle(\uparrow \text{SUBJ})\ (\uparrow \text{OBJ})\rangle\text{'} \\ \text{TENSE} & \text{PRESENT} \\ \text{PERSON} & \text{3RD} \\ \text{NUMBER} & \text{SG} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`JOHN'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{`MARY'} \end{bmatrix} \end{bmatrix} \qquad \begin{bmatrix} \text{PRED} & \text{`喜歡}\langle(\uparrow \text{SUBJ})\ (\uparrow \text{OBJ})\rangle\text{'} \\ \text{TENSE} & \text{PRESENT} \\ \text{PERSON} & \text{3RD} \\ \text{NUMBER} & \text{SG} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`約翰'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{`瑪莉'} \end{bmatrix} \end{bmatrix}$$

No further transformation on the source f-structure is required to form the target f-structure. However, for the majority of sentences, the transfer is not as straight-forward. For instance, as pointed out by Her et al. (1994), some causative verbs in English, e.g. '*sadden*', which take a subject and an object as mandatory arguments bear the structure [SUBJ + V + OBJ + XCOMP]. For instance, according to Her et al., a sentence like "John saddens Mary." is translated to:

(133) 約翰 使 瑪莉 悲傷。
John cause Mary sad.
John saddens Mary.

and their corresponding f-structures are:

English: John saddens Mary.

$$\begin{bmatrix} \text{PRED} & \text{`SADDEN}\langle(\uparrow \text{SUBJ})\ (\uparrow \text{OBJ})\rangle\text{'} \\ \text{TENSE} & \text{PRESENT} \\ \text{PERSON} & \text{3RD} \\ \text{NUMBER} & \text{SG} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`JOHN'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{`MARY'} \end{bmatrix} \end{bmatrix}$$

Chinese: 約翰使瑪莉悲傷。

$$\begin{bmatrix} \text{PRED} & \text{`使}\langle(\uparrow \text{SUBJ})\ (\uparrow \text{OBJ})\ (\uparrow \text{XCOMP})\rangle\text{'} \\ \text{TENSE} & \text{PRESENT} \\ \text{PERSON} & \text{3RD} \\ \text{NUMBER} & \text{SG} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`約翰'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{`瑪莉'} \end{bmatrix} \\ \text{XCOMP} & \begin{bmatrix} \text{PRED} & \text{`悲傷}\langle(\uparrow \text{SUBJ})\rangle\text{'} \\ \text{SUBJ} & \end{bmatrix} \end{bmatrix}$$

As shown in the above f-structures, to transfer these kinds of sentences from English to Chinese requires a transformation over the appearance of the source f-structure: inserting new attribute-value pairs, e.g. XCOMP, and establishing a link between two attributes, i.e. (↑ OBJ) = (↑ XCOMP SUBJ). Her *et al.* handle this transformation by using the information on the necessary f-structure change that has already been specified in the system as attribute-value pairs. For instance, the relevant information (i.e. transfer entry) for transferring the verb 'sadden', as presented by Her et al. (1994, page 205) is as follows:

(134) **ec_sadden ::**

$$\begin{bmatrix} \text{WORD} & \begin{bmatrix} \text{PRED} & \text{‘悲傷’} \\ \text{TECH} & \text{GENERAL} \\ \backslash\ \text{STD-T-CAUS-V} & \end{bmatrix} \\ \backslash\ \text{STD-T-TRAN} & \end{bmatrix}$$

(135) **eX_STD-T-CAUS-V ::**

$$\begin{bmatrix} \text{ACTION} & \begin{bmatrix} \langle \uparrow \text{XCOMP SUBJ} \rangle = \langle \uparrow \text{OBJ} \rangle \\ \text{TRAN } (\downarrow \text{PRED} \rangle, \langle \uparrow \text{XCOMP PRED} \rangle) \\ \text{REPVAL } (\uparrow, \text{PRED, ‘使’}) \end{bmatrix} \end{bmatrix}$$

The transfer entry (135) does not act like an ordinary attribute-value pair for representing linguistic information about a sentence. This attribute-value pair specifies a list of actions required for the transformation of the source f-structure to the target f-structure. During the transfer, the appropriate sets of attribute-value pairs (as illustrated above) are selected through instantiation and the target f-structure is produced accordingly.

This method is relatively straight-forward because the whole transformation is based on instantiation and unification. As all the information required for the transformation is specified in the system as transfer entries, this method can handle different kinds of f-structure transformation. However, this is also a disadvantage of this method. As the transfer relies solely on the information specified in the transfer entries, all the necessary transfer entries must be encoded in the system beforehand.

The verb ‘sadden’ and its equivalent in Chinese ‘使…悲傷’ demonstrated one kind of translation differences between the source and target language words which results in different source and target f-structures. There are many other kinds of these differences even between the same pair of languages (i.e. English and Chinese). To encode all these differences in terms of a number of hand-crafted transformation rules (as shown in (135)) in the system is a tedious task. Therefore this method is quite costly in terms of time and human effort. In addition, this method requires expertise on the theoretical linguistics aspects on both the source and target languages to specify all the transformation rules in the system. Therefore, someone who does not have such expertise cannot readily employ this method.

As illustrated in the above, the f-structures of an English sentence and its Chinese counterpart are not necessarily identical. Further dissimilarity is found between the f-structures of English passive sentences and their corresponding Chinese counterparts. In the following, we are going to look at this dissimilarity and how it affects the translation process for English and Chinese passive sentences. Her et al. (1994) suggested that f-structure provides a suitable medium for transfer and they demonstrated a method which uses f-structure as a medium to transfer source f-structures to target f-structures. Though Her *et al.* used f-structures as a sole medium for

the transfer, their method still relied on the use of transformation rules (e.g. the one shown in (135)) to specify how a target f-structure can be produced from the source f-structure. The main problem in using f-structure as a medium for the transfer is that the linguistic information about a sentence represented in an f-structure is not common to all languages. Thus, other means to govern the transformation (e.g. the use of transformation rules in Her *et al.*'s case) are required in the transfer. If a structure for representing linguistic information which is more universal to different languages is used for the transfer, the use of additional transformation rules for the transfer might be avoided. The application of a-structure and lexical mapping theory in aiding the transfer from English sentences to Chinese is investigated. It is found that this method enables an effective transfer from English passive sentences to Chinese without the use of additional transfer rules. This chapter will illustrate how this can be done.

6.2 Passive in English

According to Bresnan (1982*a*, page 9), passivization in English can be characterised by the following changes:

(136)

Functional change:	(SUBJ)	$\longrightarrow$	$\emptyset$ / (BY OBJ)
	(OBJ)	$\longrightarrow$	(SUBJ)
Morphological change:	V	$\longrightarrow$	$V_{[participle]}$

E.g. eat<(SUBJ) (OBJ)> $\longrightarrow$ eaten<(OBL)/$\emptyset$, (SUBJ)>

As described in (136), the main verb of a passive sentence (i.e. the verb which governs the event structure described by the sentence) appears as a *participle* and it is preceded by the auxiliary verb '*be*'. The object of the active form of the sentence is *raised* to the subject position in a passive sentence and the subject can appear as an oblique object governed by the preposition '*by*'. For instance, the English sentence "*John killed Mary.*" can be transformed to passive as either (137) or (138):

(137) Mary was killed by John.

(138) Mary was killed.

The partial lexical entries for the auxiliary verb '*was*' and the past participle form of the verb '*kill*' in English passive sentences are shown in (139) and (140) (cf. Chapter 3):

(139)

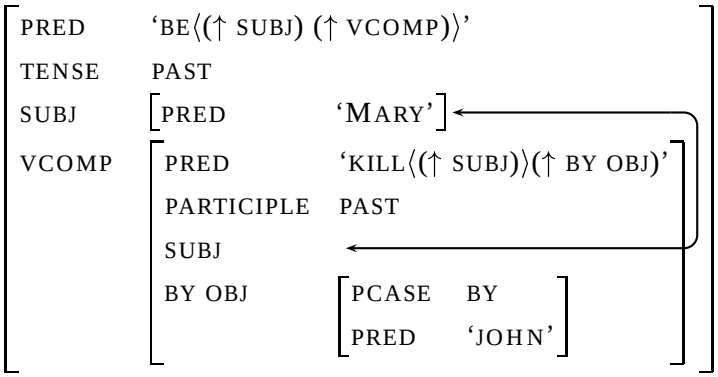
was:	V,	(↑ TENSE) = PAST
		(↑ PRED) = 'BE<(↑ SUBJ) (↑ VCOMP)>'
		(↑ VCOMP PARTICIPLE) = PAST
		(↑ VCOMP SUBJ) = (↑ SUBJ)

(140)

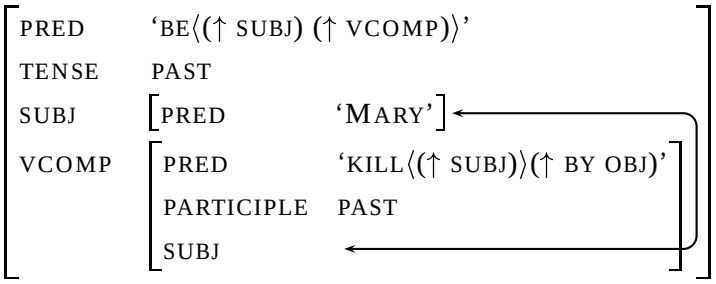
killed:	V,	(↑ PARTICIPLE) = PAST
		(↑ PRED) = 'KILL<(↑ SUBJ)> (↑ BY OBJ)'

According to (136), the oblique object 'BY OBJ' is not a mandatory argument in a passive sentence. This property is shown in the lexical entry (140) where the oblique object 'BY OBJ' appears outside the angle brackets. As defined in the lexical forms (139) and (140), the f-structures for (137) and (138) are:

Mary was killed by John.



Mary was killed.



A similar transformation applies to ditransitive sentences. A ditransitive sentence is a sentence which is made up of a ditransitive verb, e.g. :

- John *gave* Mary a book.
- John *told* Mary a story.
- John *handed* Mary a toy.
- John *lent* Mary a car.

A ditransitive verb takes a subject and two objects (i.e. a direct object and an indirect object). For instance, the ditransitive verb 'give' has the semantic form:

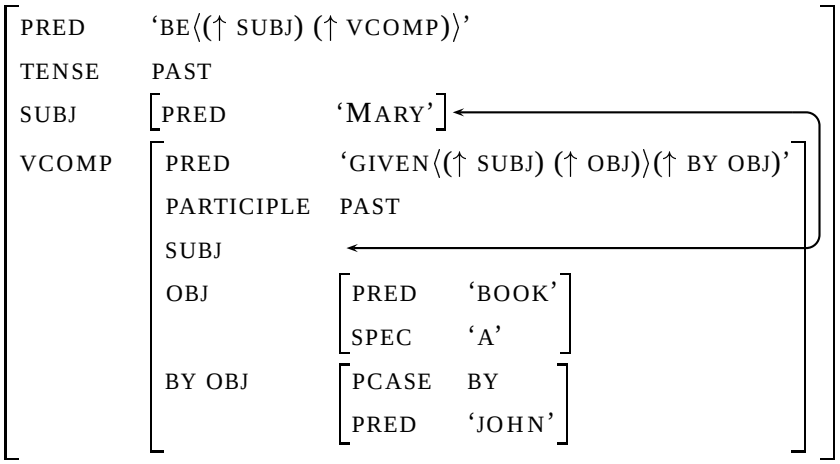
(141) give<(↑ SUBJ) (↑ OBJ2) (↑ OBJ)>

When transforming a ditransitive sentence to passive voice, the indirect object (i.e. OBJ) is raised to the subject position of the passive sentence and the direct object (i.e. OBJ2) appears immediately after the passivised verb 'given'. For instance, after passivization, the ditransitive sentence "*John gave Mary a book.*" becomes:

(142) Mary was given a book by John.

The same transformation applies to the above examples of ditransitive sentences. The f-structure for the passive sentence (142) is:

Mary was given a book by John.



In the passive sentence (142), the focus is on the recipient, i.e. the noun phrase (NP) '*Mary*', of the give-event. With the recipient of the sentence appearing as an object, the passive form of the ditransitive verb 'given' does *not* allow the theme of the sentence to be expressed as a subject. This means that the sentence "A book was given Mary by John." is grammatically incorrect. This syntactic behaviour can be explained by the a-structure of the ditransitive verb 'give' and the lexical mapping theory. The a-structure of 'give' is:

(143) give<agent recipient theme>

According to the lexical mapping theory, while a verb is undergoing passivization, the most prominent thematic role, i.e. the thematic role 'agent' in (143), in the a-structure is suppressed and the next most prominent thematic role, i.e. recipient, becomes the subject of the resulting passive sentence (cf. Section 4.3.3). The thematic role 'theme' in (143), being the least prominent thematic role, cannot be mapped to the subject position while keeping the second most prominent thematic role 'recipient' in the sentence as an object. If the focus is set on the theme, i.e. the NP '*a book*', the semantic form (144) of the verb 'give' will be required rather than the ditransitive form.

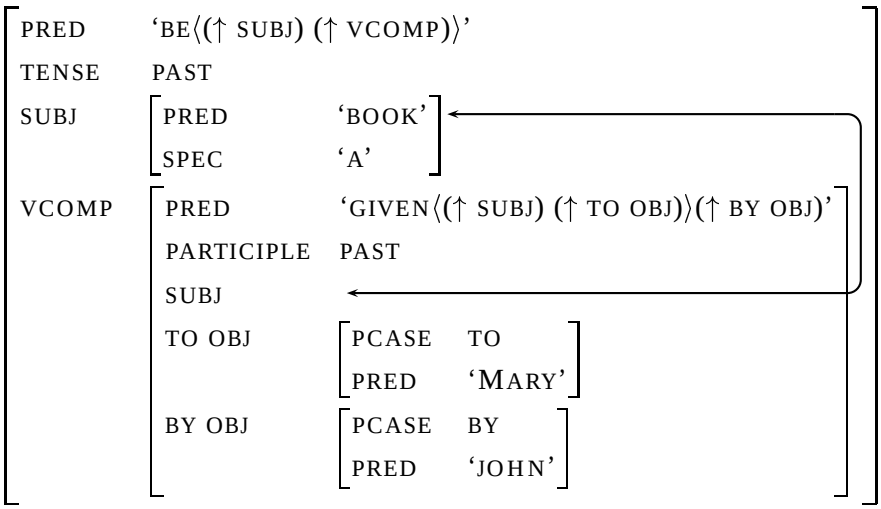
(144) give<(↑ SUBJ) (↑ OBJ) (↑ TO OBJ)>

Though using the semantic form (144) would not change the meaning of the sentence, the syntactic structure of the sentence would be changed. For instance, this semantic form yields the sentence "John gave a book to Mary.". After applying the passive transformation rules, the NP 'a book' will then become the focus of the resulting passive sentence:

(145) A book was given to Mary by John.

The f-structure for (145) is:

A book was given to Mary by John.



Note that not all ditransitive sentences can be transformed to the passive voice. For instance, as postulated by Bresnan & Zaenen (1990) and Huang (1993), the ditransitive verb '*cook*', as in the sentence "*John cooked Mary a chicken.*", cannot be transformed to passive voice as "* *Mary was cooked a meal by John.*". This can be explained by the a-structure of the ditransitive form of 'cook':

(146) cook<agent beneficiary theme>

In addition to classifying the internal argument 'beneficiary' in (146) as $[-r]$ (i.e. not thematically restricted) intrinsically, as observed by Bresnan & Zaenen (1990) and Huang (1993), the thematic role 'beneficiary' also bears the feature $[+o]$ (i.e. objective). During passivization, the thematic role 'agent' is suppressed. However, bearing the intrinsic classification $[+o]$, the next most prominent thematic role 'beneficiary', cannot be mapped to the syntactic function 'subject' which bears the classification $[-o]$ (i.e. non-objective). As a result, the passive expression of the sentence "*John cooked Mary a chicken.*" (i.e. "*Mary was cooked a chicken by John.*") is ungrammatical.

6.3 Passive in Chinese

Syntactically, the passive in Chinese is different from that in English. Most common passive sentences in Chinese are denoted by the presence of the word '被' and these sentences have similar syntactic and semantic structure to their English equivalents (cf. Wong & Hancox (1999, Sections 2 & 5.2)). While transforming a Chinese sentence from active voice to passive voice, the universal characteristic of passivization suggested by Bresnan[2] holds. For example, the Chinese sentence:

(147) 約翰 殺 了 瑪莉。
John kill ASPECT MARKER Mary.
John killed Mary.

can be transformed to the passive voice by using the word '被':

(148) 瑪莉 被 約翰 殺 了。
Mary *bei* John kill ASPECT MARKER.
Mary was killed by John.

Similar to passivization in English, the *agent* of a passive sentence can be omitted from the passive sentence. In Chinese passive '被'-sentences, the NP appearing immediately after the word '被' (e.g. the NP '約翰' in (148)) indicates the participant who initiates the event, i.e. agent. In a passive '被'-sentence, the agent of an event is not a mandatory argument. If the agent is not specified in the sentence, it is realised as '*someone*' (cf. Li & Cheng 1994, Chapter 7, Section IV). However, unlike the agent in an English passive sentence, the agent of a passive '被'-sentence is not explicitly marked by a preposition. To transform the sentence (148) to an *agentless* '被'-sentence, one simply removes the NP '約翰' and forms:

(149) 瑪莉 被 殺 了。
Mary *bei* kill ASPECT MARKER.
Mary was killed.

When comparing the English passive sentences (137) and (138) to the passive '被'-sentences (148) and (149), the word '被' seems to have a similar function to the auxiliary verb 'be' in English passive sentences. However, both syntactically and semantically, '被' and 'be' are not identical. This is shown by the fact that in addition to having a similar function to the English auxiliary verb 'be' which governs the passivization of the main verb of a sentence, '被' also marks the oblique agent of the sentence (cf. the sentences (148) & (149)).

Wong & Hancox (1999) gave a detailed discussion on how the word '被' should be represented within the LFG framework. In this thesis, the lexical form of '被' for Chinese passive sentences follows Wong & Hancox's (1999, page 176, (28)) suggestion, which is reproduced in

[2] cf. the beginning of this chapter or (Bresnan 1982*a*)

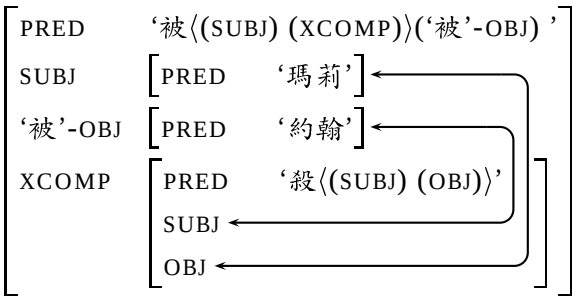
this thesis as:

(150)

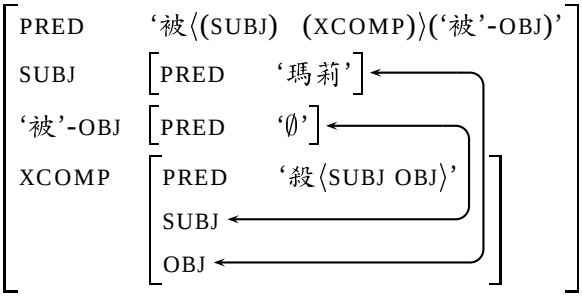
被：	V,	被‘<(SUBJ) (XCOMP)>(被-OBJ)’
		(↑ XCOMP SUBJ) = (↑ 被-OBJ)
		(↑ XCOMP OBJ) = (↑ SUBJ)

According to this lexical form, the f-structures for the sentences (148) and (149) are:

瑪莉被約翰殺了。



瑪莉被殺了。



As in English, there are verbs in Chinese which govern a subject and two objects, i.e. ditransitive verbs. The first object in a sentence with ditransitive verb is called the *indirect object* and it is mostly used to signify the *recipient* of the event described by the sentence. The second object is a *direct object* of the sentence and it is often used to express the *theme* of the event. Examples of Chinese sentences with ditransitive verbs are:

(151) 約翰 告訴 了 瑪莉 一 件 事情。
John tell ASPECT MARKER Mary a QUANTIFIER matter.
John told Mary a matter.

(152) 約翰 送 了 瑪莉 一 本 書。
John give ASPECT MARKER Mary a QUANTIFIER book.
John gave Mary a book.

(153) 約翰 問 了 瑪莉 一 題 問題。
John ask ASPECT MARKER Mary a QUANTIFIER question.
John asked Mary a question.

Some transitive verbs in Chinese can be made to behave as a ditransitive verb by adding the character '給'[3] (meaning '*to*' when it is used as a preposition or '*give*' when it is used as a verb), e.g. from '踢' (meaning '*kick*') to '踢給' (meaning '*kick-to*') and from '煮' (meaning '*cook*') to '煮給':

(154) 約翰 踢給 了 瑪莉 一 個 球。
John kick-to ASPECT MARKER Mary a QUANTIFIER ball.
John kicked a ball to Mary.

(155) 約翰 煮給 了 瑪莉 一 隻 雞。
John cook ASPECT MARKER Mary a QUANTIFIER chicken.
John cooked Mary a chicken.

Though the character '給' can appear in a sentence as a preposition meaning '*to*', it is not acting as a preposition in the above sentences because it appears before the aspect marker '了' and it does not appear immediately before an NP (e.g. the NP '瑪莉' in (154) or (155)). Aspect markers in Chinese, e.g. the words '過' and '了', are used to modify verbs only and thus they can only appear after a verb. If the character '給' is used as a preposition, it would appear immediately before the NP that it marks. In the sentences (154) and (155), the character '給' appears immediately before the aspect marker '了', but does not appear immediately before any NP; this suggests that the word '給' is acting as a part of a compound verb.

According to Li & Cheng (1994), the subject of a passive '被'-sentence represents the participant which undergoes the action described by the sentence, i.e. the *theme*. This means that, unlike the passivization of ditransitive sentences in English, instead of raising the first object in a Chinese ditransitive sentence (i.e. the recipient of the event) to the subject position, the second object, i.e. the *theme*, is raised to the subject position during passivization. For instance, the sentence (151), (152) and (154) become:

(156) 一 件 事情 被 約翰 告訴 了 瑪莉。
a QUANTIFIER matter *bei* John tell ASPECT MARKER Mary.
A matter was told to Mary by John.

(157) 一 本 書 被 約翰 送 了 瑪莉。
a QUANTIFIER book *bei* John give ASPECT MARKER Mary.
Mary was given a book by John.

(158) 一 個 球 被 約翰 踢給 了 瑪莉。
a QUANTIFIER ball *bei* John kick-to ASPECT MARKER Mary.
A ball was kicked to Mary by John.

respectively. Huang (1993) suggested that the reason for raising the second object (i.e. the direct

[3](cf. Section 4.3.3)

object) of a Chinese ditransitive sentence during passivization can be accounted for in terms of a-structure and the lexical mapping theory. Thematic roles within an a-structure are ordered according to the thematic hierarchy. According to the thematic hierarchy for Chinese stated in Section 4.3, the thematic role 'theme' appears before the thematic role 'recipient'. This means that the theme of a sentence in Chinese is more prominent than the recipient. The verbs '告訴', '送' and '踢給' all govern similar thematic roles: agent, theme and recipient, and form the a-structures:

- 告訴<agent theme recipient>
- 送<agent theme recipient>
- 踢給<agent patient recipient>

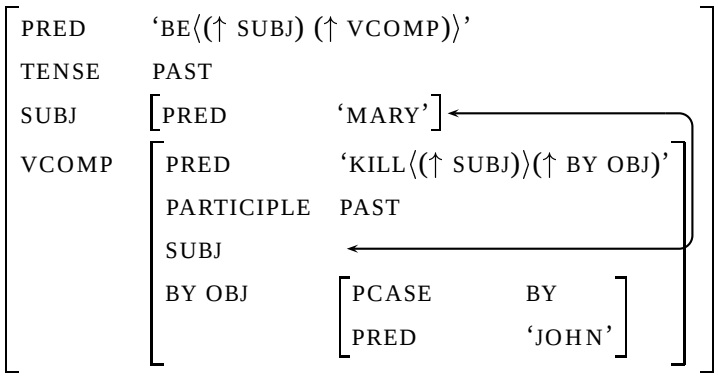
During passivization, the thematic role 'agent' is suppressed and the next most prominent thematic role, i.e. theme, is raised to the subject position. The thematic role 'theme' is intrinsically classified as $[-r]$ (i.e. non-restricted) and this classification does not conflict with the intrinsic classification for the syntactic function 'subject'. As a result, after undergoing passivization, the sentences (151) and (154) become (156) and (158) respectively.

6.4 Differences between Passive Sentences in English and in Chinese

According to the lexical forms (139) and (150) for the verbs 'be' and '被' respectively in passive sentences, the f-structures for sentence (137) "Mary was killed by John." and its equivalent in Chinese, i.e. (148) "瑪莉被約翰殺了。" are those shown in Figure 6.1 and the f-structures for the equivalent sentences (138) and (149) are those shown in 6.2. Though these f-structures are for sentences which have the same meaning and very similar syntactic behaviour, they, as shown above, have different structures. The differences between these f-structures are indicated in the lexical entries shown in (139), (140) and (150). Apart from governing the syntactic function 'subject', the auxiliary verb 'be' governs a verbal complement (VCOMP) only, but the verb '被' additionally governs an oblique agent as well as an open complement (XCOMP). Lexical entries describe the syntactic features and semantic content of lexical items and the formation of the f-structure for a sentence is dictated by the relevant lexical entries; different lexical entries for the English and Chinese passive indicators (i.e. the auxiliary verb 'be' and '被' respectively) produce different f-structures for passive sentences in English and in Chinese. One way to transfer (138) to (149) is to draw the translation correspondence between the lexical entries (139), (140) and (150) through observing the difference between these lexical entries. However, as shall see in later of this section, this method cannot be applied in some cases.

The difference between the above f-structures, again, supports the argument that f-structure is not universal across languages. As a result, when dealing with the translation of passive

Mary was killed by John.



瑪莉被約翰殺了。

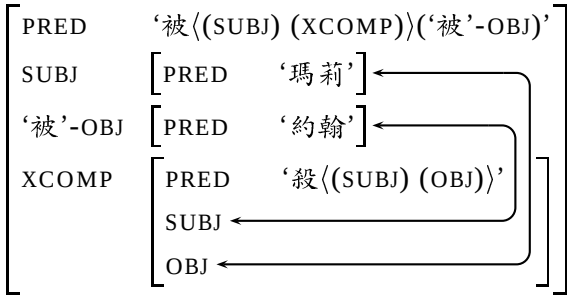


Figure 6.1: English and Chinese F-structures for “Mary was killed by John.”.

sentences from English to Chinese, the f-structure of an English sentence has to be *transformed* to the corresponding target language form for target language sentence generation. The grammar rules and lexical entries for constructing c-structures and traditional f-structures in LFG are language-dependent. Without the use of additional transformation rules for specifying how each source language f-structure can be transformed to its target language form, traditional f-structures are inadequate for using as a sole information-bearing medium for source-to-target language transfer.

Another difference between the f-structures of English and Chinese passive sentences is observed between the sentences with ditransitive verbs. Both the English verb ‘*give*’ and its Chinese equivalent ‘送’ take the same three arguments in their semantic forms:

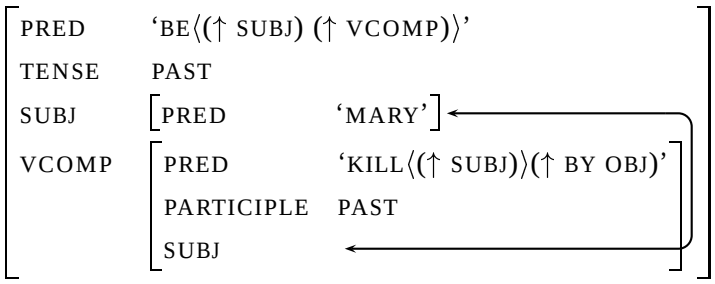
(159) give: V, (↑PRED) = ‘GIVE<(↑ SUBJ) (↑ OBJ2) (↑ OBJ)>’

E.g. John gave Mary a book.

(160) 送: V, (↑PRED) = ‘送<(↑ SUBJ) (↑ OBJ2) (↑ OBJ)>’

E.g. 約翰送了瑪莉一本書。 *(meaning “John gave Mary a book.”)*

Mary was killed.



瑪莉被殺了。

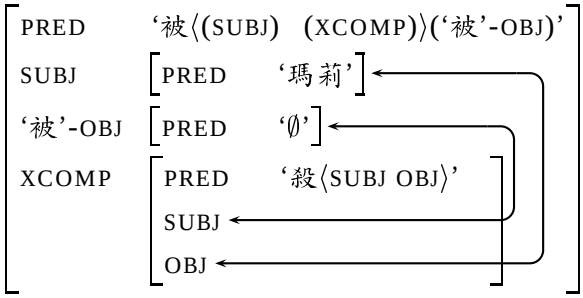


Figure 6.2: English and Chinese F-structures for “Mary was killed.”.

Though the sample sentences in (159) and (160) have the same syntactic structure and the passivizations in English and in Chinese share the same operations[4]: *object promotion* and *subject demotion* or *subject deletion*, after passivization, their syntactic structures no longer resemble each other. This is because when ditransitive sentences in English and in Chinese are undergoing passivization, as discussed in Section 6.2 and 6.3, the focus of the sentence is set on different NPs in the sentence. When a ditransitive sentence in English is undergoing passivization, the focus of the sentence is set on the *recipient* of the event by raising the second NP (i.e. the indirect object ‘*Mary*’ in (142)) to the subject position; whereas in Chinese the focus of the sentence is set on the *theme* by raising the the third NP (i.e. the direct object ‘一 本書’ in (152)) to the subject position:

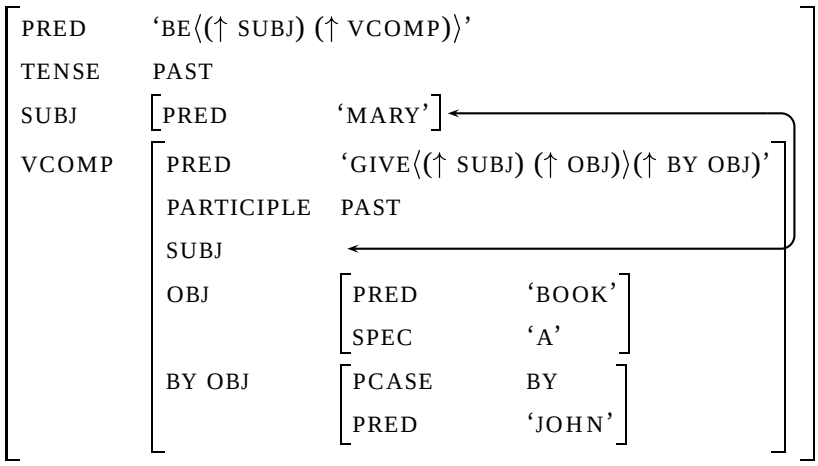
John gave Mary a book. ⟹ Mary was given a book by John.
約翰送了瑪莉一本書。 ⟹ 一本書被約翰送了瑪莉。

The corresponding f-structures for sentences (142) and (157) are shown in Figure 6.3. While the sample sentence in (159) can be translated to (160) directly, the direct translation of the passive form of the sentence in (159) (i.e. (142)), i.e. :

(161) * 瑪莉 被 約翰 送 了 一 本 書。
Mary *bei* John give ASPECT MARKER one QUANTIFIER book.

[4]cf. the Functional change during passivization in English shown in (136)

Mary was given a book by John.



一本書被約翰送了瑪莉。

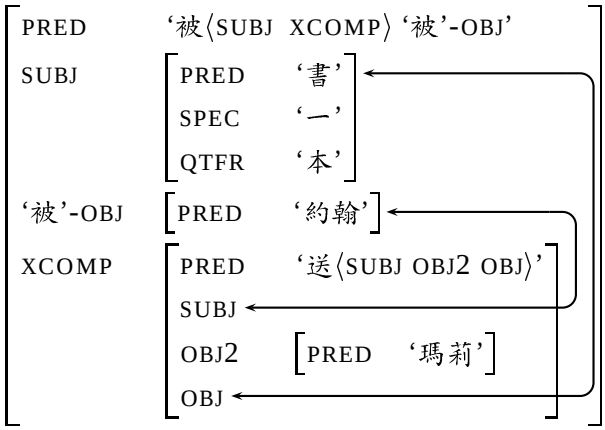


Figure 6.3: The English and Chinese equivalents of the sentence “Mary was given a book by John”

is grammatically wrong in Chinese. For English ditransitive sentences, if the second object (e.g. the NP ‘*a book*’ in “John gave Mary a book.”) is raised to the subject position during passivization, the resulting sentence “*A book was given Mary by John.*” is grammatically wrong also. If the passive sentence (157) is translated to English directly, a grammatically incorrect Chinese passive sentence will be produced. The same applies to the translation from English passive sentences with ditransitive verbs to Chinese. As shown in the above a-structures, the difference between English and Chinese ditransitive passive sentences posits another problem in re-using the source language f-structures for target language generation. Unlike the difference between the f-structures for English and Chinese passive sentences discussed at the beginning of this section, the difference between the sentences (142) and (157) is not indicated in the relevant lexical entries for these sentences. This means that, with this kind of sentences, the transfor-

mation from the source f-structure to the required target form cannot be obtained by observing the difference between the relevant lexical entries.

The inadequacy of f-structures for transfer is mainly caused by the relatively low-level linguistic information (i.e. syntactic and functional information) presented in f-structures. The linguistic information captured in traditional f-structures is quite language-dependent, thus traditional f-structures are inadequate to capture the similarity between different languages for facilitating the source-to-target language transfer. It is therefore believed that if f-structures can capture some more language-independent linguistic information about sentences, they can help to improve the lexical selection and target language sentence generation processes. With the recent development of argument structure (a-structure) and the lexical mapping theory in LFG which helps to incorporate a more language-independent linguistic information (i.e. *thematic information*) into f-structures, it is believed that the problem in transferring English passive sentences to Chinese can be alleviated.

6.5 The Transfer from English passive sentences to Chinese

As discussed in Section 6.4, passive sentences in English and in Chinese are different in some ways. These differences pose problems in transferring passive sentences from English to Chinese. The syntactic and functional information captured in a traditional f-structure is insufficient to distinguish the difference between English and Chinese passive sentences. The use of the more language-independent thematic information seems to alleviate this problem.

Huang (1993) suggested that the thematic hierarchy for Chinese is different from that for English and it is this difference which governs the difference between the grammaticality of English passive sentences and their Chinese counterparts. For instance, as mentioned in the previous section, the passive sentence "Mary was given a book by John." is grammatical, whereas its direct word-to-word Chinese translation (161) "瑪莉被約翰送了一本書。" is ungrammatical. It is correct to say the sentence (157) "一本書被約翰送了瑪莉。" in Chinese, but the sentence "A book was given Mary by John." is grammatically incorrect in English. This difference, according to Huang, can be accounted for by the difference between the a-structures for the verbs 'give' and '送'. Consider the lexical mapping between the following sentences and their corresponding a-structures:

(162) English sentence : John gave Mary a book.

Noun Phrases :		*John*	*Mary*	*a book*	
A-structure :	give $<$	agent	*recipient*	*theme*	$>$
Intrinsic :		$[-o]$	$[-r]$	$[+o]$	
Default :		$[-r]$	$[+r]$	$[+r]$	
Syntactic Functions :		SUBJ	OBJ	OBJ_θ	

(163) Chinese sentence : 約翰送了瑪莉一本書。

Noun Phrases :		約翰	瑪莉	一本書	
A-structure :	送 <	agent	*theme*	*recipient*	>
Intrinsic :		$[-o]$	$[-r]$	$[+o]$	
Default :		$[-r]$	$[+r]$	$[+r]$	
Syntactic Functions :		SUBJ	OBJ	OBJ_θ	

The order of thematic roles within a-structures, as shown in (162) and (163), shows the order of the corresponding NPs in the passive sentences. When a sentence is undergoing passivization, the NP which is mapped to the second most prominent thematic role within an a-structure is raised to the subject position of the resulting passive sentence:

(164) English sentence : Mary was given a book by John.

Noun Phrases :			*Mary*	*a book*		*John*	
A-structure :	give <	agent	recipient	theme	> by <	agent	>
Intrinsic :		$[-o]$	$[-r]$	$[+o]$		$[-o]$	
Passive :	be	$\emptyset$					
Default :				$[+r]$		$[+r]$	
Syntactic Functions :			SUBJ	OBJ_θ		OBL_θ	

(165) Chinese sentence : 一本書被約翰送了瑪莉。

Noun Phrases :			一本書	瑪莉		約翰	
A-structure :	送 <	agent	theme	recipient	> 被 <	agent	>
Intrinsic :		$[-o]$	$[-r]$	$[+o]$		$[-o]$	
Passive :	被	$\emptyset$					
Default :				$[+r]$		$[+r]$	
Syntactic Functions :			SUBJ	OBJ_θ		OBL_θ	

The order of thematic roles within an a-structure determines which NP is to be raised during passivization. From the lexical mapping (162), (163), (164) and (165), we can see that:

- the a-structures for the verb 'give' and its equivalent in Chinese, i.e. '送', govern the same thematic roles: agent, recipient and theme; and
- the order of appearance of these thematic roles within each a-structure governs which object is to be raised during passivization.

This suggests that the information encoded in a-structure is sufficient to bridge the gap between passive sentences in English and in Chinese. Hence, the use of a-structure can facilitate the transfer from English passive sentences to Chinese.

As postulated in Chapter 5.2, a-structures can aid the lexical selection process. When transferring passive sentences like (142) "Mary was given a book by John." to Chinese, the thematic roles within the a-structure 'give<agent recipient theme>' are used to select the appropriate a-structure in Chinese, i.e. '送<agent theme recipient>'. The selected a-structure, together with the lexical entries (150) and (160) for the verbs '被'and '送' respectively, can then form the skeleton of the target f-structure in Chinese (see Figure 6.4).

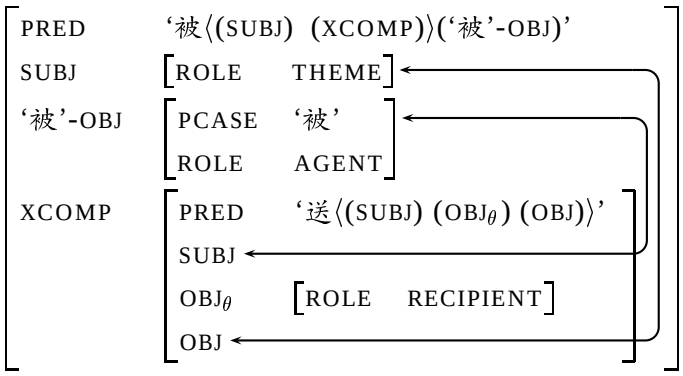


Figure 6.4: Skeleton of Chinese F-structure for "Mary was given a book by John.".

Due to the difference of the syntactic structures between English and Chinese passive sentences, the NP in the subject position of an English f-structure cannot always be mapped onto the subject position of the target Chinese f-structure. In addition to aiding the lexical selection process, the a-structures for the source and target language verbs 'give' and '送' are used to map each NP from the source sentence to the corresponding syntactic function in the target Chinese sentence. A thematic role describes the role played by a participant in an event (cf. Chapter 4.2). This description is universal over all languages and hence the role played by an NP in a sentence in one language is very likely, if not always, to be the same as the role played by the equivalent NP in another language. Therefore thematic roles can be used to link the source NPs and the corresponding syntactic functions of the target sentence. Lexical mapping theory defines how each thematic role within an a-structure is mapped to the corresponding syntactic function in a sentence. During the generation of the required target f-structure, each NP in the source English sentence is transferred to the appropriate syntactic function in the target f-structure based on the thematic role assigned with respect to the lexical mapping theory:

(166)

Source sentence : Mary was given a book by John.

Source Noun Phrases :			*a book*	*Mary*		*John*	
Target A-structure :	送 <	agent	theme	recipient	> 被 <	agent	>
Intrinsic :		$[-o]$	$[-r]$	$[+o]$		$[-o]$	
Passive :	被	$\emptyset$					
Default :				$[+r]$		$[+r]$	
Target Syntactic Functions :			SUBJ	OBJ_θ		OBL_θ	
Target Noun Phrases :			一本書	瑪莉		約翰	

Target sentence : 一本書被約翰送了瑪莉。

After the lexical mapping, each transferred source NP is inserted into the appropriate position in the initial target f-structure in Figure 6.4 and the final f-structure is formed (see Figure 6.5). This f-structure, together with the information from the c-structure of the source English sen-

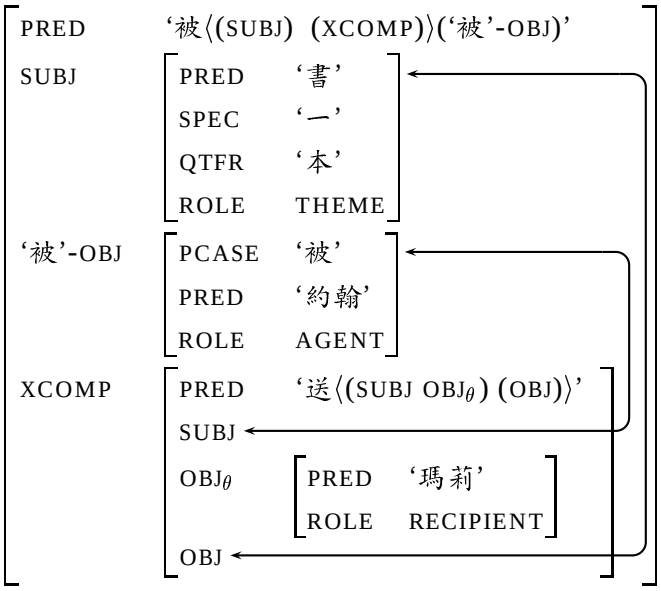


Figure 6.5: The final Chinese F-structure for "Mary was given a book by John.".

tence "Mary was given a book by John.", is then used to generate the required target Chinese passive sentence.

6.6 Discussion

The representation method of f-structure provides a seemingly appropriate medium for source-to-target language transfer. However, the linguistic information provided by traditional f-

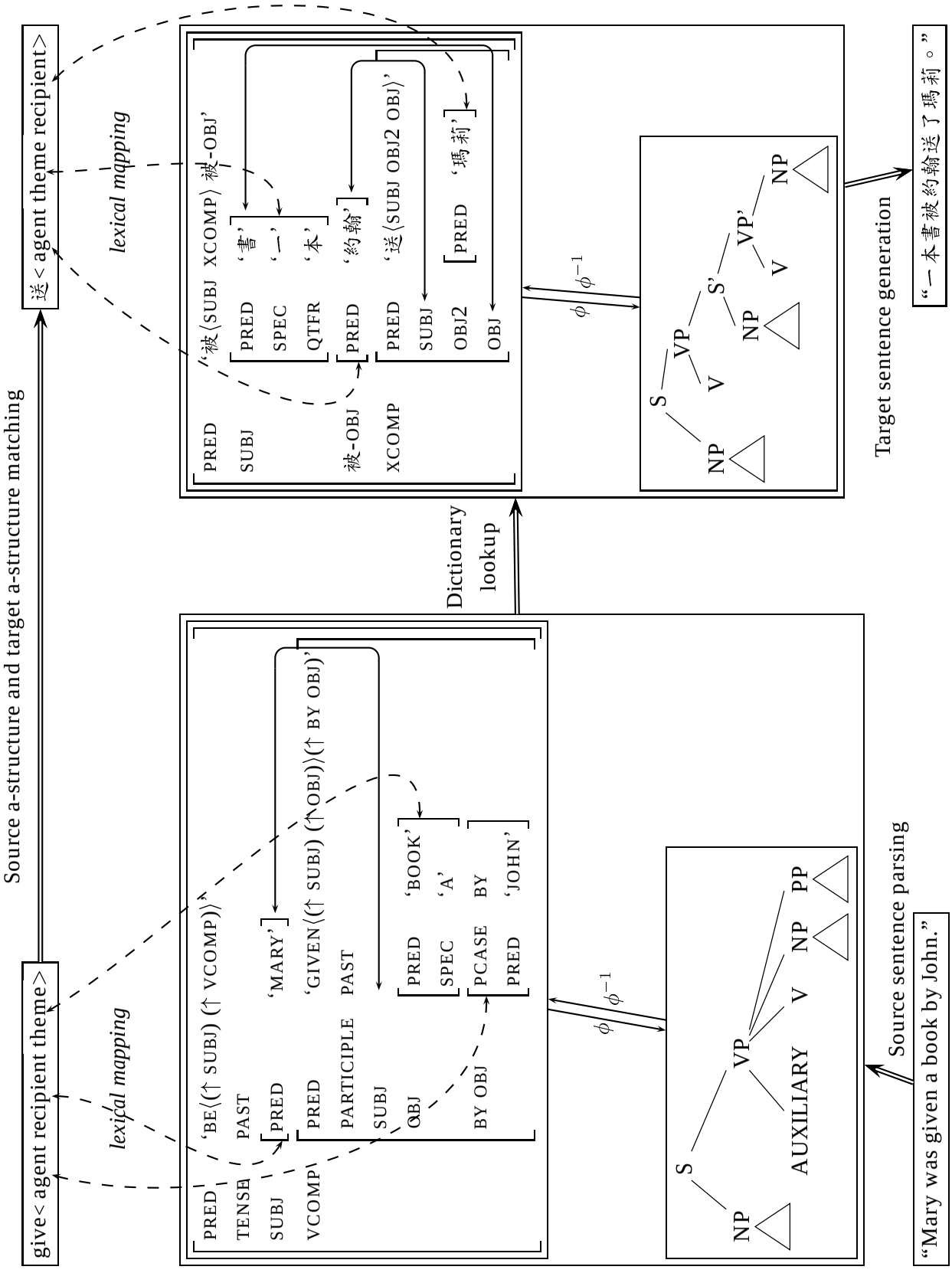


Figure 6.6: Transferring English passive sentence into Chinese using a-structure and lexical mapping theory

structures alone is insufficient for source language word disambiguation. Although f-structures are believed to be relatively language-independent, not all f-structures are universal across languages. Therefore, after lexical transfer, some source language f-structures are required to be transferred to their target language form before they can be used for target language sentence generation. This poses the question about the ability of f-structure to act as a suitable medium for carrying out source-to-target language transfer.

As discussed in Chapter 4, a-structure shows the necessary participants involved in an event in terms of thematic roles and these thematic roles carry some semantic properties which can be used to determine if an object is qualified to play certain role in the event (e.g. an agent must be an animate object and a locative is either a physical object or time). When compared with f-structure which represents the higher syntactic and functional information (cf. Section 3.1.2), a-structure is more language independent because it implicitly carries some kind of real-world knowledge. Thus it is more suitable to bridge the gap between the syntax of source and target languages.

Lexical mapping theory defines the mapping between each thematic role in an a-structure and the corresponding syntactic function in a sentence through matching the $[\pm r]$ and $[\pm o]$ features (cf. Section 4.3). The use of a-structure and lexical mapping theory in transferring sentences from source language to target language means that, unlike Kaplan et al.'s (1989) approach (cf. Section 3.2.2), there is no need to explicitly define equations for capturing all the correspondences between the syntactic functions of the source and target sentences (e.g. $(\tau \uparrow \text{SUBJ}) = \tau(\uparrow \text{SUBJ})$); because these correspondences have already been implicitly captured by the source and target a-structures as well as the lexical mapping theory (cf. Figure 6.6). A syntactic function in the source language sentence would be automatically mapped to the appropriate syntactic function in the target language sentence through the application of a-structure and the lexical mapping theory on both the source language and target language. The transfer process therefore involves:

1. matching source language a-structure to the corresponding target language form,
2. obtaining the required target language lexicon from the bilingual dictionary using the source language lexicon, and
3. carrying out lexical mapping according to the lexical mapping theory to form the required target f-structure.

All these tasks can easily be implemented by means of simple matching and unification. This transfer method can be easily realised in conventional Prolog.

In addition to allowing a systematic transfer from English passive sentences to Chinese, the transfer method described in Figure 6.6 also allows Chinese passive sentences to be translated to English in the same manner. That is to say, the initial f-structure for the required English

passive sentence can be produced from the f-structure of the translating Chinese sentence by the same transfer method.

Due to the fact that a-structure is relatively language independent, the transfer between English and Chinese passive sentences can be a bidirectional process. In the previous section, we looked at how English *passive* sentences are transferred to Chinese through the use of a-structure and lexical mapping theory. Figure 6.6 illustrates this translation process in more detail, i.e.:

1. Parse the source English passive sentence (i.e. "Mary was given a book by John.") according to LFG and obtain the c-structure and f-structure representation of the sentence.
2. From the database of English and Chinese a-structure pairs, select the appropriate a-structure for the source English sentence (i.e. "give<agent recipient theme>").
3. Perform the lexical mapping as defined by the lexical mapping theory. If the selected a-structure is correct, each thematic role within it should be mapped to a syntactic function in the source English sentence.
4. From the dictionary of English and Chinese a-structure pairs, select the appropriate target language a-structure (i.e. "送<agent theme recipient>"). With the information provided by the target language a-structure and the bilingual dictionary, obtain the target language lexical entries from the source language words (e.g. "BE<SUBJ VCOMP>" $\longrightarrow$ "被 <SUBJ XCOMP> 被-OBJ").
5. Establish the translation correspondence between source language syntactic functions and their corresponding target language syntactic functions by performing lexical mapping between the thematic roles in the selected target language a-structure and the syntactic functions of the target language sentence. For instance, the SUBJ of the source English sentence (i.e. Mary) is to be translated as the XCOMP OBJ2[5] of the target sentence.
6. Generate the target language sentence according to the lexical entries for the Chinese lexicon obtained earlier.

This translation process does not require the co-description of bilingual transfer rules. The lexical entries involved are mono-lingual, i.e. they either describe the English lexicon or the Chinese lexicon. The link between English and Chinese in this translation process is established by the English and Chinese a-structures pairs and the bilingual dictionary entries. Both of these bilingual data, when implemented as symmetrical bidirectional dictionaries, allow a bidirectional transfer. This transfer is made possible by the fact that a-structure describes information about the necessary participants of an event. This information is relatively language independent owing to the fact that an event described in one language is very likely to contain the same necessary participants as the same event described in another language. This

[5]In earlier work of LFG, the syntactic function 'OBJ2' is the term used for describing the object$_\theta$ of the sentence (cf. Bresnan & Kanerva 1989, Note 30).

allows a-structures to act as a bridge between English and Chinese. Since the detailed mapping of syntactic functions to the appropriate arguments in the a-structures is taken care of by the lexical mapping on each side (i.e. English and Chinese in Figure 6.6), there is also no need to establish detailed translation correspondences between syntactic functions of source and target languages by means of transfer rules.

The translation process shown in Figure 6.6 is symmetrical by nature. The linguistic information about the source and target languages encoded in this process are independent of each other. This means that the parser and the generator do not require the knowledge of the target language and the source language respectively. Therefore, the source language analysis and the target language generation within the translation process can be made fairly independent of each other. As sentence generation can be viewed as a reverse process of sentence parsing, and vice versa, when both the parser and the generator for this translation process are implemented as bidirectional processes, the resulting system would be able to support bidirectional translation.

6.7 Conclusion

In this chapter, we exemplified one of the problems in using f-structures as the sole medium for the transfer. This problem is due to the fact that f-structure is *not* universal across languages. The kind of information encoded in f-structure (i.e. syntactic information) makes f-structure relatively language dependent and thus it is not suitable to be used as a sole medium for the transfer. If f-structure is used for the transfer, a considerable amount of work is required to define transfer rules for bridging the gap between different languages.

In order to alleviate this problem, a-structure and lexical mapping theory were introduced into the transfer. This chapter showed how a-structure and lexical mapping theory can be applied to effectively improve the transfer between source and target language sentences which have different structures. Unlike f-structure, a-structure contains information which is less language dependent. This makes it more suitable to be used as a medium to bridge the gap between different languages. As a result, when carrying out the transfer, less human effort (e.g. in defining the detailed translation correspondences) is required to bridge the syntactic and lexical gap between the source and target languages.

The transfer method proposed in this chapter is mainly based on simple matching of features and unification. As computers are suited for symbol processing and matching, this method is relatively simple and easy to implement in computing terms. This transfer method also allows the transfer from English passive sentences to Chinese and vise versa. Thus, the resulting translation process can be implemented as a bidirectional process.

Chapter 7

Conclusion and Future Work

In this study, several investigations have been carried out in applying a-structure and the lexical mapping theory to alleviate some of the problems identified in various Machine Translation (MT) processes: *source language sentence parsing, source-to-target language transfer and target language sentence generation*. The problems being dealt with in this study can be roughly divided into two categories: linguistic ambiguity and the differences between source and target languages. Chapter 5 and 6 discussed the results of the investigations carried out in this study.

Chapter 6 reported on another investigation carried out in this study which looked at the ability of a-structure and the lexical mapping theory in aiding the transfer of English sentences to Chinese. One of the problems observed in transferring English sentences to Chinese occurs in the transfer between English ditransitive sentences and their Chinese counterparts. Owing to the differences in the ways of expression and the syntax between English and Chinese, the transfer between ditransitive sentences in English and Chinese is not straightforward and a means to bridge the gap between source and target sentences is required in order to facilitate the transfer (cf. Chapter 6). It was observed that a-structure and the lexical mapping theory are a good means to alleviate this problem. In addition to helping an MT system to transfer sentences from one language to another, as illustrated in Chapter 5, a-structure is also capable of alleviating the problem caused by some kinds of lexical and structural disambiguation, especially the disambiguation involving different compositions of verbs and prepositions.

7.1 Problems in Using A-structure and Lexical Mapping Theory in MT

In this approach, a-structures play a crucial role in selecting the most appropriate target language lexical items and in carrying out lexical mapping. If there is no matching target language a-structure for a source language a-structure of a predicator and/or an a-structure is inappropriately established in an MT system, the system is likely to fail in selecting the appropriate

target language equivalent and it will also fail to generate the appropriate target language f-structure for target language sentence generation. This section further discusses the potential problems observed in applying a-structure and lexical mapping theory to aid MT processing.

7.1.1 No Matching Source-and-Target Language A-structures

The approach to lexical and structural disambiguation proposed in this thesis exploits the ability of a-structure to describe linguistic information which is relatively less language dependent. Recall that the motivation behind this approach is that the participants took part in an event described in one language are very likely to be the same as the same event described in another language. This assumption has been proven by studying the a-structures of the same event described in English and Chinese (cf. Section 5.2). Therefore, if a-structure is 'universal' across languages, i.e. the same event described in different languages should subcategorise the same a-structure arguments (i.e. thematic roles), it will undoubtedly be able to act as a link between source language verbs and their target language equivalents. Bresnan (1994) suggested that the LFG formalism, which comprises of c-structure, f-structure and a-structure, is capable of acting as an alternative architecture of Universal Grammar. There is no doubt that the representation of thematic information in a-structure is applicable to different natural languages. However, there is doubt as to whether a-structure is universal across languages.

It is observed that there is a small portion of verbs in both English and Chinese which *do* have equivalents in the other language, but have no matching a-structures. For instance, as was mentioned in Section 5.4, the Chinese ditransitive verb '通知' subcategorises three arguments and it bears the a-structure '通知<agent theme recipient>', as in:

(167)

約翰	通知	警察	一	件	劫案。
John	notify	police	one	QUANTIFIER	robbery.

but its English counterpart '*notify*' subcategorises only two arguments[1] and it bears the a-structure 'notify<agent theme>' (as in "John notified the police."). As a-structure forms a link between lexical semantics and syntactic structure, to a certain extent a-structure is syntax dependent. This makes a-structures incapable of relating some source language verbs and their target language equivalents.

Due to the absence of immediate matching a-structures between the verbs 'notify' and '通知', when translating a sentence like "John notified the police." to Chinese, the two-argument a-structure for 'notify' would not match the three-argument a-structure for '通知'. Thus, the transfer which is based on source and target a-structure matching would fail. Owing to this problem, one might argue that the proposed transfer method is not worth pursuing because if

[1]Though with the introduction of the preposition 'of', the verb 'notify' can take three arguments (as in "John notified the police of the robbery."), unlike the verb 'give' or 'tell' which requires the subcategorisation of three arguments, it is not necessary for the verb 'notify' to bear the oblique argument 'of-OBJ'.

the sentence "John notified the police." is translated to Chinese in a word-for-word manner, the resulting Chinese sentence:

(168) 約翰 通知 了 警察。
John notify ASPECT MARKER police.

is adequate to convey the meaning of the original English sentence. Though, to a certain extent, the sentence (168) is understandable to average Chinese speakers, many of them would consider this sentence as incomplete or ungrammatical. It is similar to the utterance "John gave a book." which sounded incomplete to most English speakers. It is debatable whether a good MT system should allow the production of an incomplete or ungrammatical sentence when the resulting sentence can convey the meaning of the source sentence. If the answer is yes, this means that the a-structure '通知<agent recipient>' is acceptable to the grammar. This two-argument a-structure can therefore be introduced to the database of a-structure entries. The sentence "John notified the police." would no longer cause any problem to the transfer. However, if the required MT system refuses the production of any ungrammatical target language sentence, the sentence "John notified the police." would simply not be translated.

As mentioned earlier, the verb 'notify' can bear the oblique function 'of-OBJ' to form a sentence like "John notified the police of the robbery.". When translating sentences involving 'notify ...of' to Chinese and/or vise versa, as discussed in Section 5.4, the selection of the appropriate target language translation relies on the one-to-one correspondence between 'notify of' and '通知'. As the set of thematic roles appearing in the a-structures 'notify<agent theme> of<instrument>' and '通知<agent theme recipient>' are different (e.g. the *theme* of the sentence "John notified the police of the robbery" is in fact the *recipient* of the Chinese sentence (167)), it is not possible to transfer syntactic functions in a source sentence to the appropriate position in the target sentence based on the matching of thematic roles. The mismatch between these thematic roles suggests that the use of a-structures for source-to-target verb transfer might not be applicable in all cases. With verbs without matching a-structure arguments, the transfer would have to rely on the use of translation correspondence defined in the MT system.

Another pair of English and Chinese verbs which poses similar problem while using a-structures for MT is 'envy' and its Chinese counterpart '妒忌'. In English, one can say:

(169) John envied Mary.

(170) John envied Mary's success.

(171) John envied Mary her success.

However, in Chinese, one can say:

(172) 約翰 妒忌 瑪莉。
John envy Mary

(173) 約翰 妒忌 瑪莉的 成功。
John envy Mary's success.

However, it is ungrammatical to say:

(174) 約翰 妒忌 瑪莉 她的 成功。
John envy Mary her success.

While the use of a-structure and lexical mapping theory supports a fairly straight-forward translation between the sentences (169) and (172) and the sentences (170) and (173), the translation of the English sentence (171) to Chinese would fail. As in the case of the sentence "John notified the police", this translation problem is due to the non-existence of target language equivalent rather than the introduction of a-structure into the transfer.

As there is no immediate Chinese equivalent for the sentence (171) "John envied Mary her success.", one possible translation which would preserve the writing style and the meaning of this sentence is:

(175) 約翰 妒忌 瑪莉。 約翰 妒忌 她的 成功。
John envy Mary. John envy her success.
John envied Mary. John envied her success.

In order to facilitate this translation, a correspondence is required to be drawn between the three-argument a-structure for 'envy' and the corresponding Chinese a-structures.

7.1.2 Difficulty in Establishing Appropriate A-structures

Another potential problem in using a-structures for aiding the source-to-target language transfer is that it is difficult to establish appropriate a-structures. If an inappropriate a-structure is assigned to a verb in one language, the chance for it to fail in matching with the appropriate a-structure in another language would be very high. Though there is a wide range of literature written about the formation of argument structures and the assignment of thematic roles to noun phrases in sentences, e.g. Jackendoff (1972), Givón (1984), Foley (1984) and Dowty (1991), the guidelines given in this literature still appear to be inadequate to govern the formation of a-structures for some verbs. The English ditransitive verb 'envy', as in (171) "John envied Mary her success.", is a good example of such a verb.

It is arguable as to what the a-structure of the English ditransitive verb 'envy' is. The indirect object of a ditransitive verb in English which acts as a conscious participant of the event is often marked as a '*beneficiary*' or a '*recipient*'. However, the participant described by the NP '*Mary*' in (171) does not behave like a 'beneficiary' or 'recipient'. In (171), the NP 'Mary' is the '*possessor*' of 'her success'. One might then propose the a-structure for the sentence "*John envied Mary her success.*" to be: 'envy<experiencer possessor theme>', where the NP 'John' is the *experiencer*

and the NP 'her success' is the *theme*. However, the thematic role 'possessor' is not defined in the thematic hierarchy. This definition creates a problem in the lexical mapping because during the lexical mapping, the order of thematic roles presented in an a-structure affects the assignment of thematic roles to the syntactic functions in a sentence. It has not yet been proven that the thematic role '*possessor*' is relevant to the thematic hierarchy or that the hierarchy:

"$\cdots$ experiencer $>$ possessor $>$ theme $\cdots$"

is universal for all verbs in a language. As a result, the a-structure 'envy<experiencer possessor theme>' does not seem to be an appropriate a-structure for the ditransitive verb 'envy'[2].

The sentence (171) appears to be a union of both (169) and (170). When considering the types of participants involved in the event described by (169) "John envied Mary.", it is clear that the a-structure for (169) is 'envy<experiencer patient>' because John is the one who experienced the emotion 'envious' towards Mary and Mary is the one who displays the locus of the effect of this emotion. However, when considering the role of each participant in the event described by (170), it is not clear whether the NP '*Mary's success*' should be considered as the patient or the theme of the event.

One interesting point worth noting is that with verbs of emotion, e.g. the verbs 'love', 'like' and 'hate', it is appropriate to say:

- John loves this job.
- John likes many jobs.
- John hates failures.
- John hates that job.

in which abstract patients of the events (e.g. 'job' and 'failures') are not required to bear an ownership like '*his* job' or '*their* failures'. However, with the verb 'envy', though it is under the category of *verb of emotion*, it sounds odd to utter:

- * John envied success.
- * John envied this fortune.
- * John envied that skill.

If the matter that causes someone to be envious is abstract, the resulting sentence would sound more appropriate when the ownership of this matter is specified in the sentence, e.g.:

- John envied Mary's success.
- John envied Tom's fortune.

[2]Some of the ideas of the discussion about the a-structure for the English ditransitive verb 'envy' flowed from a discussion with Alex Alsina conducted via electronic mail.

- John envied the skill that Jill has.

This observation shows that although the verb 'envy' is a verb of emotion, it has different linguistic behaviour from usual verbs of emotion. Verbs of emotion often subcategorise the thematic roles '<experiencer patient>' and the patient can be an abstract object (e.g. 'failure') or an animate object (e.g. 'cats'). However, with the sentence (170), instead of behaving like a patient which "*displays the locus of the effect*", the NP '*Mary's success*' appears to be expressing the argument "*of which location or state is predicated*" (cf. Section 4.1.4). Thus, it is more appropriate to describe the event structure of the sentence (170) by the a-structure 'envy<experiencer theme>'.

With the a-structure assignments:

1. 'envy<experiencer patient>' $\Longrightarrow$ *John envied Mary.*
2. 'envy<experiencer theme>' $\Longrightarrow$ *John envied Mary's success.*

and the similarity between the sentence (171) and the union of (169) and (170), it seems appropriate to derive the a-structure for (171) as:

(176) envy<experiencer patient theme>

This a-structure would not cause any problem to the lexical mapping with the syntactic functions of (171):

(177) Sentence : John envied Mary her success.

A-structure :	*envy* <	experiencer	patient	theme	>
Intrinsic :		$[-o]$	$[-r]$	$[+o]$	
Default :		$[-r]$		$[+r]$	
Syntactic Functions :		SUBJ	OBJ	OBJ_{th}	
NPs :		John	Mary	her success	

The a-structure (176) also has the advantage of capturing the similarity between the transitive forms and the ditransitive form of 'envy'. The problem with transferring the English sentence (171) to Chinese discussed in the previous section can also be alleviated through the use of this a-structure:

1. The a-structure correspondence between the sentences (171) and (175) is:

   envy<experiencer patient theme> $\Longrightarrow$
   妒忌<experiencer patient> and 妒忌 <experiencer theme>

2. The experiencer of (171) 'John' will be translated to '約翰' and form the SUBJ of (175). Similarly, the patient (i.e. 'Mary') and theme (i.e. 'her success') of (171) will form the OBJ of the sentences in (175) and be translated to '瑪莉' and '她的成功' respectively.

Although the a-structure (176) seems to describe the event structure of the sentence (171) appropriately, it is not common that a sentence subcategorises both a patient and a theme. Some linguists might find this construct too unusual and it is not certain whether this construct would cause problems elsewhere. Even though the a-structure (176) is suitable for the MT task described, thus making it appropriate for the required purpose[3], a considerable amount of study and analysis was required to derive this a-structure. In this approach where a-structure is used as a medium for carrying out the transfer, with verbs which do not have an obvious and easy-to-define a-structure, the applicability of this approach would be diminished. However, it is believed that with continuous effort by both theoretical and computational linguists in generalising the the linguistics of natural languages, the problem encountered in this approach would be alleviated.

7.2 What makes this investigation successful?

The results of this investigation showed that some of the results in the formalisation of natural languages obtained by theoretical linguists can, to a certain extent, be readily applied to MT processing. The generalisation of different linguistic behaviours across languages by theoretical linguists, as examined in this study based on a relatively new extension of LFG (i.e. a-structure and lexical mapping theory), allows computational linguists to spend less effort on observing and describing the differences between natural languages. As a result, the time and effort required to develop a linguistic-based MT system is lessened. However, as discussed in Chapter 4, some modifications and fine-tunings of the results of theoretical linguistic studies are required so as to facilitate their application in the domain of computational linguistics.

This thesis described an alternative approach to applying LFG in MT through the use of a-structure and the lexical mapping theory. This approach performs MT based on the linguistic information captured in the c-structure, f-structure and a-structure of sentences as defined in the LFG formalism. Although traditional c-structures and f-structures together provide a seemingly good medium for carrying out the source-to-target language transfer, the syntactic and functional information about sentences captured in these structures does not provide sufficient information for handling the disambiguation of source language words during the transfer. With the use of lexical mapping theory and the semantic features possessed by each thematic role[4], thematic information encoded in a-structures can be incorporated into the traditional f-structures. This improves the capability of f-structures to provide additional linguistic information for improving the lexical selection process during the transfer. A-structure forms a link between the source and target language event structures. This helps to bridge the gap between source and target language sentences due to their syntactic difference, thus facilitating the translation of sentences from one language to another.

[3] cf. Section 4.1
[4] cf. Section 4.1

The natural languages that this study was based on are English and Chinese: with English as the source language and Chinese as the target language. Though this study was conducted based on the different linguistic behaviours of the English and Chinese languages, it is believed that the method described in this thesis is not restricted to aiding the translation between English and Chinese sentences only. The main aim of this study is to investigate the ability of a-structure and lexical mapping theory to aid disambiguation in MT processing. The results obtained, as exemplified in Chapters 5 and 6, showed that this method is capable of capturing some information about the real-world through describing the participants involved in real-world events described by natural language sentences. This characteristic of a-structure allows the proposed method to aid lexical and structural disambiguations in MT processing. Unlike conventional methods to MT which rely on observing and describing the differences in the syntax of different languages (e.g. Kaplan et al. (1989) and Her et al. (1994)), the disambiguation is made successful by the fact that the information captured in a-structure is relatively language independent and this information formed an implicit link between different languages. As the disambiguation is not purely syntactic-oriented, this approach is believed to be suitable for aiding the translation process of arbitrary language pairs.

Many linguistic-based approaches to MT rely on the use of some kinds of equations to capture the correspondence between each pair of source and target lexical units for the transfer, e.g. Whitelock's (1992) Shake-and-Bake approach. Unlike these approaches, the use of a-structure and lexical mapping theory for MT does not require the introduction of additional equivalence equations to bridge the gap between the source and target languages during the transfer. This is made possible by the fact that lexical mapping theory helps the formation of target language sentence structure by defining what syntactic function each thematic role should be mapped with (cf. Chapter 6). Each syntactic function in the source language sentence is related to its corresponding syntactic function in the target language sentence implicitly through the source and target language a-structures matching and the lexical mapping between a-structure arguments and syntactic functions. The gap between the source and target languages is therefore narrowed by the matching a-structures in these languages. While developing the required MT system, there is no need to go through a lengthy and complicated process in defining translation correspondences like '$(\tau \uparrow \text{SUBJ}) = \tau(\uparrow \text{SUBJ})$' (cf. Section 3.2.2). As a result, the required system development process is relatively simpler. Furthermore, less human effort and time, especially in terms of observing the linguistics of the relevant natural languages and preparing the grammar, is required during the system development process.

7.3 Future Work

This thesis shows how a-structure and lexical mapping theory can be used to facilitate the transfer of sentences which do not have a matching target language equivalent. However, the ability of this method to alleviate other problems in MT has not yet been explored.

7.3.1 Disambiguating nouns

A-structure and lexical mapping theory is capable of aiding the disambiguation and lexical selection of verbs and prepositions. Can the same disambiguation method be used to aid the disambiguation of other kinds of homographs, e.g. nouns? As discussed in Section 4.1, each thematic role carries some kind of semantic properties. These semantic properties are capable of helping the disambiguation of verb and preposition pairs. As illustrated in Section 1.1.4, the introduction of semantic markers to an MT system helps the MT system to 'understand' the underlying meaning of a word. Through verifying the meaning of each word in a sentence, an MT system can select the most appropriate word sense of each word which is used in the sentence. Therefore, it seems plausible that the thematic information encoded in a-structure can help an MT system to disambiguate the meaning of other kinds of homographs.

Consider the meanings of the English noun 'ball'. The noun 'ball' often means "a round or roundish body, either solid or hollow, of a size and composition suitable for any of various games" (Makins 1994), as in:

(178) John kicked a *ball* to Mary.

However, this noun can also mean "a social function for dancing" (Makins 1994), as in:

(179) John went to a *ball* with Mary.

When translating sentences that contain this kind of nouns, some way to distinguish the meaning used in each case is required. When considering the a-structure of the above sentences involving the noun 'ball', we can see that the NP 'a ball' in each case is assigned with different thematic roles, i.e. patient and locative respectively. The event described in the sentence (178) involves hitting a physical object, thus the meaning of the noun 'ball' in this sentence cannot be "a social function for dancing", which is an abstract object. Similarly, the event described by the sentence (179) refers to travelling to a location. However, the word sense "a round or roundish body, either solid or hollow, of a size and composition suitable for any of various games" cannot be a location for someone to travel to. Therefore, the noun 'ball' in 179 would be more likely to mean "a social function for dancing", which can serve as an abstract location.

Note that the use of a-structure to aid the lexical disambiguation of nouns has an advantage over the use of simple semantic markers. Previous attempts in using simple semantic markers to aid lexical disambiguation (e.g. Wilks 1976) tended to introduce semantic markers in an *ad hoc* manner, i.e. to derive the set of semantic markers manually according to the lexicon in the system. However, a-structure offers a systematic method to capture the right amount of thematic information of a sentence. Each thematic role appearing in an a-structure carries some semantic properties and it can also be mapped with the syntactic structure of a sentence systematically according to the lexical mapping theory. Unlike thematic roles in a-structure, there is no formal mechanism to link semantic markers with the syntactic information of sen-

tences. Therefore, when compared with the use of simple semantic markers alone for lexical disambiguation, the use of a-structure seems to be more convenient and effective.

The disambiguation of the noun 'ball' with the use of thematic information in a-structure seems to suggest that a-structure can aid the disambiguation of nouns. However, without a detailed study on the ability of this method to aid this kind of disambiguation, it is difficult to conclude whether a-structure would allow an effective solution to the disambiguation of nouns.

7.3.2 Automatic extraction of a-structures from a corpus

Another possible extension to this research project is to develop a method to extract a-structures from a corpus automatically. In the proposed approach to MT, a-structure forms part of the lexicon for aiding the analysis, transfer and generation of sentences. To establish all a-structures required by an MT system manually can be a laborious task. Therefore, if a-structures can be extracted automatically from a corpus, the time and labour required to implement the proposed method to MT would be reduced. Nowadays, corpora which are prepared in electronic forms are highly available. Many researchers have been studying the extraction of different kinds of information from a corpus. It is believed that given a set of guidelines for establishing a-structures from sentences (e.g. the guidelines illustrated in Section 4.1), it is possible to extract a-structures automatically from a corpus.

7.3.3 Reducing the processing time

This thesis illustrated that thematic information captured in a-structures is suitable for solving some of the problems arising in MT. However, due to the limited time available for this research project, the question as to how to implement the proposed method to MT efficiently has not yet been answered.

MT has a well-known reputation for consuming a large amount of processing time. This is mainly due to the fact that MT is mostly implemented using symbolic processing. In the processing techniques currently available, any mismatch, either in grammar rules or words, in any part of the processing will force the system to undo some of the previous operations and redo them with other options. When any problem of ambiguity is encountered during MT processing, a large amount of undoing and redoing on part of the MT process is required.

The reason that MT is such a lengthy computing task is partially due to the limited facilities offered by the conventional programming languages like C++ and Prolog. Amongst them, Prolog is better in handling a large amount of structured data and carrying out pattern matching by unification and backtracking. The dictionaries required for MT can be stored systematically

in a form of Prolog structures[5]. Prolog supports a depth-first search with backtracking facility which is designed based on the generate and test algorithm (Hentenryck 1989). It is this unplanned generate-and-test behaviour of Prolog which leads to a large amount of processing time being wasted during the system backtracking[6].

The approach to MT proposed in this thesis is based on symbolic matching. The fact that the facilities offered by conventional programming languages are inadequate to perform symbolic matching efficiently makes the proposed method to MT a potentially lengthy computing task, and thus affects the practicality of this method. In the early stage of this research study, the ability of the Constraint Logic Programming (CLP) paradigm to improve the matching method provided by conventional Prolog was investigated. However, due to the limited time available, this investigation was ceased after the feasibility study. The result obtained from the feasibility study showed that CLP has the potential to alleviate the problem created by the unplanned generate-and-test behaviour of Prolog. It is believed that by using the CLP techniques to implement the MT method discussed in this thesis (i.e. the use of a-structure and lexical mapping theory for MT), the resulting MT system would be more efficient.

7.4 Conclusion

This investigation studied the ability of a-structure and lexical mapping theory to aid various MT tasks. From the results obtained, it is concluded that this relatively new extension of the LFG formalism, when used in MT processing, enhances the capability of LFG to act as the linguistic backbone of a linguistic-based MT system. The proposed approach to MT allows a relatively straight-forward and simple way to implement MT systems and to perform machine translation. In addition to reporting how and to what extent a-structure and lexical mapping theory alleviate some of the problems in MT, this study also demonstrated that, with slight adaptation (e.g. the adaptation suggested in Section 4.2), the results of linguistic generalisation obtained by theoretical linguists can be easily and effectively applied to the development of practical MT systems.

[5] A Prolog structure is a kind of storage medium for storing links between entities, e.g. words from certain language and some grammatical notions. This is more often used to represent database in Prolog as it is easy to access during Prolog run-time. Examples of Prolog structures are:

```
e_noun(cat, singular).
eg_noun(cat, 'Katze', singular).
```

[6] Although Prolog allows the programming of a careful planned search for reducing the amount of backtracking involved during the system processing, as described by (Hentenryck 1989), not all kinds of systems enable an easy incorporation of a planned search.

Bibliography

Alsina, A. (1996*a*), Resultatives: a joint operation of semantic and syntactic structures, *in* M. Butt & T. H. King, eds, 'Proceedings of the LFG '96 conference, Grenoble, 26–28 August 1996', [Online]. Available: http://csli-publications.stanford.edu/LFG/1/lfg1.html [2000, January 6].

Alsina, A. (1996*b*), *The role of argument structure in grammar: evidence from romance*, Center for the Study of Language and Information, Stanford, CA, CSLI lecture notes series; 62.

Alsina, A. & Mchombo, S. A. (1993), Object asymmetries and the Chicheŵa applicative construction, *in* S. A. Mchombo, ed., 'Theoretical aspects of Bantu grammar', Center for the Study of Language and Information, Stanford, CA, CSLI lecture notes series; 38, pp. 17–45.

Ananiadou, E. (1990), The use of sublanguages in machine translation, *in* 'Proceedings of a workshop on machine translation, UMIST, Manchester, 2–3 July 1990', Speech and Language Technology Club, Department of Trade and Industry, London. Unpaged.

Anderson, J. M. (1994*a*), Case, *in* R. E. Asher, ed., 'The encyclopedia of language and linguistics', Vol. 1, Pergamon Press, Oxford, pp. 447–453.

Anderson, J. M. (1994*b*), Case grammar, *in* R. E. Asher, ed., 'The encyclopedia of language and linguistics', Vol. 1, Pergamon Press, Oxford, pp. 453–464.

Arnold, D., Balkan, L., Humphreys, R. L., Meijer, S. & Sadler, L. (1994), *Machine translation: an introductory guide*, Blackwells/NCC, London.

Bresnan, J. (1982*a*), The passive in lexical theory, *in* J. Bresnan, ed., 'The mental representation of grammatical relations', MIT Press, Cambridge, MA, pp. 3–86.

Bresnan, J. (1994), 'Locative inversion and universal grammar', *Language* **70**(1), 72–131.

Bresnan, J. (1995), Lexicality and argument structure, *in* 'The Paris syntax and semantics conference, 12–14 October 1995', [Online]. Available: http://www-lfg.stanford.edu/lfg/archive/archive.html [2000, January 6].

Bresnan, J., ed. (1982*b*), *The mental representation of grammatical relations*, MIT Press, Cambridge, MA.

Bresnan, J. & Kanerva, J. M. (1989), 'Locative inversion in Chicheŵa: a case study of factorization in grammar', *Linguistic Inquiry* **20**(1), 1–50. Also appeared in: Stowell T. and Wehrli E. , eds (1992) *Syntax and the lexicon*. Academic Press, San Diego. Syntax and semantics 26, pp. 53-101.

Bresnan, J. & Zaenen, A. (1990), Deep unaccusativity in LFG, *in* K. Dziwirek, P. Farrell & E. Meijas-Bikandi, eds, 'Grammatical relations: a cross-theoretical perspective', Center for the Study of Language and Information, Stanford, CA, pp. 45–57.

Bruce, B. & Moser, M. G. (1992), Grammar, case, *in* S. C. Shapiro, ed., 'Encyclopedia of artificial intelligence', 2nd edn, John Wiley, Chichester, pp. 563–570.

Carlson, G. N. (1984), 'Thematic roles and their role in semantic interpretation', *Linguistics* **22**(3), 259–279.

Chang, C. H.-H. (1991), 'Thematic structure and verb copying in Mandarin Chinese', *Language Sciences* **13**(3–4), 399–419.

Comrie, B. (1977), In defense of spontaneous demotion: the impersonal passive, *in* P. Cole & J. M. Sadock, eds, 'Grammatical relations', Academic Press, London, Syntax and semantics; 8, pp. 47–58.

Cooper, R. P. (1994), Head-driven Phrase Structure Grammar, *in* R. E. Asher, ed., 'The encyclopedia of language and linguistics', Vol. 3, Pergamon Press, Oxford, pp. 1532–1535.

Dowty, D. (1991), 'Thematic proto-roles and argument selection', *Language* **67**(3), 547–619.

EAGLES (1996), *EAGLES preliminary recommendations on subcategorisation*, [Online]. Available: www.ilc.pi.cnr.it/EAGLES96/synlex/synlex.html [2000, January 6].

Fass, D. & Pustejovsky, J. (1992), Lexical decomposition, *in* 'Encyclopedia of artificial intelligence', 2nd edn, John Wiley, Chichester, pp. 806–812.

Fillmore, C. J. (1968), The case for case, *in* E. Bach & R. T. Harms, eds, 'Universals in linguistic theory', Holt, Rinehart and Winston, New York, pp. 1–88.

Fillmore, C. J. (1977), The case for case reopened, *in* P. Cole & J. M. Sadock, eds, 'Grammatical relations', Academic Press, London, Syntax and semantics; 8, pp. 59–81.

Foley, W. A. (1984), *Functional syntax and universal grammar*, Cambridge University Press, Cambridge.

Gazdar, G., Klein, E. H., Pullum, G. K. & Sag, I. A. (1985), *Generalized Phrase Structure Grammar*, Harvard University Press, Cambridge, MA.

Gazdar, G. & Mellish, C. (1989), *Natural language processing in Prolog — an introduction to computational linguistics*, Addison-Wesley, Wokingham, chapter 7, pp. 230–237.

Givón, T. (1984), *Syntax: a functional-typological introduction*, Vol. 1, John Benjamins, Amsterdam.

Halvorsen, P.-K. (1988), Situation semantics and semantic interpretation in constraint-based grammars, *in* 'Proceedings of the International Conference on Fifth Generation Computer Systems (FGCS '88), Tokyo, 28 November – 2 December 1988', Institute for New Generation Computer Technology, Tokyo, pp. 471–478. Also appeared in: Dalrymple, M., Kaplan, R. M., Maxwell, J. T. & Zaenen, A., eds (1995), *Formal issues in Lexical-Functional Grammar*, Center for the Study of Language and Information, Stanford, CA, pp. 293–309.

Halvorsen, P.-K. & Kaplan, R. M. (1988), Projections and semantic description in LFG, *in* 'Proceedings of the International Conference on Fifth Generation Computer Systems (FGCS '88), Tokyo, 28 November – 2 December 1988', Institute for New Generation Computer Technology, Tokyo, pp. 1116–1122. Also appeared in: Dalrymple, M., Kaplan, R. M., Maxwell, J. T. & Zaenen, A., eds (1995), *Formal issues in Lexical-Functional Grammar*, Center for the Study of Language and Information, Stanford, CA, pp. 293–309.

Hentenryck, P. V. (1989), *Constraint satisfaction in logic programming*, MIT Press, Cambridge, MA.

Her, O.-S. (1989), 'An LFG account for Chinese bei sentences', *Journal of the Chinese Language Teachers Association* **23**(3), 67–89.

Her, O.-S., Higinbotham, D. & Pentheroudakis, J. (1994), Lexical and idiomatic transfer in machine translation: An LFG approach, *in* 'Research in Humanities Computing', Vol. 3, Oxford University Press, Oxford, pp. 200–216.

Horby, A. S., ed. (1984), *Oxford advanced learner's English-Chinese dictionary*, Oxford University Press; Keys Publishing, Hong Kong.

Huang, C.-R. (1993), 'Mandarin Chinese and the lexical mapping theory — a study of the interaction of morphology and argument changing', *The Bulletin of the Institute of History and Philology* **62**(2), 337–388.

Hutchins, W. J. & Somers, H. L. (1992), *An introduction to machine translation*, Academic Press, London.

Ingria, R., Boguraev, B. & Pustejovsky, J. (1992), Dictionary / lexicon, *in* 'Encyclopedia of artificial intelligence', 2nd edn, John Wiley, Chichester, pp. 341–365.

Jackendoff, R. (1972), *Semantic interpretation in generative grammar*, MIT Press, Cambridge, MA.

Kaplan, R. & Bresnan, J. (1982), Lexical-Functional Grammar: a formal system of representation, *in* J. Bresnan, ed., 'The mental representation of grammatical relations', MIT Press, Cambridge, MA, pp. 173–281.

Kaplan, R. M. (1989), The formal architecture of Lexical-Functional Grammar, *in* C. R. Huang & K. J. Chen, eds, 'Proceedings of the Republic of China Computational Linguistics Conference (ROCLING II), Taipei, 1989', Academia Sinica, Taipei, pp. 1–18. Also appeared in: Dalrymple, M., Kaplan, R. M., Maxwell, J. T. & Zaenen, A., eds (1995), *Formal issues in Lexical-Functional Grammar*, Center for the Study of Language and Information, Stanford, CA, pp. 7–27.

Kaplan, R. M. & Maxwell, J. T. (1988), An algorithm for functional uncertainty, *in* 'Proceedings of the 12th International Conference on Computational Linguistics (COLING '88), Budapest, 22–27 August 1988', John von Neumann Society for Computing Sciences, Budapest, pp. 297–302.

Kaplan, R. M., Netter, K., Wedekind, J. & Zaenen, A. (1989), Translation by structural correspondences, *in* 'Proceedings of the 4th Conference of the European Chapter of the Association for Computational Linguistics, UMIST, Manchester, 10–12 April 1989', Association for Computational Linguistics, New Brunswick, NJ, pp. 272–281. Also appeared in: Dalrymple, M., Kaplan, R. M., Maxwell, J. T. & Zaenen, A., eds (1995), *Formal issues in Lexical-Functional Grammar*, Center for the Study of Language and Information, Stanford, CA, pp. 311–319.

Kay, M. (1973), 'Automatic translation of natural languages', *Daedalus* **102**(3), 217–230.

Kay, M. (1984), Functional Unification Grammar: a formalism for machine translation, *in* 'Proceedings of the 10th International Conference on Computational Linguistics (COLING '84), Stanford, CA, 2–6 July 1984', Stanford University, Stanford, CA, pp. 75–78.

King, M. (1982), Eurotra: an attempt to achieve multilingual MT, *in* V. Lawson, ed., 'Practical experience of machine translation: proceedings of a conference, London, 5–6 November 1981', North-Holland, Amsterdam, pp. 139–147.

Kudo, I. & Nomura, H. (1986), Lexical-Functional transfer: a transfer framework in a machine translation system based on LFG, *in* 'Proceedings of the 11th International Conference on Computational Linguistics (COLING '86), Bonn, 25–29 August 1986', University of Bonn, Bonn, pp. 112–114.

LFG research group in CSLI (1995), *Lexical-Functional Grammar: exploring the facets of linguistics knowledge*, Center for the Study of Language and Information, Stanford, CA. A short online introduction to the research on LFG in CSLI. No longer available.

Li, D. & Cheng, M. (1994), *A practical Chinese grammar for foreigners*, Sinolingua, Beijiang.

Lyons, J. (1968), *Introduction to theoretical linguistics*, Cambridge University Press, Cambridge.

Lyons, J. (1977), *Chomsky*, Harvester Press, Hassocks.

Makins, M., ed. (1994), *Collins English dictionary*, 3rd edn, Harper Collins, London.

Malmkjoer, K. (1991), Case grammar, *in* K. Malmkjoer, ed., 'The linguistics encyclopedia', Routledge, London, pp. 65–70.

Neidle, C. (1994), Lexical Functional Grammar (LFG), *in* R. E. Asher, ed., 'The encyclopedia of language and linguistics', Vol. 3, Pergamon Press, Oxford, pp. 2147–2153.

Nyberg, E. H. I. & Mitamura, T. (1992), The KANT System: fast, accurate, high-quality translation in practical domains, *in* 'Proceedings of the 14th International Conference on Computational Linguistics (COLING '92), Nantes, 23–28 August 1992', University of Nantes, Nantes, pp. 1254–1258.

Palmer, M. & Wu, Z. (1995), 'Verb semantics for English-Chinese translation', *Machine translation* **10**(1–2), 59–92.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985), *A comprehensive grammar of the English language*, Longman, London.

Rappaport, M. & Levin, B. (1988), What to do with θ-roles, *in* W. Wilkins, ed., 'Thematic relations', Academic Press, London, Syntax and semantics; 21, pp. 7–36.

Rohrer, C. (1986), Linguistic bases for machine translation, *in* 'Proceedings of the 11th International Conference on Computational Linguistics (COLING '86), Bonn, 25–29 August 1986', University of Bonn, Bonn, pp. 353–355.

Sadler, L. (1990), Codescription and transfer, *in* 'Proceedings of a workshop on machine translation, UMIST, Manchester, 2–3 July 1990', Speech and Language Technology Club, Department of Trade and Industry, London. Unpaged.

Sadler, L. & Arnold, D. (1992), 'Unification and machine translation', *Meta* **37**(4), 657–680.

Sells, P. (1985*a*), *Lectures on contemporary syntactic theories: an introduction to Government-Binding Theory, Generalized Phrase Structure Grammar, and Lexical-Functional Grammar*, Center for the Study of Language and Information, Stanford, CA, CSLI lecture notes series; 3, chapter 4, pp. 135–191.

Sells, P. (1985*b*), *Lectures on contemporary syntactic theories: an introduction to Government-Binding Theory, Generalized Phrase Structure Grammar, and Lexical-Functional Grammar*, Center for the Study of Language and Information, Stanford, CA, CSLI lecture notes series; 3, chapter 2, pp. 35–38.

Sinclair, J., ed. (1987), *Collins COBUILD English Language dictionary*, Harper Collins, London.

Sinclair, J., ed. (1989), *Collins COBUILD dictionary of phrasal verbs*, Harper Collins, London.

Sinclair, J., ed. (1990), *Collins COBUILD English grammar*, Harper Collins, London.

Somers, H. (1990), Current research in machine translation, *in* 'Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language, Austin, Texas, 11–13 June 1990', Lingusitics Research Center, University of Texas at Austin, Austin, pp. 1–12. Also appeared in: Proceedings of a workshop on machine translation, UMIST, Manchester, 2–3 July 1990. Speech and Language Technology Club, Department of Trade and Industry, London. Unpaged.

Tan, F. (1987), The predicate argument structure of bei, *in* 'Proceedings of the 13th Annual Meeting of the Berkeley Linguistics Society, Berkeley, CA, 14–16 February 1987', Berkeley Linguistics Society, Berkeley, CA, pp. 285–295.

Tanenhaus, M. K. & Carlson, G. N. (1989), Lexical structure and language comprehension, *in* W. Marslen-Wilson, ed., 'Lexical representation and process', MIT Press, Cambridge, MA, chapter 18, pp. 529–561.

Trask, R. L. (1993), *A dictionary of grammatical terms in linguistics*, Routledge, London.

Warren, H., ed. (1994), *Oxford learner's dictionary of English idioms*, Oxford University Press, Oxford.

Whitelock, P. (1992), Shake-and-bake translation, *in* 'Proceedings of the 14th International Conference on Computational Linguistics (COLING '92), Nantes, 23–28 August 1992', University of Nantes, Nantes, pp. 784–790.

Whitelock, P. & Kilby, K. (1995), *Linguistic and computational techniques in machine translation system design*, 2nd edn, UCL Press, London, chapter 10, pp. 147–170.

Wilks, Y. (1976), Parsing English II, *in* E. Charniak & Y. Wilks, eds, 'Computational semantics: an introduction to artificial intelligence and natural language comprehension', North-Holland, Amsterdam, pp. 155–184.

Wong, S. H. S. & Hancox, P. (1998*a*), An investigation into the use of argument structure and lexical mapping theory for machine translation, *in* 'Proceedings of the 12th Pacific Asia Conference on Language Information and Computing (PACLIC 12), Singapore, 18–20 February 1998', Chinese and Oriental Languages Information Processing Society, Singapore, pp. 334–339.

Wong, S. H. S. & Hancox, P. (1998*b*), Using a-structure and lexical mapping theory in LFG for machine translation, *in* 'Pre-proceedings of the 9th Artificial Intelligence / Cognitive Science Conference (AICS '98), Dublin, 19–21 August 1998', University College Dublin, Dublin, pp. 55–63.

Wong, S. H. S. & Hancox, P. (1999), What is the lexical form of '*bei*'?, *in* 'Proceedings of the 13th Pacific Asia Conference on Language Information and Computing (PACLIC 13), Taipei, 10–11 February 1999', National Cheng Kung University, Tainan, pp. 169–176.

Wood, M. M. (1993), *Categorial Grammars*, Routledge, London.

Yang, J. & Gerber, L. (1996), Chinese-English machine translation system, *in* 'International Conference on Chinese Computing '96 (ICCC'96) —- the latest technological advancement & application, Singapore, 4–7 June 1996', Chinese and Oriental Languages Information Processing Society. Also appeared in: [Online]. Available: http://www.systransoft.com/Papers/ppr_cess.htm [2000 January 6].