UNIVERSITÉ DE GENÈVE

Archive ouverte UNIGE
https://archive-ouverte.unige.ch

Thèse 2013

Open Access

This version of the publication is provided by the author(s) and made available in accordance with the copyright holder(s).

Structural Asymmetries in Machine Translation: The case of English-Japanese

Kauffmann, Alexis Joseph Azar

How to cite

KAUFFMANN, Alexis Joseph Azar. Structural Asymmetries in Machine Translation: The case of English-Japanese. Doctoral Thesis, 2013. doi: 10.13097/archive-ouverte/unige:34540

This publication URL: https://archive-ouverte.unige.ch/unige:34540
Publication DOI: 10.13097/archive-ouverte/unige:34540

© This document is protected by copyright. Please refer to copyright holder(s) for terms of use.

Université de Genève
Faculté des lettres

Structural Asymmetries in Machine Translation: The case of English-Japanese

Thèse présentée à la Faculté des lettres de l'Université de Genève par

Alexis Kauffmann

pour l'obtention du grade de Docteur ès lettres

Membres du jury:
Directeur de thèse: Eric Wehrli, Université de Genève
Président du jury: Jacques Moeschler, Université de Genève
Timothy Baldwin, Université de Melbourne
Christian Boitet, Université Joseph Fourier (Grenoble 1)
Sadao Kurohashi, Université de Kyoto

28 juin 2013

Acknowledgements

After spending about 5 years dedicated to this thesis, I realise that many people have contributed to it in one way or the other, and I feel very grateful to them.

First of all, I would like to thank greatly my supervisor Prof. Eric Wehrli for his teaching, his help and his great support, and for giving me the opportunity to work in an excellent atmosphere where motivation has always been given and pressure has never been put. I would also like to thank him for giving me motivation not only for my research, but also for using my bicycle and practising several kinds of sports, with a special thought for the afternoon when we went ice skating in *Parc des Bastions* with other lab members.

I would like to thank Prof. Sadao Kurohashi a lot for his supervision, his good and clear advice and for receiving me so well in Kyoto University. There also, I have enjoyed the fact that motivation was not only given for research, but also for sport practise, and I would like to remember more particularly the day when we went windsurfing on Lake Biwa with other lab members.

I would also like to thank Luka Nerima for his supervision and help on the questions related to lexical databases and his good advice on many things, Prof. Daisuke Kawahara for his supervision and great help for my research throughout my stay at Kyoto University, and Christopher Laenzlinger for his advice and for answering all the little linguistic questions I have had throughout the years.

I would like to thank the members of the jury, especially Prof. Jacques Moeschler for reading one of the preliminary draft versions of this work, and Prof. Christian Boitet for answering my questions about MT history and for his advice.

Of course, I would also like to thank a lot Prof. Izumi Tahara for correcting my Japanese and giving me her opinion on Japanese linguistic questions, and Frank Horton for correcting my English.

I would also like to thank all my colleagues and former colleagues at LATL and CLCL and at Kurohashi-Kawahara lab, especially Asheesh Gulati that has been working with me on experiments about statistical post-editing, and all those who helped me either when I needed data or human assessment, when I had technical questions or when I needed proofreading: Tanja Samarzic, Lonneke

Van der Plas, Toshiaki Nakazawa, Nobuhito Tamaki and Hiroshi Manabe. As lab life is not only made of research and teaching, I would like to express my gratitude to Eva Capitao for all the cakes and biscuits she has brought and made for us, Jean-Philippe Goldman for his unforgettable fondue events, Andrea Gesmundo for his funny jokes and wonderful guitar playing, and of course Yves Scherrer for all the lunches and the fun times we have had together.

I also would like to remember all the glorious snowy times spent at Swiss, French and Japanese mountains with lab members (especially professors and senior researchers), where almost all of them were skiing while I was riding my snowboard.

Last but not least, I would like to thank for their support my always encouraging parents and family, my beloved wife and my dear friends, with a special thought to the ones who completed like me their DEUG or licence in *Université Joseph Fourier* in Grenoble and have then defended or started working on their PhD thesis.

Contents

List of Figures

List of Tables

List of Acronyms

- **AP** Adjective Phrase
- **ATN** Augmented Transition Networks
- **ATR** Advanced Telecommunication Research institute international
- **AdvP** Adverb Phrase
- **BLEU** BiLingual Evaluation Understudy
- **CETA** Centre d'Etudes sur la Traduction Automatique
- **CFG** Context-Free Grammar
- **CJKI** Chinese Japanese Korean Institute
- **CP** Complementizer Phrase
- **ConjP** Conjunction Phrase
- **DLT** Distributed Language Translation
- **DP** Determiner Phrase
- **EBMT** Example-Based Machine Translation
- **EDR** Electronic Dictionary
- **FP** Functional Projection
- **GAT** Georgetown Automatic Translation project
- **GETA** Groupe d'Etude pour la Traduction Automatique
- **GG** Generative Grammar
- **HMT** Hierarchical Machine Translation
- **HTER** Human-targeted Translation Error Rate

- **KNP** Kurohashi-Nagao Parser
- **LATL** Laboratoire d'Analyse et de Technologie du Langage
- **LBMT** Linguistics-Based Machine Translation
- **MIT** Massachusetts Institute of Technology
- **MMT** Multi-lingual Machine Translation
- **MT** Machine Translation
- **NEC** Nippon Electric Company
- **NLP** Natural Language Processing
- **NP** Noun Phrase
- **NTT** Nippon Telegraph and Telephone
- **PP** Preposition (or Postposition) Phrase
- **PT** Personal Translator
- **RBMT** Rule-Based Machine Translation
- **S** Sentence
- **SBSMT** Syntax-Based Statistical Machine Translation
- **SMT** Statistical Machine Translation
- **SOV** Subject-Object-Verb
- **SQL** Structured Query Language
- **SVO** Subject-Verb-Object
- **TP** Tense Phrase
- **TWIC** Translation of Words In Context
- **UNL** Universal Networking Language
- **VP** Verb Phrase

Chapter 1

Introduction

1.1 Machine translation

The main topic of this thesis is text machine translation (MT). The principle of text MT is to translate automatically text from a *source* language to a *target* language. One may argue that computers cannot be good translators because they cannot translate texts as humans can. Even if this statement may be true, state-of-the-art non-specialised text machine translation is nowadays practically useful for two main purposes:

- translating a text fast, even if the generated translation is not completely correct; the generated translation can help the reader understand the source text, especially if he or she cannot read the text in its source language; this is particularly true with the use of MT systems available on the internet;
- helping the human translator in the translation task.

The quality and accuracy of machine translation outputs depend on the MT system used and on the type of source text: some MT systems are more effective for the translation of technical documents, while others also handle well texts that contain everyday language. The choice of the language pair and translation direction[1] is another significant parameter. Some MT systems are conceived for the translation of only one language pair, while others are multilingual systems. The quality of the translation produced by multilingual systems depends on the language pair and translation direction. These variations in translation quality

[1] A pair of languages can give rise to translation in two different directions. For example, English and Japanese can give rise to Japanese-to-English or English-to-Japanese translation. The different directions will imply different problematics, especially in regard to the number of underspecified parameters (see Bond (2005)) to overcome in the source language.

are due to both practical and linguistic reasons. A typical practical cause is the unequal amount of lexical data available for each language pair. Another practical cause is the unequal level of implementation of the MT systems depending on language pairs.

1.2 Structural asymmetries in machine translation

The main linguistic cause for the variations of translation quality depending on language pairs is the level of structural and lexical similarity between the two languages of the language pairs: while closely related languages may share either closely related vocabulary or quite similar syntactic structures and morphological variations, structurally unrelated languages need a more advanced treatment of structural differences and asymmetries.

In this thesis, we will focus on the treatment of structural differences (see Hutchins and Somers (1992), p.103) in the translation process, and more particularly on the treatment of asymmetrical cases, at a lexico-syntactic level. Structural differences are often related to word and component order. In the following example, the word order of the English verb phrase "goes to school" is exactly reversed in the Japanese translation:

(1.1) Keiji goes to school.

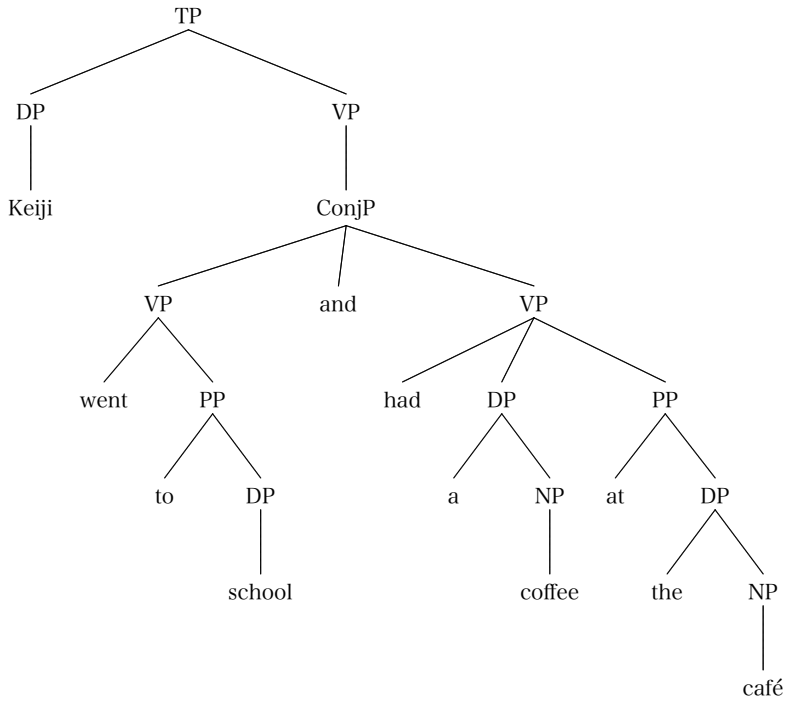
圭司	は	学校	に	行きます。
keiji	ha	gakkou	ni	ikimasu
keiji	[topic/subject]	school	to	goes

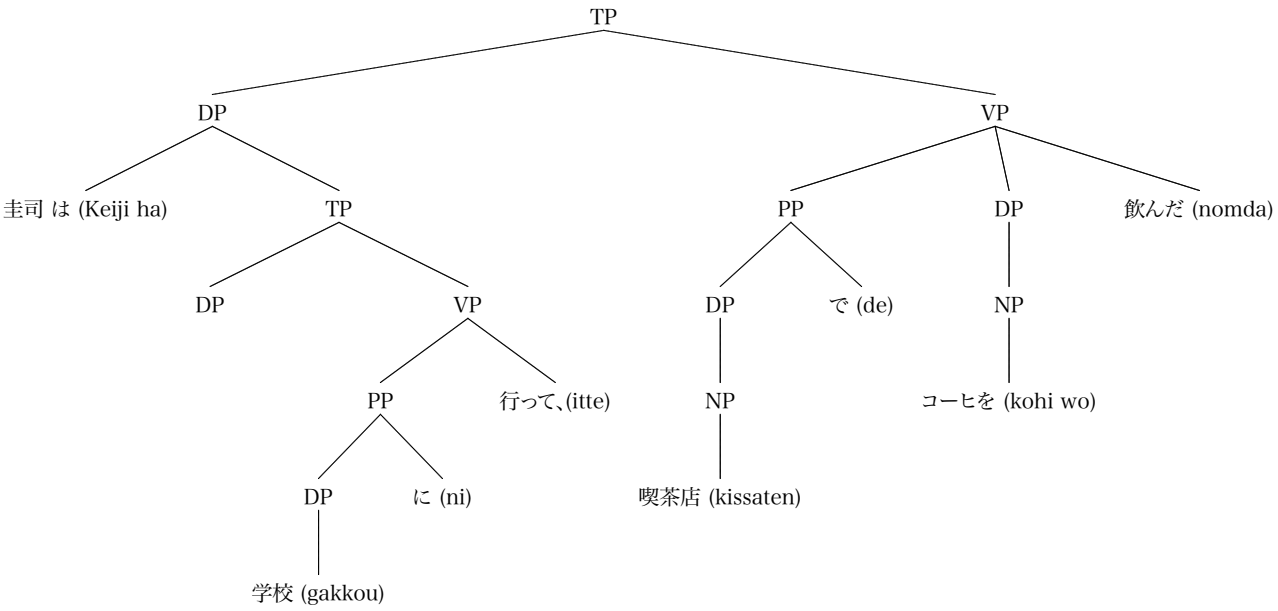
Structural asymmetries (also described as *translation asymmetries* in (Pause, 1997), as *machine translation divergences* in (Dorr, 1994), *systematic divergences* in (Brockett et al., 2002), and as a non-trivial type of *structural differences* in Hutchins and Somers (1992)) happen in cases where a simple reordering of word and constituents would not be sufficient to generate a correct and fluent output in the target language. For example, the following English sentence contains the conjunction *and* between two verbal clauses. The Japanese translation does not contain such a conjunction, but requires the verb of the first clause to be in gerundive form, preferably followed by a Japanese comma:

(1.2) Keiji went to school and drank a coffee at the café.

圭司	は	学校	に	行って、	喫茶店	で	コーヒ	を
keiji	ha	gakkou	ni	itte,	kisaten	de	kôhi	wo
keiji	[nominative]	school	to	go(gerund),	café	at	coffee	[object]

飲んだ。
nomda.
drank





The Japanese translation, that does not contain a coordination conjunction, has a syntactic structure which is different from the English source sentence: instead of finding two verbal phrases (VP) governed by a conjunction in a conjunctive phrase (ConjP), the first VP is a gerundive one, affixed to the subject and ended with a comma, and the last VP is the main one of the sentence. It is an example of structural asymmetry in translation.

1.3 English-Japanese MT

We will focus on English-to-Japanese MT. Japanese on the one hand, and English on the other hand represent two different language families: Western languages[2] and Far-East Altaic languages[3].

Western and Far-East Altaic languages are structurally different at several levels, such as word and component order, morphological inflexion and politeness levels. The typical word order in English is Subject-Verb-Object (SVO), whereas it is Subject-Object-Verb (SOV) in Japanese. Moreover, Japanese verbal clauses do not necessarily have their subjects or objects expressed (Saint-Jacques, 1966) and definite or indefinite articles do not exist in Japanese (Bond, 2005), which often causes a lack of information in Japanese-to-English MT. On

[2]We consider here West Germanic languages (such as English or Scots), Romance languages (such as Italian, Spanish or French), and, to a lesser extent, other Germanic or Scandinavian languages.

[3]Japanese and Korean.

the other hand, the Japanese verbal inflexion gives much more importance to the politeness level than the English does.

English-to-Japanese MT, even if it deals with less underspecified input than Japanese-to-English MT, requires a good handling of structural language differences in word and component order (as in example 1.1). Moreover, it requires a special attention for asymmetrical cases, where the structures employed in the Japanese sentence are not the same as the ones found in the English one (as in example 1.2). Hence, Japanese translations of English sentences might contain adjectival clauses, nominalised clauses, or specific politeness or modality expressions that would not be found in the source sentence. The different types of English-Japanese structural asymmetries will be discussed in Chapter 3.2.

English-to-Japanese MT has a long and rich history (see Hutchins (1986); Wilks (2009)), closely related to Japanese-to English MT and the global research on natural language processing (NLP) in Japan. Consequently, many lexical resources, MT and NLP related tools are available for English-to-Japanese MT. In comparison, we'll also mention French-to-Japanese and Japanese-to-French MT, a language pair for which only a handful of applications have been specifically developed.

1.4 An approach based on linguistic knowledge

Since its beginnings, machine translation has evolved into several methods (Hutchins (1986), Boitet et al. (2009), Lee et al. (2010)). Some of these methods, such as *rule-based machine translation* (RBMT) are based on human knowledge of languages and usually rely on a syntactic and/or semantic analysis of the source sentence. We will refer to these as *linguistics-based machine-translation* (LBMT[4]). During the last three decades, other methods based on machine learning such as *statistical machine translation* (SMT) or *example-based machine translation* (EBMT) have considerably progressed. We will refer to these as *data-driven approaches*.

Basic statistical methods and basic rule-based methods have often showed disappointing results on the translation of structurally different, non-related languages. The first ones tend to fail to handle well the source sentence grammar, especially long-distance relationships and proper reordering, while the latter ones tend to generate sentences with a lack of fluency and a structure which is too similar to the source sentence structure.

Recently, much research has been led to include syntactic representations and improve reordering in statistical or data-driven MT (Lee et al., 2010). Other

[4]These approaches are often referred to as *rule-based methods* in reference to *rule-based machine translation* (RBMT). We will prefer here the appellations *linguistics-based machine-translation* (LBMT) or *handcrafted approaches* (see Boitet et al. (2009)) that cover a wider scope of techniques.

research results have been obtained by the addition of statistical components to handcrafted approach systems, thus leading to the creation of *hybrid* systems (Thurmair, 2009). We have chosen here to do linguistics-based machine translation (LBMT), with a multilingual system designed in a classical procedural transfer-based architecture, using deep syntactic tree structures for input and output sentence representations. The classical LBMT paradigm will be enhanced with linguistic data acquired by statistical methods.

On the basis of linguistic knowledge and large-scale lexical data partly obtained by statistical techniques, we will take advantage of deep syntactic LBMT to try to solve problems inherent in English-to-Japanese translation such as local and long-distance reordering, focusing on structural asymmetries at the lexico-syntactic level, and especially on verbal argument structure translation, Japanese adjectival clause generation, and modality translation.

Hence, our approach will be close the ones followed by Nagao and Tsujii (1986) with the Mu MT system, Kaji (1987) with the Hitachi Hicats MT system, Amano et al. (1989) with the Toshiba Transac MT system, and by Kinoshita et al. (1992). The novelty of our approach, compared to the ones mentioned here, will be the use of statistically acquired data for the treatment of structural asymmetries.

1.5 Application

Its-2 is a multilingual LBMT system, with a classical transfer-based architecture, using large-scale electronic lexicons. We will give more details about this system and its architecture in Chapter 2.3.3.

As a practical application, we will describe throughout the chapters the implementation of a basic prototype of English-Japanese version of Its-2. This is the first time that Its-2 is used for the generation of a Far East Altaic language. Hence, the experiments will give an opportunity to test its potential adaptability. However, developing a full-scale working, wide-covering English-Japanese LBMT system will remain out of the scope of the thesis. Such implementations on a large scale have been done already, notably by and by Fujitsu with the Atlas-2 system[5].

The different steps of the implementation will consist in preparing large-scale electronic lexicons for the system; preparing theoretical transfer rules[6] and implementing them, going from basic simple sentences to more complex sentences

[5]The Atlas-2 system (Fujitsu (2012)) had about 586'000 lexical entries in 2001, 1'000'000 in its 2003 version, 1'500'000 in its 2004 version, 5'500'000 in its 2007 version, and even more in its current version, Atlas V14.

[6]A *transfer rule* is a rule that explains how to translate a linguistic component of the source sentence into another (possibly empty) component in the target sentence. It is applied during the *transfer phase* of the translation process.

and complex verbal structures; and using statistically acquired linguistic data to improve the translation quality, enhancing both the lexicons and the MT system itself.

Manual tests and automatic evaluation of the outputs will be done at several stages of the implementation. Some comparisons with other existing MT systems will also be carried out. The tests and evaluation should give us an overview of the system ability to handle structural asymmetries and the effects of using statistically acquired linguistic data.

1.6 Research questions

On the basis of our linguistic study and the results of the tests and evaluations of our experiments, we will try in this research to consider the following questions:

- Can a multilingual linguistics-based MT system such as Its-2, that has been usually used for MT between Western languages, adapt well for translation into a Far-East Altaic language like Japanese?
- Can a procedural linguistics-based approach without interlingual representation overcome syntactic structure differences and structural asymmetries?
- Can statistically obtained data help a linguistics-based MT system to handle structural asymmetries?

1.7 Overview of chapters

The thesis is organised as follows:

Chapter 2 is an introductory chapter. The first section, that can be skipped by readers familiar with Japanese, contains a description of fundamental notions about the Japanese writing system and language. Then, the second section discusses the state of the art of English-Japanese, French-Japanese and multilingual MT and describes the architecture of the Its-2 MT system that is used in this thesis.

Chapter 3 is about translation asymmetries at the lexical level. First, it describes the different types of English-Japanese translation asymmetries, discussing cases described in (Dorr, 1994) classification and other possible cases, and mentioning the consequences on MT lexicons. Then, it describes, as an application, the development of large-scale monolingual and bilingual lexicons for the English-Japanese version of Its-2. Finally, a method is proposed for bilingual verb subcategorisation detection using monolingual subcategorisations

and basic bilingual verb correspondences. This method is applied for the improvement of Its-2's bilingual English-Japanese lexicon, using Japanese verb subcategorisation data that had been automatically acquired from the web.

Chapter 4 describes basic word and component reordering and selection for English-to-Japanese simple sentence translation. This step is a required preliminary step before the translation of longer and more complex sentences.

Chapter 5 focuses on the treatment of structural asymmetries, with the example of the generation of Japanese adjectival sentences in English-to-Japanese and French-to-Japanese MT. It first gives a linguistic description of the phenomenon, then exposes a proposed solution and in the end carries out a comparative evaluation on a selection of appropriate sentences.

Chapter 6 deals with the translation of complex sentences in English-to-Japanese MT. It describes possible cases of clause reordering and analyses the structural differences between English and Japanese. A statistical classification of Japanese conjunctive words launched on web-extracted text is used to improve lexical classification and consequently achieve a more accurate clause translation and reordering. A small evaluation is discussed at the end of the chapter.

Chapter 7 focuses on the treatment of modality and complex verbal structures in English-to-Japanese MT. Transfer rules for modality are deduced from a statistical analysis of modality translation in aligned Japanese-English parallel corpora. Verb subcategorisation data mentioned in Chapter 3 are used for a better translation of complex verbal structures. Finally, an automatic evaluation is launched and mentions improvements obtained with the applications of the methods described in Chapter 6 and 7, as well as the remaining drawbacks of Its-2 for English-to-Japanese translation.

Finally, conclusion and future directions are discussed in **Chapter 8**.

Chapter 2

Overview of the Japanese writing system and related research

2.1 Introduction

In this chapter, we will explain some basic notions about Japanese writing system.

Then, we will describe some research related to the thesis. First, we will give a brief summary of MT history. Then we will describe the state of the art in English-Japanese, Japanese-English and multilingual MT. Then, we will mention research related to French-Japanese and Japanese-French MT. Finally, we will describe the Its-2 MT system, that has been used for the thesis practical application.

2.2 Overview of the Japanese writing system

We will explain here some basic notions about Japanese writing: about the direction of writing, the characters, punctuation and word segmentation. The reading of this section is useful for the reader who is not familiar with Japanese or with its writing system.

Part of the material contained in this section was already described by the author in (Kauffmann, 2008b).

2.2.1 Direction of writing

Japanese can be written either the traditional way: vertically from right to left, or the Western way: horizontally from left to right. The traditional way of writing is used mostly for books and newspapers[1], whereas the Western way is more used for usual, official, scientific and technical documents. Signs in the streets are written either way.

2.2.2 Writing system

The Japanese writing system uses three character sets: 漢字 (Kanji), ひらがな (Hiragana) and カタカナ (Katakana)[2].

Kanji

Kanji are ideograms. They almost all come from Chinese. In English, they are also called *Chinese characters* and in Japanese, the word 漢字 (Kanji) literally means *Han characters*, which reminds us of their Chinese origin. Each kanji has one or several meanings and one or several pronunciations. The pronunciation depends on the word where the character appears. There are two types of pronunciation: 音読み ("onyomi": Chinese reading) which is derived from the pronunciation of the character in Chinese, and 訓読み ("kunyomi": Japanese reading) which is derived from traditional Japanese pronunciation.

There are 2136 officially recognised characters: the Jôyô Kanji[3]. 983 other characters are also allowed in proper names: the Jinmeiyô Kanji.

Thousands of other characters exist but have not been included in the set of Jôyô and Jinmeiyô Kanji. A graduate student in science is expected to know about 3000 Kanji, and a graduate student in language and literature is expected to know about 5000 Kanji. The famous Nelson dictionary contains about 5000 Kanji. Computer systems typically contain 5000 Kanji. The total amount of Kanji since their origin is about 80000, and most of them are traditional or archaic ones.

Hiragana

Hiragana (see Table 2.1) is a Japanese syllabary, which consists of about 50 symbols (46 normal symbols + 4 small ones + 2 diacritics). These symbols come from calligraphic forms of Kanji with the same pronunciation. They were first used mainly by court ladies, who usually did not have the access to Kanji education that men had.

[1]Both writing directions can be used in newspapers, and often appear in the same page.

[2]The Latin alphabet is also sometimes used in Japanese, especially for acronyms such as *CD*. Numbers are usually written in arabic numerals and sometimes in numeral Kanji.

[3]The number of Jôyô Kanji used be of 1945 characters and has been increased to 2136 in 2010.

They are now taught in school to the children who use them to write words when they do not know kanji yet. They are used in texts to write verbal and adjectival endings, grammar words such as determiners, conjunctions and particles. Some nouns, adjective and verbs are fully written in Hiragana.

They are used in Kanji dictionaries to write the *kunyomi* of the Kanji.

Figure 2.1: *Hiragana*

Katakana

Katakana (Table 2.2) is another way to represent the same syllabary (+ the long vowel symbol ー). They also come from Kanji with the same pronunciation and were created by Buddhist monks.

These characters are now mainly used to write foreign nouns or Japanese nouns that have a foreign origin. They are also sometimes used on commercial signs, replacing difficult kanji, and in manga to insist emphatically on some words or to represent onomatopoeias.

They are used in Kanji dictionaries to write the *onyomi* of the Kanji.

Figure 2.2: *Katakana*

2.2.3 Japanese punctuation

Japanese has some special punctuation symbols, such as the Japanese comma: 、 and the Japanese dot: 。("maru"), the Japanese quotation marks: 「」 and double quotation mark: 『』 .

The question mark ？ and exclamation mark ！ are used too, though only some colloquial interrogative sentences need a question mark instead of a usual Japanese dot.

2.2.4 Japanese words

The Japanese text is divided in 文 ("bun": sentences), that end with a Japanese dot : 。. There are no spaces between Japanese words. Japanese traditional grammar considers that a sentence is composed by 文節 ("bunsetsu": phrases), that are themselves divided in 単語 ("tango": words). A bunsetsu is formed by a 詞 ("shi": independent word), often followed by one or several 辞 ("ji": non-

independent words), such as particles or copula[4] (see Nakamura-Delloye (2003)).

As there are no spaces between words in a Japanese text, readers are helped in understanding word segmentation by seeing changings of character sets. For example, most nouns and verb radicals are written in kanji, whereas hiragana are used mainly for terminations or grammar words, and katakana words are mainly nouns.

2.2.5 Romanisation

A transcription of Japanese into the Latin alphabet exists: it is called *romanisation*. Several romanisation norms have been defined, such as the *Hepburn*, *Kunrei-shiki* and *Nihon-shiki*. Usually, when Japanese sentences are transcribed to Latin characters, following the romanisation process, bunsetsus are not taken into account and every word is separated from the next one with a space.

In this thesis, the examples of Japanese sentences will be displayed as in example 2.1: the first line is the Japanese sentence, written in Japanese characters; the second line is the same sentence with its words transcribed in Latin characters following the Hepburn romanisation norm; the third line gives the translation of the sentence words into English; and the fourth line is the full sentence translation in English.

(2.1)

これ	は	最初の	文	です。
kore	ha	saisho no	bun	desu
This	[topic/subject]	first	sentence	it is

This is the first sentence

In the examples of translation of an English sentence into Japanese, the presentation will be as in example 2.2: the source sentence comes first, and is then followed by the Japanese translation, the romanised transcription and the word translation[5].

(2.2) This is the 2nd sentence.

これ	は	二番目の	文	です。
kore	ha	niban me no	bun	desu
This	[topic/subject]	2nd	sentence	it is

[4]The copula だ ("da") is the main Japanese auxiliary. It is often equivalent to the English *to be*. A more detailed description of its properties is given in section A.0.3.

[5]The Japanese translation will sometimes be followed by a *literal translation of the translation*, which will be slightly different from the original source sentence.

2.3 Related research

In this section, we will describe related research, focusing mainly on MT. Other related research fields such as syntactic parsing and other NLP tasks, Japanese and English grammar or lexical databases will be mentioned in following chapters. The field of translation with computers is divided into machine translation (MT) or machine-assisted translation (MAT), text or speech translation and specialised or general-use translation. Here, our study is mainly confined to general-use text MT systems.

First, we will give a brief summary of MT history, with a particular interest for the work achieved in Japan. Then, we will describe the state of the art in English-Japanese, Japanese-English and multilingual MT, distinguishing commercial and research systems and classifying systems by their architectures. We will also mention research related to French-Japanese and Japanese-French MT. Finally, we will describe in detail the Its-2 MT system and its architecture.

2.3.1 English-Japanese, Japanese-English and multilingual MT

Brief overview of MT history

In 1933, Petr Petrovich Smirnov-Troyanskii issued in Soviet Union a patent that described the idea of a multilingual text translation machine. He is considered by MT historian W.J. Hutchins as "genuine precursor of machine translation" (Hutchins (1986), p.22). Unfortunately, computers were not available at that time for Troyanskii to help him realise entirely his project.

The first experiments on MT with computers took place in the USA in the late forties and represented the beginning of the MT research field (see Hutchins (1986), p.27-28).

Then, after a demonstration of Russian-English MT called "the Georgetown-IBM demonstration" given in 1954[6], the global interest in MT increased and this field began to be studied by researchers around the world (Hutchins (1986), p.25-37): between 1954 and 1956, other researches on MT started in the USA, USSR, United Kingdom, Canada and Japan; between1957 and 1966, researches started in Eastern Europe country (especially Czechoslovakia and East Germany), other Western Europe countries (especially France, West Germany and Italy), Hong Kong, China and Mexico (Hutchins (1986), p.61-149).

In Japan, research on MT started in 1956. Soon, the first English-Japanese and Japanese-English experimental systems were developed. The Yamato system was designed for English-to-Japanese translation of sentences from school

[6]demonstration which was the result of the Georgetown Automatic Translation project (GAT), which had started in 1951.

textbooks. It performed a basic syntactic analysis of the English sentence, before reordering in the transfer phase. However, its domain was limited to simple or simplified sentences, hence reducing the amount of vocabulary and syntactic structures to handle. It was designed and enhanced from 1960 to 1970. Yamada worked in 1964 on another experimental system for Japanese-to-English translation, which was the first one to include a Japanese segmentation and syntactic parsing component (Hutchins (1986), p.146-147). In the seventies, Nagao worked at Kyoto University on English-to-Japanese MT of scientific paper titles, with the Titran system (Hutchins (1986), p.317-319).

In 1964, Toma started to implement the Systran Russian-English MT system in West Germany, which would be used in 1970 by the US Air Force and in 1974 by the National Aeronautic and Space Administration. The Logos system, created by Scott, was developed since 1965, sponsored by the US Air Force, for English-Vietnamese and Vietnamese-English MT. In the seventies, these two systems started to be used for the translation of other language pairs, becoming multilingual commercial systems.

From 1971 to 1988, the Ariane MT system (which was later renamed Ariane 78, Ariane 85 and then Ariane G5) was developed and improved by the GETA team in France at Grenoble University (Vauquois and Boitet, 1985). Its most developed version is a Russian-to-French MT system, and it has also been used for French-English MT and other language pairs (or specialised domains), at various levels of development. Ariane and its transfer structures, that combine multilevel representations of the sentence, have had a strong and influential impact in Europe as well as Japan and other Asian countries.

For example, the Mu MT system, which was developed from 1982 to 1986 at Kyoto University and the JICST (Nagao and Tsujii, 1986) (and which has been later followed by the Majestic system at JICST (Ashizaki, 1989)), for the translation of scientific paper abstracts or titles, had a transfer structure quite similar to the ARIANE transfer structure (Hutchins (1986), p.319-321).

In Europe, from 1982 to 1992, the Eurotra project[7] was an international project aiming to achieve multilingual translation for the European Community. Even if its final results were poor in comparison with the initial expectations, this project helped to set new bases for MT research in Europe.

The late seventies, eighties and nineties saw a development of commercial MT systems, linked with the emergence of mini and then micro-computers. This trend has been especially strong in Japan (Nagao, 1989) where several commercial systems such as, for instance, the Fujitsu Atlas 2 system (an interlingua-based Japanese-to-English LBMT system (Uchida, 1989)), the Toshiba Transac system (an English-to-Japanese and Japanese-to-English LBMT system, with a deep syntactico-semantic parsing, inspired in part by the ARIANE system

[7] which followed the 1977-1982 Leibniz project.

(Amano et al., 1989)), the Sharp Duet-E/J system (an English-to-Japanese LBMT system, with a deep syntactico-semantic parsing (Sata, 1993)), and the Hitachi Hicats/JE system (a Japanese-to-English LBMT system, with a deep syntactico-semantic parsing (Kaji, 1987)) have been developed and commercialised.

Most of the commercial systems of that era (including Western systems and Japanese systems) are LBMT systems (Boitet et al., 2009). They however have different architectures, some of them using an interlingual semantic layer and others doing a simpler analysis of the source sentence. Many of them (such as Atlas 2 (Fujitsu, 2012) or Systran (Systran, 2012)) have been updated and are still commercialised or available in free online versions now.

Since 1996, a pivot language called Universal Networking Language(UNL) has been developed. The idea inherent to this pivot language is to be able to give semantic representations of source texts written in any human natural language, thus enabling to achieve multilingual interlingua-based LBMT (see Boitet (2005); Boitet et al. (2007); Uchida and Zhu (2005)).

Since the eighties, and even more in the last fifteen years, statistical systems (Koehn et al., 2007) have considerably progressed (Boitet et al., 2009), taking advantage of computational capacity of modern computers to perform machine learning on large text corpora. This technological evolution has also been helpful for EBMT, which is another data-driven approach.

Furthermore, in the last fifteen years, there has been a huge development of free MT websites on the web (Wilks, 2009), making MT a more accessible service for users around the world everyday. One of the most famous free online MT systems, Google Translate (Google, 2012a), has been available on the web since 2007 and is SMT (or hybrid[8]) system. Other free online systems (or free on-line versions of commercial systems such as Lucy Kwik Translator (Lucy LT, 2012) or Personal Translator Demo (Linguatec, 2013)) are either data-driven, linguistics-based or hybrid systems.

More recently, a growing interest has been shown in statistical systems integrating syntactic components. This approach enables to achieve better word and component reordering than basic SMT systems, especially for structurally distant language pairs such as Chinese-English (Zollmann et al., 2008). Another recent trend is the addition of statistical components to RBMT systems, sometimes called hybridisation (Boitet et al. (2009); Gulati (2011)).

Research systems

While the way commercial translation systems work is not always publicly explained, the principle of operation of the systems developed for research purposes

[8]Google Translate can be considered as hybrid for the language pairs containing languages with non-trivial morphology.

is always described in detail in scientific publications. We will show here an enumeration of some research MT systems, classified by types. We will give a more detailed description about the principle of operation for some of them.

Linguistics-based approaches LBMT was the dominant paradigm in research MT systems (such as Ariane (Zaharin, 1990)) before the development of statistical and machine-learned techniques. We distinguish here 3 types of handcrafted linguistic approaches: rule-based MT, other transfer-based LBMT and interlingua-based LBMT.

Rule-based MT RBMT generally consists of at least three steps: analysis[9] of the source sentence, transfer and generation of the target sentence[10]. The translation of the syntactic and/or semantic representation of the source sentence is achieved by the use of transformational grammar rules of a specific formalism that depends on the MT system (Zaharin, 1990).

Most recent RBMT research systems are implementations of open-source MT toolkits. For example, the Opentrad project has coordinated the creation of two open-source RBMT systems for Spanish languages and dialects: Apertium and Matxin. Apertium (Forcada et al. (2009); Costa-Jussa et al. (2010)) is a multilingual system designed for translation between related languages. It was at first used only for translation between Spanish languages such as Catalan-Spanish, without reaching than the level of quality of the Lucy output on this language pair. Then, its derived implementations have been applied to other European language or dialect pairs, such as Swedish-to-Danish (Tyers and Nordfalk, 2009).

Matxin (Alegria et al., 2007) is relying on a deeper sentence representation, for the translation of more distant language pairs, such as Spanish-Basque[11]. In a compared evaluation, Matxin has obtained better HTER scores[12] than the SMT system Matrex on Spanish-to-Basque translation.

Openlogos (Barreiro et al., 2011) is the open-source version of the Logos multilingual rule-based commercial system. It has been used for translation from English to German and to four Romance languages and from German to English or Italian, and an adaptation of OpenLogos has been developed for English-to-Hindi translation. Logos and OpenLogos rely on a source sentence analysis that contains both syntactic and semantic information, and that is computed in an incremental way (see Figure 2.3). This representation is supposed to solve

[9]The term *parsing* is also used.

[10]The term *synthesis* is sometimes preferred to *generation*, as in (Zaharin, 1990).

[11]Unlike other Spanish dialects that are closely related to standard Castilian Spanish, Basque, which is not a Romance language, is not related to Spanish.

[12]The HTER (Human-targeted Translation Edit Rate) score evaluates the quantity of manual post-edition needed on MT output. A better output is supposed to require less manual post-edition (Zaidan and Callison-Burch, 2010).

problems caused by structural ambiguities (Scott (1977), p.14).

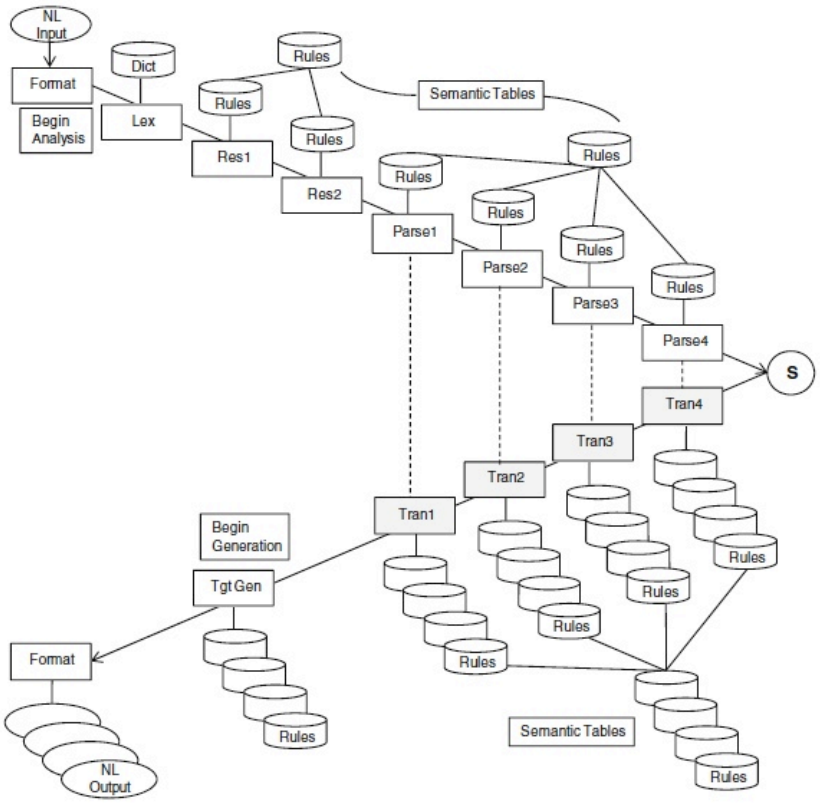


Figure 2.3: *OpenLogos system architecture (Barreiro et al., 2011)*

Even though RBMT is usually based on handcrafted transfer rules, other alternatives also exist. In (Imamura et al., 2003), a method is proposed for automatic acquisition and improvement of transfer rules for a Japanese-to-English RBMT system, showing positive results.

Other transfer-based LBMT Some other transfer-based LBMT systems are closely related to RBMT but their architecture differs in some points. They do not depend on formal transformational grammar rules, and their transfer phase is implemented in a procedural way.

This kind of procedural complex transfer phase was found in the MT system described in (Kinoshita et al., 1992). This LBMT system builds a logical syntactico-semantic representation of the English input sentence (which still contains the English vocabulary from the input sentence) and applies the bilingual transfer to the logical representation. This method is especially useful

for the treatment of structural asymmetries in English-to-Japanese translation (Kinoshita et al. (1992); Tsujii and Fujita (1991)).

The ALT-J/E system (Ikehara et al. (1991); Nakaiwa et al. (1994); Bond (2005)) is another transfer-based procedural LBMT system, with advanced semantic analysis and disambiguation, for Japanese-to-English translation.

Its-2 (Wehrli et al., 2009a) also has a classical transfer-based architecture, with transfer transformational rules implemented in a procedural way. Its architecture will be described in detail in section 2.3.3.

Interlingua-based LBMT Interlingua-based LBMT achieve intermediate steps where the source sentence is translated into an *interlingua* or *pivot language* and the interlingua is translated into the target language before the generation of the target sentence.

The pivot language can either be a human language (such as Esperanto in the DLT MT system (Witkam, 1988)) or an abstract syntactico-semantic representation (such as UNL). With an interlingual representation which is much more independent from the source language than the syntactico-semantic representations mentioned previously, interlingua-based LBMT reduces the influence of the source sentence structure on the output. Therefore, if the pivot language is an abstract structure where only interlingual relations or features are kept, the method can be particularly relevant for handling structural asymmetries between distant languages.

This idea has been described in (Barnett et al., 1991) and (Dorr, 1994). It has been applied in the Japanese-to-English interlingua-based LBMT system Atlas-2.

Later, the UNL language has been developed in the same spirit as the Atlas-2 interlingua, and aiming at being a fully universal pivot language. So far, UNL interlingual LBMT can generate translation between 47 languages, at various level of implementation, giving better results for translation of English, Japanese and Bengali (UNDL Foundation (2013)). However, most publications about UNL interlingual LBMT only mention UNL-to-target language automatic deconversion, which suggests that most researches about automatic source language-to-UNL enconversion have not led to high quality results yet.

Example-based approaches Knowing the difficulty involved in structural asymmetries in translation between distant languages, especially in Japanese-to-English translation, another approach has been taken with EBMT (Somers, 1999): the use of example sentences.

Instead of translating mainly words or collocations and applying reordering and structural modification to the generated sentence, EBMT systems try, as much as possible, to translate the whole sentence using sentence translation

examples, selecting the most similar example sentence and replacing words or phrases that do not match with the example. This method was first proposed in (Nagao, 1984) and presented as a possible solution for the translation of distant languages such as English and Japanese.

We present here EBMT systems that represent three different variants of example-based approaches: syntactical EBMT, interlingual EBMT and analogical EBMT.

Syntactical EBMT The Microsoft Research system is a multilingual system, and its development was particularly advanced for English-to-Japanese translation (Brockett et al., 2002). It performs a syntactic analysis augmented with a logical semantic representation of the source sentence. It uses example sentences of the source and target languages, which have also been been parsed in logical semantic trees. Transfer is performed with a unique algorithm, automatically learned (for both translation directions) from the parsed example sentences, transforming the source tree into a target tree. In an evaluation, it showed results that were equivalent to those obtained by non-tuned English-to-Japanese LBMT systems. It has been used by Microsoft for user documentation translation for several language pairs.

However, obtaining good results with its syntax-augmented statistical system Treelet in (Quirk and Menezes, 2006), it seems that the Microsoft Research team has been working on syntax-based statistical machine translation (SBSMT) rather than syntactical EBMT since that period (see the *Syntax-based SMT* paragraph).

The Kyoto EBMT system (Kurohashi et al. (2005); Nakazawa and Kurohashi (2010) is also a syntactical EBMT system, following many of the principles defined in (Nagao, 1984). It performs a syntactic parsing in syntactico-semantic sentence representations in dependancy grammar trees (with the KNP parser (Kurohashi and Nagao (2003)) that focus on Japanese case grammar, when Japanese is the output language) (see Figure 2.4). It makes use of parsed example sentences, and transfer is achieved measuring similarity with the example sentences and choosing the most similar ones for the translation of full sentences or linguistic components. Heuristics and statistical corrections are also applied in the transfer and generation phases. In a hand-made evaluation task on Japanese-to-English scientific paper abstract translation, the Kyoto system gave positive results in semantic adequacy. Those results were roughly equivalent to those obtained by a RBMT system, and clearly better than those of a phrase-based SMT system (Nakazawa and Kurohashi, 2010).

Interlingual EBMT The MMT system based on the EDR dictionary was an interlingual EBMT system, presented in (Yasuhara, 1993). After syntactic

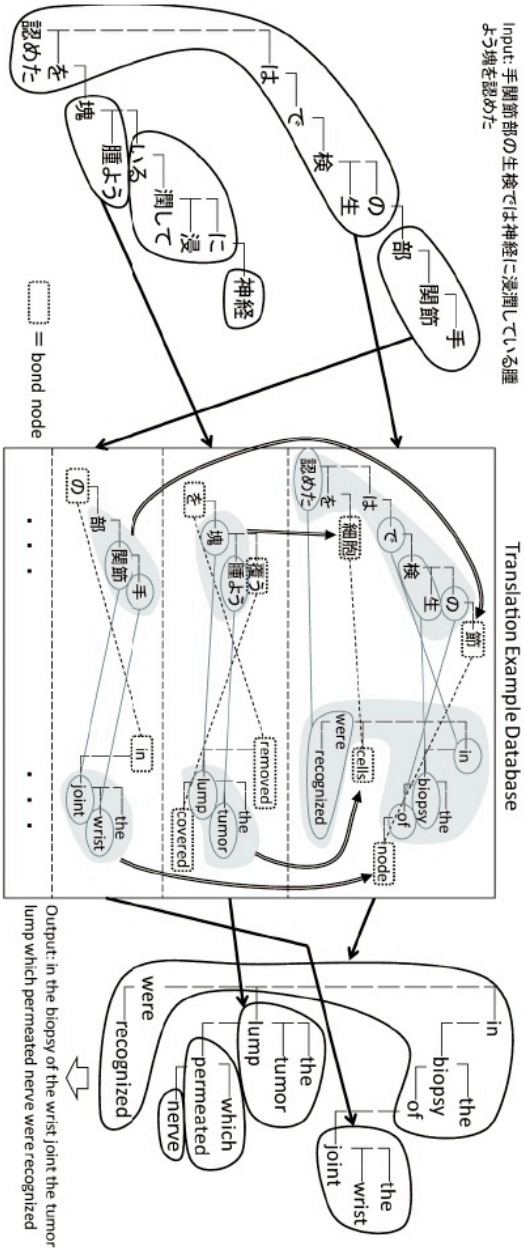


Figure 2.4: *KYOTO system architecture (Nakazawa and Kurohashi, 2010)*

and semantic analysis, source sentence as well as examples were converted into the Atlas-2 conceptual interlingua (see the *Japanese systems* paragraph), making use of the EDR electronic lexicons and their semantic classification. The similarity was then calculated on the interlingual conceptual representations.

Analogical EBMT The ALEPH and GREYC systems (Lepage and Denoual (2005a); Lepage and Denoual (2005b); Lepage and Lardilleux (2007)) are fully analogical EBMT systems. They were based on the analogy technique mentioned in (Nagao, 1984), but did not achieve any syntactic analysis. They just referred to proportional analogies detected in example sentences, which were processed as simple character strings.

The ALEPH system obtained results that were equivalent to those obtained by the best other systems in the IWSLT 2005 Japanese-to-English translation task (Lepage and Denoual, 2005a). However, its evolution (the GREYC system) failed to obtain good ranking in both automatic and manual evaluations for the IWSLT 2007 Japanese-to-English translation task, using training data that was four times smaller than in the 2005 task (Lepage and Lardilleux, 2007). This seems to show that the required amount of training data is higher for analogical EBMT than for some other techniques such as phrase-based SMT, because of the need for a training corpus with enough *analogical density*.

Statistical systems In the early nineties, the first SMT models were based on word translation, calculating the probability of translating a word into another one (Koehn (2010), p.118). They allowed limited word reordering. This technique was not adapted for the translation of structural asymmetries, especially for language pairs that need long-distance reordering, such as English and Japanese. However, it has brought a big improvement in lexical selection for a more fluent output, especially with the introduction of n-gram language models, that give an estimation of word phrase fluency. Language model improvement is still studied in current research (Xiong et al., 2011).

Phrase-based SMT Phrase-based SMT has been introduced in 1998 and has been one of the dominant approaches recently (Koehn (2010), p.148). It is based not only on word translation probabilities, but also on phrase translation probabilities. It is therefore closer to EBMT than word-based SMT is. However, unlike most EBMT approaches, original phrase-based SMT is not based on syntax, and pays attention to non-linguistically motivated phrases as well as linguistically motivated phrases or sentences[13].

[13]Here, we distinguish *linguistically motivated phrases* (such as any noun phrase (NP), determiner phrase (DP), verb phrase (VP), etc.), and *non-linguistically motivated phrases* (such as "of the", "but we", "and a", "milk and", etc.), which can be found in text without forming syntactic constituent.

Due to its capacity to capture language asymmetries in non word-by-word translation and despite limitations in handling long-distance reordering, phrase-based SMT has often been chosen for Japanese-to-English and Chinese-to-English translation tasks (Zhang et al. (2006), Shen et al. (2006)).

Moreover, open-source toolkits such as Moses (Koehn et al., 2007) have enabled users to make their own phrase-based SMT system, by training the (non trained) toolkit component on learning MT for the language pair (and language domain) of their choice, using aligned bilingual corpora. This task can be achieved quite fast by the users, as long as bilingual corpora are available for the chosen language pair.

Hybrid systems These systems are based on a combination of different paradigms. One of the most frequent hybrid approaches is RBMT (or LBMT) corrected or mixed with data-driven MT (Thurmair, 2009). This strategy is divided into several versions such as combination of LBMT and EBMT, RBMT with statistical post-editing, RBMT with statistical components. Other hybrid approaches consist in a combination of several MT systems. A description of the different existing hybrid methods has been shown in (Thurmair, 2009) and in (Gulati, 2011).

Combination of LBMT and EBMT Combination of linguistic-based and example-based MT was proposed in (Shirai et al., 1997), for Japanese-to-English translation. This hybrid system was doing example-based MT when matching examples were available and LBMT in other cases, reusing the ALT-J/E LBMT system. Furthermore, it was using rule-based methods to replace non-matching elements in selected translation examples.

A similar technique was later proposed in (Sanchez-Martinez et al., 2009), adding and using translated sentence examples with the RBMT system Apertium, slightly improving the results of the RBMT system for English-to-Spanish and Spanish-to-English translation.

RBMT with statistical post-editing More recently, another method has been presented and has been the focus of many works: rule-based systems with statistical post-editing (Simard et al. (2007); Dugast et al. (2007)). This method is based on RBMT and machine learning of manual post-edition of the RBMT output. Thus, a SMT system is trained to learn automatically the typical corrections that need to be done on the RBMT system output. This training can be achieved either with manually post-edited translations or reference translations (Ueffing et al., 2008). This technique has been applied to the commercial RBMT system Systran for Western languages translation (Simard et al. (2007); Dugast et al. (2007)) and for Chinese-English translation(Ueffing

et al., 2008).

It has also been used with the LBMT system Its-2, improving null subject translation (Russo et al., 2012). Results have shown improvements in translation quality, in comparison with pure SMT or pure RBMT translation.

This method has also been applied to Japanese-to-English translation in (Ehara, 2007), (Suzuki, 2011) and (Toue et al., 2011). In (Ehara, 2007), statistically post-edited output was not necessarily better than pure RBMT output for Japanese-to-English patent translation: opposite results were found, depending on the chosen evaluation method. In (Toue et al., 2011), manual evaluations show that statistical post-editing on the output of the deep semantico-syntactic LBMT system Toshiba Transac decreases the global translation quality while giving higher scores with automatic translation quality evaluations, both questioning the validity of the method and the validity of automatic evaluations for Japanese-to-English translations. (Suzuki, 2011) proposes an automatic sentence quality assessment, in order to evaluate the need for statistical post-editing of RBMT produced output.

RBMT with statistical components This other method requires the addition of statistical selections and corrections at different possible stages along the rule-based translation process (Font Llitjós and Vogel, 2007), while statistical post-editing is achieved only after the translation process. (Eisele et al., 2008) describes the improvement of a lexicon for RBMT by statistical techniques.

(Font Llitjós and Vogel, 2007) describes a RBMT system improvement at five steps: automatic generation rule refinement, use of language model scores for sentence selection, semi-automatic lexicon improvement, statistical word selection and statistical rule selection. However, these works remain dedicated to translation between Western languages. (Ahsan et al., 2010) have experimented a deep combination of SMT and RBMT techniques for English-to-Hindi translation[14], where both rule-based and statistical methods are applied for analysis, transfer (local reordering and long-distance reordering) and generation (see Figure 2.5). They have obtained very positive results.

(Jin, 2010) presents a Chinese-to-English RBMT system improved by the integration of semantic disambiguation components. The author presents the strategy as a hybrid RBMT and semantic strategy. However, it is unclear if the term *hybrid* is really appropriate for this system, as semantics layers have often been included to classical LBMT systems (Barreiro et al. (2011); Kinoshita et al. (1992)).

[14]English-to-Hindi translation requires not only local, but also long-distance constituent reordering, as English-to-Japanese translation does.

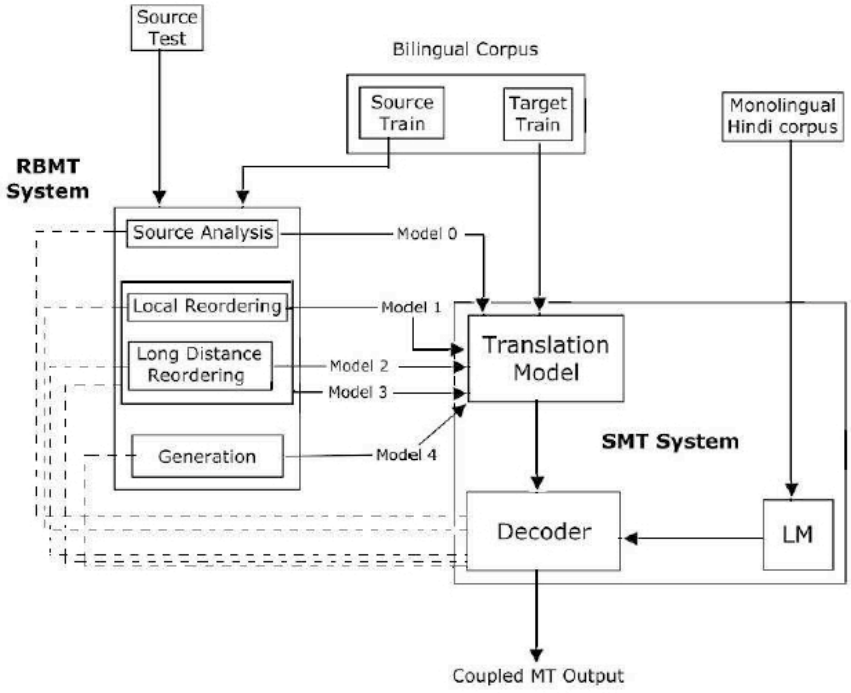


Figure 2.5: *LTRC English-Hindi system architecture (Ahsan et al., 2010)*

MT system combination Hybrid MT can also be achieved by MT system combination, selecting or merging the outputs produced by the different systems (Frederking and Brown (1996); Hildebrand and Vogel (2010); Eisele et al. (2008)).

Syntax-augmented statistical systems Statistical systems have good results in handling lexical selection and fluency in the output. However, basic SMT systems have sometimes weaknesses in producing grammatical output and in handling syntax and morphology. In order to solve these problems, a recent trend has consisted in adding linguistic and especially syntactic components or techniques into statistical systems. This has resulted in the development of a particular type of hybrid systems. This paradigm is divided into several approaches such as hierarchical machine translation (HMT), SMT with reordering pre-processing, syntax-based SMT (SBSMT).

Hierarchical phrase-based MT HMT is considered to be the simpler paradigm in syntax-augmented SMT. It consists, instead of learning flat phrase-to-phrase translation as it is done in phrase-based SMT, in allowing the system to learn transfer rules that include non terminal symbols (Chiang et al. (2005), Chiang (2007)). These machine-learned transfer rules are transformational grammar rules in the *synchronous context-free grammar* (CFG) formalism (see 2.6), hence recalling the RBMT paradigm. However, a notable difference with RBMT is found in the fact that HMT transformational rules are not necessarily linguistically motivated. In (Chiang et al., 2005), it is shown that the Hiero HMT system gives better results than a phrase-based MT system for Chinese-to-English translation, giving significant improvement in word reordering.

(Gesmundo and Anderson, 2011) present a Chinese-English HMT system built with the free open-source cdec HMT toolkit (Dyer et al., 2010), applying a non bottom-up decoding algorithm[15] that slightly improves the results for Chinese-English translation.

(Wu and Tsujii, 2011) described a HMT system for English-to-Japanese translation including a katakana/Latin characters transliteration component for technical term translation training.

HMT can also be combined with RBMT in deeply hybrid approaches. (Chen and Eisele, 2010) presents a hybrid HMT-RBMT system that includes rules derived from translations produced by the RBMT system Lucy (ex-METAL) in the transfer rules of the HMT system. The author observes better results with this hybrid system than with HMT or RBMT systems, for German-to-English

[15] *Decoding* is the phase where the best translation is estimated and chosen among the possible choices, in SMT.

$\langle S_{[1]}, S_{[1]} \rangle \Rightarrow \langle S_{[2]} X_{[3]}, S_{[2]} X_{[3]} \rangle$
$\Rightarrow \langle S_{[4]} X_{[5]} X_{[3]}, S_{[4]} X_{[5]} X_{[3]} \rangle$
$\Rightarrow \langle X_{[6]} X_{[5]} X_{[3]}, X_{[6]} X_{[5]} X_{[3]} \rangle$
$\Rightarrow \langle$Aozhou $X_{[5]} X_{[3]}$, Australia $X_{[5]} X_{[3]} \rangle$
$\Rightarrow \langle$Aozhou shi $X_{[3]}$, Australia is $X_{[3]} \rangle$
$\Rightarrow \langle$Aozhou shi $X_{[7]}$ zhiyi, Australia is one of $X_{[7]} \rangle$
$\Rightarrow \langle$Aozhou shi $X_{[8]}$ de $X_{[9]}$ zhiyi, Australia is one of the $X_{[9]}$ that $X_{[8]} \rangle$
$\Rightarrow \langle$Aozhou shi yu $X_{[1]}$ you $X_{[2]}$ de $X_{[9]}$ zhiyi, Australia is one of the $X_{[9]}$ that have $X_{[2]}$ with $X_{[1]} \rangle$

Figure 2.6: *Partial derivation of a context-free grammar (Chiang et al., 2005)*

translation.

SMT with reordering pre-processing Syntactic reordering preprocessing consists in reordering the source sentence before the translation process, following the word and constituent order of the target language. This approach can be particularly useful for MT methods that cannot handle long-distance reordering well, such as phrase-based SMT. (Collins et al., 2005) showed such a preprocessing for German-to-English translation: a set of linguistically motivated handcrafted reordering rules are been applied after the source sentence syntactic parsing; then, the reordered German sentence is translated into English by the phrase-based SMT system. Results were clearly better with the reordering preprocessing than with pure phrase-based SMT[16]. Similar work was presented by the authors for Chinese-to-English translation, again giving good results (Wang et al., 2007a). Similar approach was applied for Japanese-to-English translation, achieving reordering on the basis of a syntactico-semantic analysis of the Japanese sentences, giving positive results (Komachi et al., 2006).

Syntax-based SMT SBSMT aims at taking advantage of information given by syntactic parsing to improve statistical translation (Yamada and Knight (2001); Cowan et al. (2006)). Instead of learning phrase-to-phrase translation (like phrase-based MT) or learning hierarchical transformation rules without the guidance of a syntactic parse (like HMT), the source sentences are parsed and translation from syntactic tree to target phrase is learned. This method, called *tree-to-string* SBSMT enables the learning of long distance reordering that could not be captured in basic phrase-based SMT and prevents from learning

[16]The technique of reordering preprocessing is not a new one. The innovation there was to apply it for SMT.

non-linguistic transformations as it can happen in HMT. It is especially relevant for the translation of structural asymmetries language pairs such as English-Japanese, as it is shown in Figure 2.7.

Since 2002, the Microsoft Research team has also worked on tree-to-string (or string-to-tree) SBSMT called Treelet (Quirk and Menezes, 2006). They apply logical form parsing to English sentences and learn automatically the transfer phase from the trees to aligned target sentences. This method has helped them develop MT from or into languages where data for syntactic parsing was lacking.

(Wang et al., 2007b) presented Chinese-to-English tree-to-string SBSMT with binarized syntactic trees[17], showing that the binarization process can significantly improve the translation results.

In order to handle both source and target language syntactic or semantico-syntactic structures, *tree-to-tree* SBSMT has also been studied. Problems related to syntactic differences between source and target tree structures have been partially overcome, evaluations giving positive results for German-to-English (Cowan et al., 2006) or Chinese-to-English and Arabic-to-English (Chiang, 2010) tree-to-tree translation.

As parsing errors can occur and damage the quality of SBSMT, the *forest* concept has been introduced (see Figure 2.8). It consists in considering several possible syntactic trees for a same sentence and thus reducing the negative effects of parsing errors. It has been used to improve results in tree-to-tree, tree-to-string or string-to-tree SBSMT ((Liu et al., 2009); (Zhang et al., 2009); (TaroWatanabe and Sumita, 2011)).

The open-source toolkit Joshua (Li et al., 2009) enables users to create SBSMT or HMT systems.

SBSMT, as well as HMT, are closer to syntactical EBMT than basic SMT is. Still, differences remain and syntactical EBMT remains more deeply syntactical, as shown in Nakazawa and Kurohashi (2010). However, the fact that HMT and SBSMT are independent or less dependent on syntactic parsing than syntactical EBMT and LBMT can also be seen as an advantage, as syntactic parsers often give incorrect or incomplete analysis (see Bond (2005)).

Comparisons A comparison of the different SMT and syntax-augmented SMT approaches has been achieved in (Zollmann et al., 2008). It has shown that HMT gave better results than phrase-based MT and that SBSMT gave even better results, for Chinese-to-English translation. However, phrase-based SMT with local reordering achieved by the use of language models gave the best results for Arabic-to-English translation, where much less component reordering is needed.

In (Lee et al., 2010), a combination and comparison of syntax-augmented

[17]A binarized tree is a tree where every node leads to, at most, 2 branches.

Figure 2.7: *English-to-Japanese Tree-to-String translation channel for the sentence "He adores listening to music." (Yamada and Knight, 2001)*

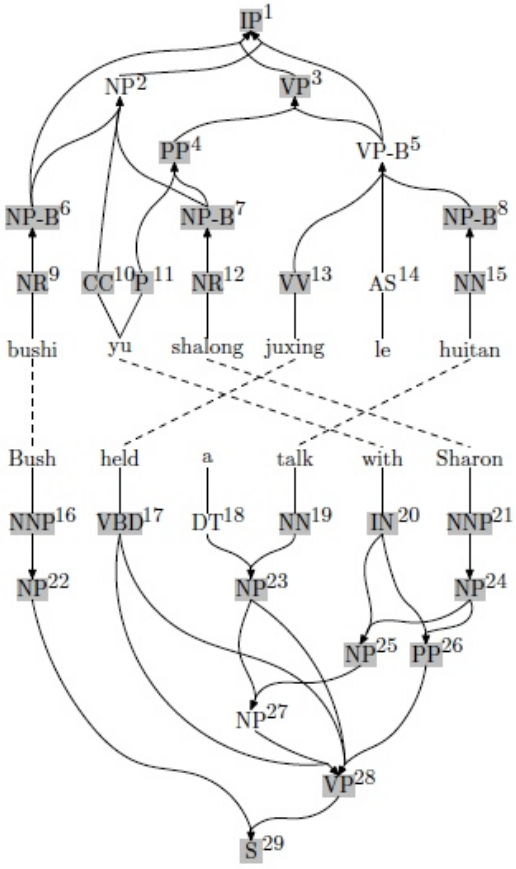


Figure 2.8: *An example of two aligned packed forests (Liu et al., 2009)*

SMT methods for English-to-Japanese translation, that requires more reordering than Chinese-to-English translation, has been described. Results have shown that reordering preprocessing on phrase-based MT (as in Figure 2.9) can outperform SBSMT and HMT, and that the best results can be obtained by a combination of reordering preprocessing and tree-to-string SBSMT.

Current commercial systems

We will give here a list of current Japanese and Western multilingual commercial MT systems. As much less detailed updated technical descriptions are available for current commercial systems than for research systems, we will only give brief comments on their architecture.

Japanese systems Commercial MT has a long history in Japan where many commercial systems have been commercialised since the eighties and nineties (Wilks (2009), p.128-133). They have been used in the country, especially for professional technical and formal document translation. Most Japanese commercial systems have been programmed for Japanese-to-English and English-to-Japanese translation. Some, such as HonYaku Pikaichi, now also achieve multilingual translation. Some have been dedicated to the Japanese-Korean or Japanese-Chinese language pairs.

Faced with the problem of structural asymmetries between Japanese and English, commercial Japanese LBMT systems have integrated techniques such as deep syntactic transfer, semantic analysis and transfer, or interlingual approach. Large-scale lexicons have been developed for these systems, including semantic data, syntactic data, collocation and other multi-word expressions (Wilks (2009), p.126-133). Furthermore, they almost all make use of translation memories in addition to the transfer-based or interlingua-based translation. We present here some of the currently commercialised Japanese MT system for the English-Japanese language pair. Among the 8 systems, the first 4 systems are all transfer-based LBMT systems, the 5^{th} one can be considered as a a hybrid EBMT-RBMT system, and the 6^{th} one is an interlingua-based LBMT system.

Honyaku Pikaichi is Cross Language's transfer-based LBMT system. The Honyaku Pikaichi free online version is used in popular Japanese internet portal such as Rakuten-Infoseek, Excite and Yahoo! Japan (Yahoo! Japan, 2012). It uses English as a pivot language for other European language translation. Cross Language also sells other specialised professional MT tools (Cross language, 2012).

Yakuse!! Goma (MT Labs, 2012) is the Nagano- based company MT Lab's transfer-based LBMT system. It has inherited from the Bravice MT system technology. It has been used on the SDL website Freetranslation.com for English-Japanese and Japanese-English translation.

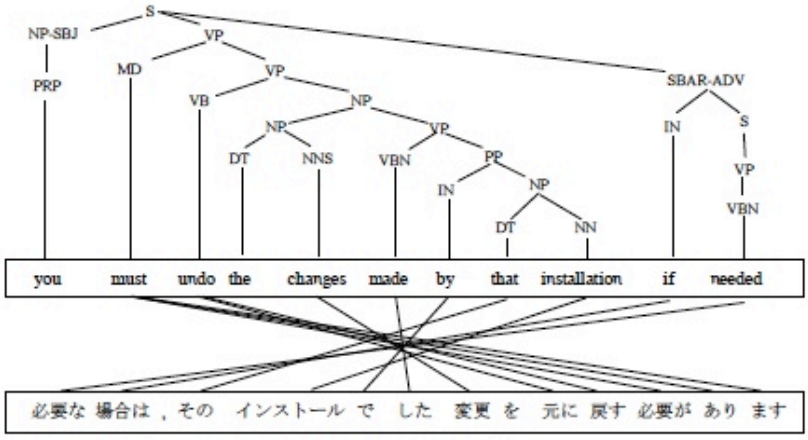


Figure 1. Parse Tree and Word Alignment before Reordering

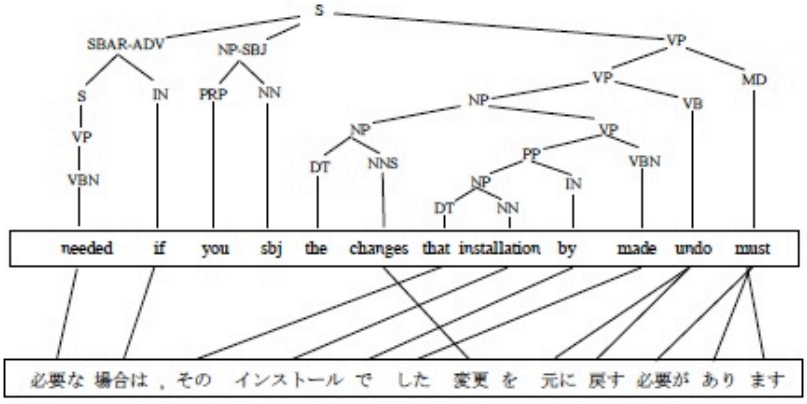


Figure 2. Parse Tree and Word Alignment after Reordering

Figure 2.9: *English-Japanese alignment modification after reordering preprocessing (Lee et al., 2010)*

IBM's system Honyaku no Ôsama V5 (IBM, 2012) is a famous professional transfer-based LBMT system, based on a dependancy parser. It can now also handle dialogue or day-to-day language text translation, a function that has often been lacking in professional Japanese MT systems (Boitet et al., 2006).

Toshiba's MT system The Honyaku V15 (Toshiba (2012); Sakai et al. (2003)) is another famous professional transfer-based LBMT system. It is based on an ATN[18] parser.

Yakushite.net is Oki's free online translation page (Oki, 2012). It has replaced Oki's older system Pensée for Internet. Yakushite.net is described as a *pattern based* MT system, which can be considered as a hybrid system between syntactical EBMT system and RBMT. It achieves syntactic parsing, uses example sentences and word dictionaries, and accomplishes the transfer phase making use of lexicalised CFG transfer rules (called *translation patterns*). Some of these rules are handcrafted and others are automatically learned from the parsed example sentences. Finally, it applies handcrafted rules that correct verbal conjugation in the output sentence(Sasaki and Murata, 2005). Yakushite.net uses crowd-sourcing from users' feedback to increase the number of example sentences and improve the translation quality of the system.

Atlas V14 is Fujitsu's MT system (Fujitsu (2012)). It is an interlingua-based LBMT system. It has inherited from the Atlas-2 technology (Uchida, 1989), which has defined an interlingual representation that has inspired the UNL language (Boitet et al. (2006)). Atlas V14 offers to users many specialised lexicons and several possible writing styles that can be selected for the output text.

Crossroads for Enterprise (Nec (2012)) is NEC's professional MT system, doing translation between English and Japanese, Japanese and Chinese and Japanese and Korean. It has replaced the NEC Pivot interlingua-based LBMT system (which was commercialised under the name Honyaku Adaptator II).

翻訳J・E・T ("Honyaku JET") is Kodensha's MT system for translation between English and Japanese, that has replaced in 2004 Kodensha's older system J・London that was commercialised since 1994 (Kodensha, 2012). Kodensha also sells systems for translation between Japanese and Chinese, and Japanese and Korean.

Multilingual systems Multilingual MT systems have had an increasing worldwide success since free versions have been available online on the internet (Wilks (2009) p. 225; Boitet et al. (2009)). The seven following systems are fully multilingual and all achieve English-to-Japanese translation among the language pairs

[18]ATN: Augmented Transition Networks. An ATN is a network in which parsing is described as the transition from a start state to a final state in a transition network corresponding to the grammar of a language. ATN grammars can generate all recursively enumerable sets.

they can translate. Among these systems, five are SMT ones and it seems that the kind of SMT (phrase-based or syntax-augmented (Zollmann et al., 2008)) they are using has not been explicitly revealed. This choice may depend on language pairs and system versions.

Systran is a multilingual MT system that has been available since 1968. It used to be a RBMT system (Boitet et al., 2006). In 2010, it has been converted into a hybrid system with the launch of Systran Server 7, a move that followed research on statistical post-edition done by Systran researchers (Dugast et al., 2007). It achieves translation between 14 languages for 52 bi-directional language pairs (between European languages and between English and non-European languages). It also is available in a free online version, now called SYSTRANet (Systran, 2012). It was used as a free translation service called Babelfish, for the internet search engines Yahoo! and Altavista, until June 2012.

Google Translate (Google, 2012a) is a free online multilingual SMT (or sometimes hybrid, depending on the language pair) system, available since 2007[19]. It is maybe the most famous MT system, in part due to the notable success of the Google search engine. Google Translate is trained on multilingual corpora such as UN or EU documents or web-extracted texts (Google, 2012b) and achieves translation between 65 languages. Translation for the 4160 possible unidirectional language pairs is proposed to users[20]. As the system has not been trained for most of the proposed language pairs, untrained language pairs such as French-Japanese are translated using English (or other languages) as a pivot language (see in section 2.3.2). Furthermore, Google Translate offers an intuitive interface for lexical post-editing.

SDL-LW (Language Weaver, 2012) is a multilingual phrase-based SMT system that has inherited from the technology of the Language Weaver MT system. It currently achieves translation from or into English (or sometimes French or Spanish) with 38 languages. Freetranslation.com is the free online version of SDL-LW. It produces translation between 34 languages, allowing all the possible 1122 unidirectional language pairs, using as well English as a pivot language for untrained language pairs.

Bing Translator (Microsoft, 2012), is Microsoft's free online multilingual SBSMT system, which has inherited from the Treelet system technology. It achieves translation between 38 language pairs, allowing 1406 unidirectional language pairs and also uses English as a pivot language for untrained language pairs.

Asia Online's Language Studio (Asia Online, 2012) is a multilingual SMT system, producing translation between 31 languages (mainly European and

[19]Before the end of 2007, Google's free translation service already existed, but the translation was produced by Systran and not yet by Google's own system.

[20]For n languages, $n*(n-1)$ unidirectional language pairs exist. In the case of Google Translate!, n=65, which makes 65*(65-1)=4160 possible unidirectional language pairs.

Asian languages). It does not uses English as a pivot language and achieves direct translation for more than 600 unidirectional language pairs, such as for example Japanese-to-German. It does not offer any online version.

Personal Translator (Linguatec, 2013) is a German MT system for translation between English and five Western languages, between German and French, and between English and Chinese.

Other multilingual commercial MT systems remain specialised in Western languages. ProMT is a Russian LBMT system, (Carrera et al., 2009), claiming to produce consistent output, even when the input contains unknown words. It has been developed since the late eighties, and now domain-specialised hybrid versions of the system are also developed. It achieves translation between Western languages and Russian. Its free online version is available on the online-translator.com web page and on the Reverso (Softissimo, 2012) website, that uses ProMT for the translation of European languages.

Lucy (Lucy LT, 2012) is a LBMT system (formerly called Siemens METAL system) that achieves translation between Western languages and Spanish language and dialects. Kwik Lucy is its free online version.

Lionbridge iTranslator (LionBridge, 2012) is a free online LBMT system, that has inherited from Weidner MAT tool. It achieves translation between Western languages.

GramTrans (Wiechetek, 2008) is a rule-based LBMT system that achieves translation between Western languages, Scandinavian languages and Esperanto. It offers a free online version.

2.3.2 French-Japanese MT

MT between Japanese and French (French-to-Japanese or Japanese-to-French) has been studied less than MT between Japanese and English. However, research in this domain has existed for more than 50 years.

It started with Yamada's work on Japanese-French MT at the CETA in Grenoble in 1961 (see Hutchins (1986), p.129). Then, in 1970 in Grenoble again, Makinouchi (1970) described algorithms for transfer-based Japanese-to-French MT. However, those works remained theoretical ones.

In 1982, an experimental Japanese-to-French version of Atlas-2 was developed by H. Uchida at Fujitsu. It contained limited lexicons, containing between 10'000 and 20'000 entries. This prototype has never been extended to a more complete version. Later, a French-Japanese sentence alignment tool has been developed (Nakamura-Delloye (2005), Nakamura-Delloye (2007)).

French-Japanese MT has also been studied from a lexical point of view in (Mangeot and Kuroda, 2003) and in (Robitaille et al., 2006), where compilation of French-Japanese terminologies from the web has been achieved.

In spite of such research works dedicated to the French-Japanese language pair, no MT system has been fully implemented for direct French-Japanese or Japanese-to-French translation so far. Currently, only Asia Online states that versions of its system are now under development for French-to-Japanese and Japanese-to-French direct translation. As we have already mentioned, other current multilingual systems such as Google Translate, Reverso or Honyaku Pikaichi achieve French-to-Japanese translation using English as a pivot language: sentences are first translated from French to English, and then from English to Japanese. Even if this double translation process may increase translation errors, it can also produce satisfying output for simple sentence translation, due to the proximity of the French-English language pair (see Chapter 5.4.3).

2.3.3 Its-2

Its-2 is a multilingual linguistics-based MT system (Wehrli et al., 2009a). It has been developed (at LATL in Geneva University) following a classical transfer-based architecture, with transfer and transformational rules implemented in a procedural way. Its-2 has first been programmed for Western language translation (between French, English, German, Italian (Russo and Wehrli, 2011) and Spanish). The Its-2 French-to-Japanese and English-to-Japanese versions have been under development since 2008, and this development will be described in this thesis.

Architecture

Following the classic transfer-based LBMT architecture, the translation process is divided into three steps: syntactic parsing, transfer and generation (Wehrli et al., 2009a). Developed in an object-oriented approach, the system is made of main core modules that lead the translation process in a generic way, and language specific modules that adapt it to the syntactic, morphological and lexical properties of the translated languages. It makes use of monolingual and bilingual word and collocation lexicons at the three steps of the translation process (see Figure 2.10).

Syntactic parsing

The parsing phase is achieved by the multilingual syntactic parser Fips, which is as well divided into core modules and language-specific modules (Wehrli (2007), Wehrli and Nerima (2009)).

The source sentence is analysed in a deep syntactic structure, following a grammar which is inspired from Generative Grammar. Predicate-argument structures are recognised, following lexical information contained in the lexicons (see Chapter 3.5). Collocations can as well be detected.

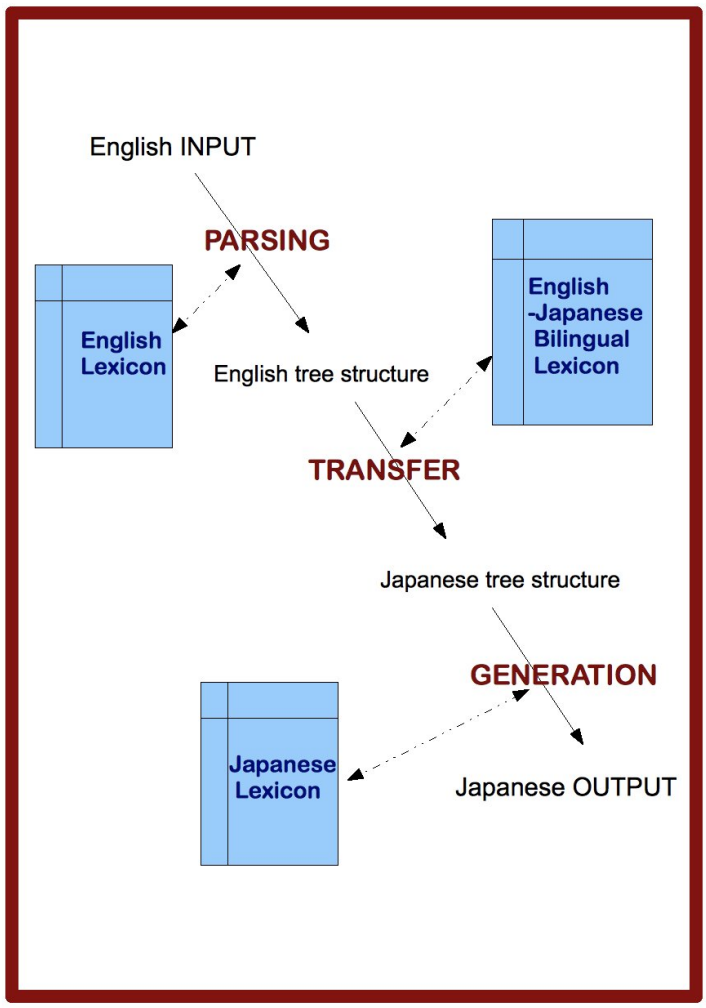


Figure 2.10: *The Its-2 translation process (English-to-Japanese example)*

The Fips and Its-2 grammar is inspired by theories in Generative Grammar from the seventies (Chomsky, 1972), also borrowing concepts from Lexical-Functional Grammar (Bresnan (1982); Bresnan (2001)) and from more recent works such as the Minimalist model (Chomsky, 1995) and the Simpler Syntax model (Culicover and Jackenfoff, 2005).

In its syntactic structures, every node has a (possibly empty) head as well as a left and right constituent list. Even if these syntactic structures are flatter than those found in current generative syntax, they can still contain empty categories such as null subjects or traces of constituents that have been moved from their canonical position. A more detailed description of the Fips and Its-2 grammar is given in (Wehrli and Nerima, 2009).

Fips has been programmed for the parsing of 24 languages, at various levels of development. Its parsing accuracy on the different languages depends not only on the level of implementation but also on the amount of data recorded in the monolingual lexicons. Experiments have enabled some simple Japanese sentence parsing with Fips, (Kauffmann (2008a); Kauffmann (2008b)), but the implementation has not been continued deep enough to reach a State-of-the-Art accuracy level. The English and French versions of Fips used in this thesis are much more advanced, and have consequently been used not only for syntactic parsing, lexical tagging and machine translation, but also in several NLP applications such as collocation extraction, text summarization, information extraction, or translation of word in context (see (Wehrli and Nerima, 2009)).

Transfer

The transfer phase is the phase where the tree structure obtained from the parsed source sentence is converted into a target language tree structure. This conversion is done by recursively traversing and transferring the source language tree and subtrees, in this order: head, right subconstituents, left subconstituents.

By default, the core transfer module has been programmed to do an isomorphic transfer and to produce a tree with the target language words or collocations and the same structure as the source tree, using lexical data contained in bilingual lexicons.

However, the specific transfer module (which is specific to a directional language pair, such as French-to-Italian or English-to-French) redefines the transfer phase, adding non-isomorphic transfer procedures and producing a target language tree that matches with the target language syntax. In this target tree, syntactic components have been either reordered, modified, added or deleted, depending on the transfer rules that have been implemented procedurally in the specific transfer module. These transfer rules may be systematic or depend on the syntactic, semantic or lexical context. Closely related languages such as French and Spanish can be approximately translated with only few implemented

rules, while distant languages such as English and Japanese need a heavy specific transfer.

The transfer process is slightly different for arguments of a recognised predicate: their properties in the target language depend on the subcategorisation frame of the target language predicate.

Generation

The generation phase is the last phase where, in a direct application of generative grammar theory, the target sentence is generated from the deep syntactic structure. The generation core module allows the extraction of target sentence words from the target language full form monolingual lexicon[21], while the target language specific generation module supervises elision or addition of link words, agreement between words and morphological selection between the different possible word forms. When the generation is completely achieved, the target sentence is outputted.

Lexicons

The architecture of Its-2 relies on rich lexical information. The aim of storing enough syntactic and semantic information in the lexicon is to reduce the complexity of the transfer step, reducing as much as possible the need for lexicalised transfer rule implementation.

The lexical databases of Its-2 are organised this way: monolingual lexicons are divided into three parts: the lexeme[22] table, the full-form word table and the collocation table which indicates which lexemes are found together in each collocation; bilingual lexicons store bilingual correspondences between lexemes or collocations, giving a score to every correspondence in each translation direction; those scores enable Its-2 to make a default ranking of the possible translations for every translated word or collocation (Wehrli et al., 2009a).

The monolingual lexicons of Its-2 contain about 41'000 French lexemes and 15'000 French collocations and about 57'000 English lexemes and 8000 English collocations. The bilingual lexicons contain more than 81'000 French-English bilingual correspondences and, for example, 40'000 French-Italian correspondences.

An example of sentence translation

In order to illustrate the Its-2 translation process, we can look at the different steps of the English-to-French translation of the sentence *The movement toward*

[21] A full form lexicon is a lexicon that stores all the inflected word forms.

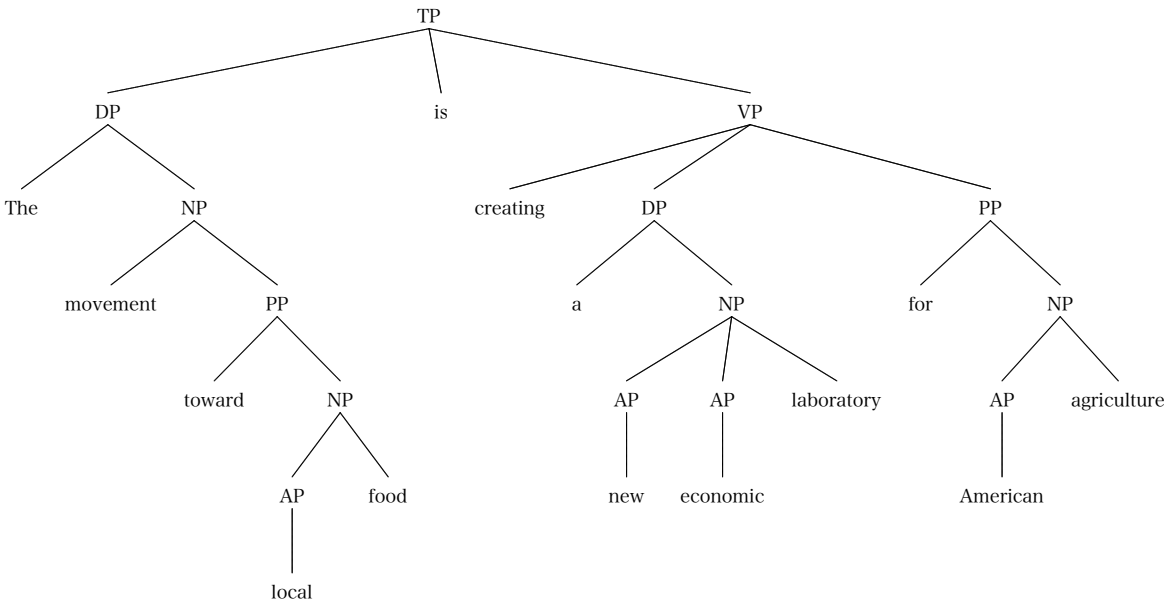
[22] lexemes are the canonical forms of words, the ones that are expected to be found in dictionaries (van der Plas, 2008). The opposite of a canonical form is a conjugated (or inflected) form.

local food is creating a new economic laboratory for American agriculture. :

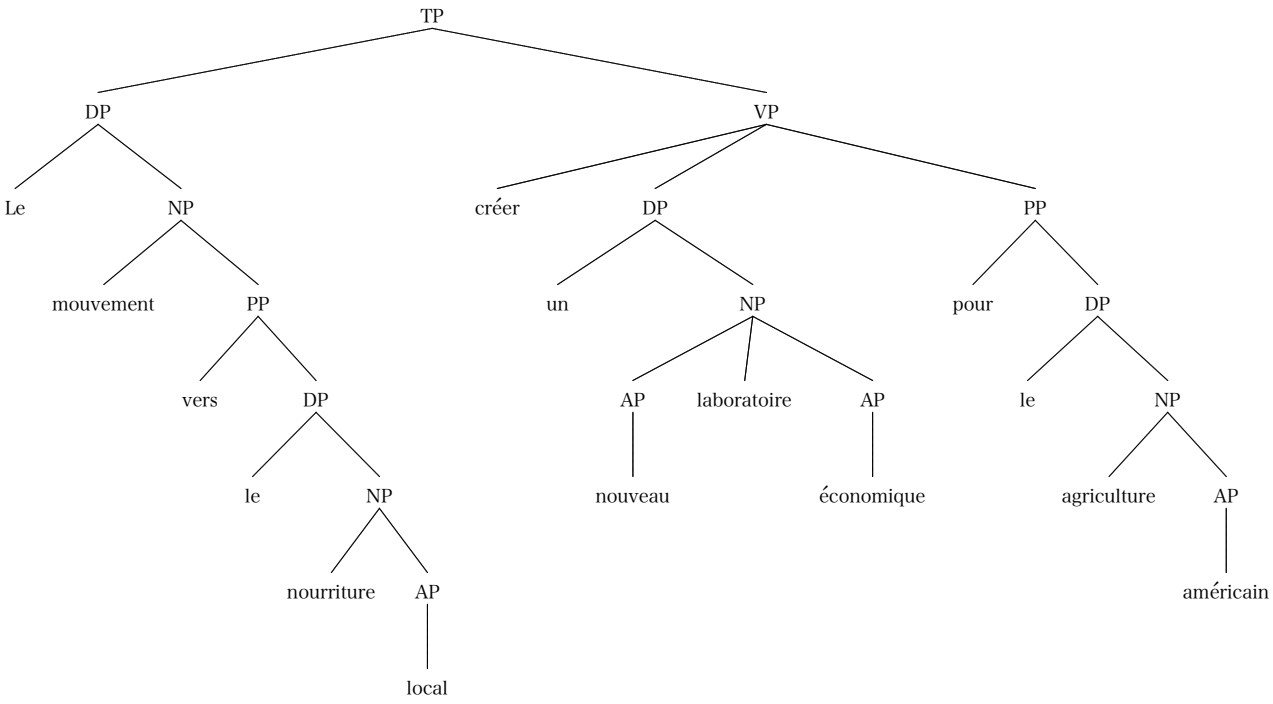
- First, the English input sentence is read:

The movement toward local food is creating a new economic laboratory for American agriculture.

- After syntactic parsing, an English tree structure is obtained:



- Then, in the transfer phase, a French tree structure is produced. We can notice that local reordering have been applied, changing the position of some adjectives:



- After the generation phase, the French output sentence is written:

Le mouvement vers la nourriture locale crée un nouveau laboratoire économique pour l'agriculture américaine.

Recent evolution

Recent research works involving Its-2 have dealt with collocation translation (Seretan and Wehrli (2007)), Western language translation with a yearly participation into the WMT competition (Wehrli et al., 2009a), and on statistical post-editing of the Its-2 output using Moses for an improvement in null-pronoun sentence translation (Russo et al., 2012). Its-2 is also used for sentence translation in TWIC, a word-in-context translation application that exists as a web browser plug-in or as a smartphone application (Wehrli et al. (2009b)).

2.4 Conclusion

In this chapter, after showing the fundamental principles of the Japanese writing system, we have described the evolution and state-of-the-art works in English-Japanese and multilingual MT, presenting the rich history and the wide panel of existing systems, including LBMT, SMT, EBMT, hybrid and SBSMT systems.

We have seen that all of these five paradigms can be used for the English-Japanese language pair, and that too basic RBMT and SMT techniques were less appropriate, due to the structural difference between these two languages.

We have described in more detail the Its-2 multilingual LBMT system, which will be used throughout this thesis. We will test the capacity of this system to adapt to the Japanese language for English-to-Japanese translation, and try to optimise the translation quality, using data obtained by statistical methods.

Chapter 3

Handling translation asymmetries at the lexical level

3.1 Introduction

Structural asymmetries cannot be treated in MT without rich and detailed lexical resources.

Lexical resources, such as bilingual or monolingual corpora and lexicons, play a very important role in MT. They contain the necessary information for a good translation quality. In this chapter, we discuss different aspects of the lexicon that have a strong impact on the treatment of English-Japanese structural asymmetries, such as verb subcategorisation, collocations and other multi-word expressions, or agreement features.

In the first section, we describe the different cases of translation asymmetries in English-Japanese MT and the information which is consequently expected from the lexical data in MT.

In the second section, we present the lexicons that have been built for the English-Japanese version of the Its-2 MT system.

In the last section, we focus on verb subcategorisation, presenting a method for verb subcategorisation mapping between two languages and applying it to the Its-2 English-Japanese lexicon.

3.2 Translation asymmetries at the lexical level

In this section, after mentioning the impact of lexical classification on linguistic-based machine translation, we will describe several cases of translation asymmetries, discussing the classification presented in (Dorr, 1994) and (Mahesh et al., 2005).

3.2.1 Impact of the lexical classification

Lexical classifications define word categories and subcategories. They are influenced by the type of grammars they rely on. In Japanese, as words are not delimited by spaces, the notion of *word* itself is influenced by the grammar. Word boundaries can vary depending on *word segmentation*, which is directly determined by the grammar.

Some grammars are intended to be fully language-specific. For example, (Hashimoto, 1934) described a grammar of Japanese with its own lexical classification. *Functional structural grammars* are also intended to be fully language-specific (see Saint-Jacques (1966) for Japanese). Language-specific grammars use language-specific lexical classifications. These classifications are supposed to be independent from standard Western classifications, and therefore to respect the specific lexical properties of the language.

Other grammars, such as Generative Grammar (GG), are intended to be adaptable to many (or all) types of languages (see Tsujimura (1996) for Japanese with GG). With these grammars, lexical classification is, as much as possible, using the same word categories for all languages. An approach of that kind is taken with the multilingual parser Fips and multilingual MT system Its-2, where a simplified version of GG is used, using related lexical classifications for all languages. This consistency in classification makes the comparison of lexical entities and syntactic structures between languages easier. From this comparison, structural asymmetries can be detected and studied.

We will now present a classification of structural asymmetries between English and Japanese, discussing Dorr's classification (1994), in which six categories of machine translation divergences have been defined: categorial, promotional (or demotional), conflational, structural, thematic, and lexical divergences.

3.2.2 Asymmetries in word category

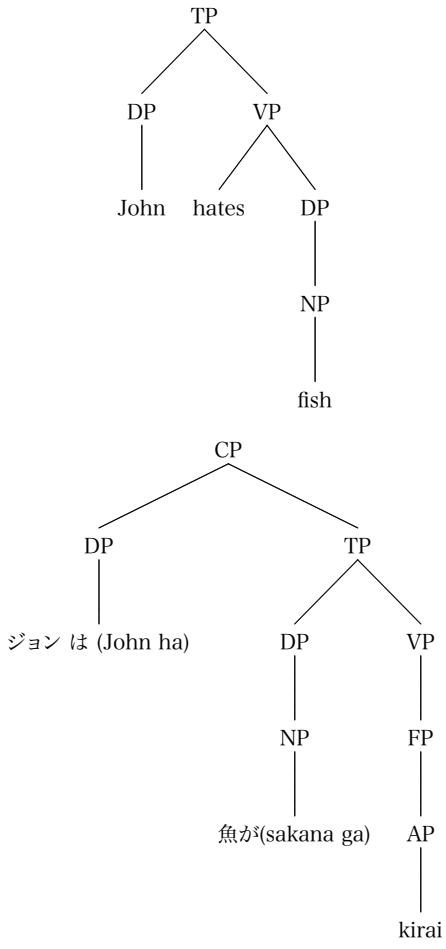
This phenomenon happens when a source word is translated into a word from a different category, causing an asymmetry between the source and target syntactic structure. This corresponds to *categorial divergences* in Dorr's classification.

For example, in the following English sentence, the verb *to detest* is used and translated into the adjective 嫌い ("kirai" : hateful) in Japanese. We will come back to the generation of Japanese adjectival clauses in Chapter 4.

(3.1) John hates fish.

ジョン	は	魚	が	嫌い。
John	ha	sakana	ga	kirai
John	[topic]	fish	[nominative]	hateful

literally: As for John, fish is hateful.



Specific cases where an adverb of the source language becomes a verb, hence being promoted to a central position in the target language verb phrase[1], have

[1]or in the reverse direction, when a verb becomes an adverb.

been classified by Door as *promotional divergences* or *demotional divergences*. In the following example, the adverb *too much* becomes the verb すぎる ("sugiru": overdo) in Japanese:

(3.2) She has eaten too much.

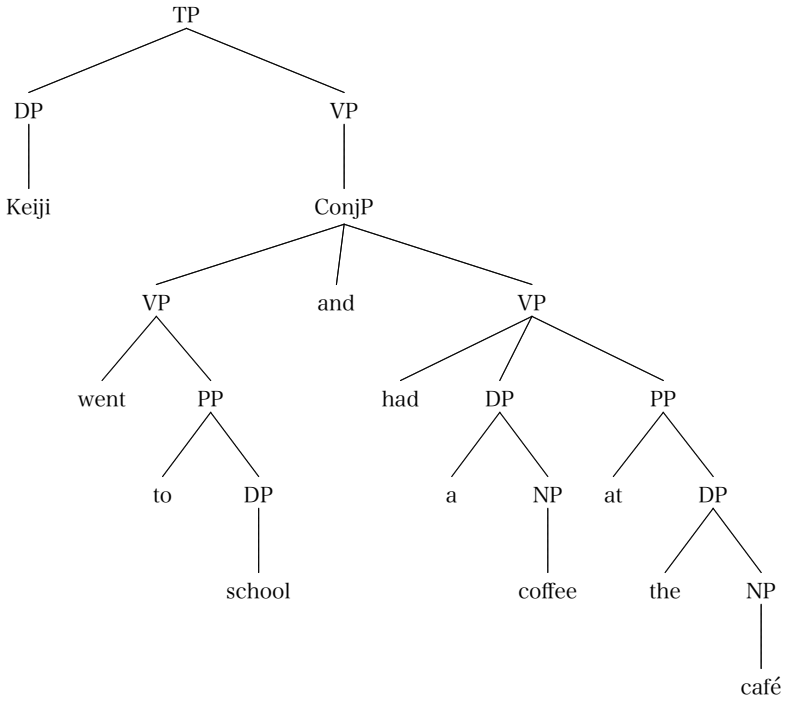
彼女	は	食べ	すぎた。
kanojo	ha	tabe	sugita.
she	[topic/subject]	eating	overdid

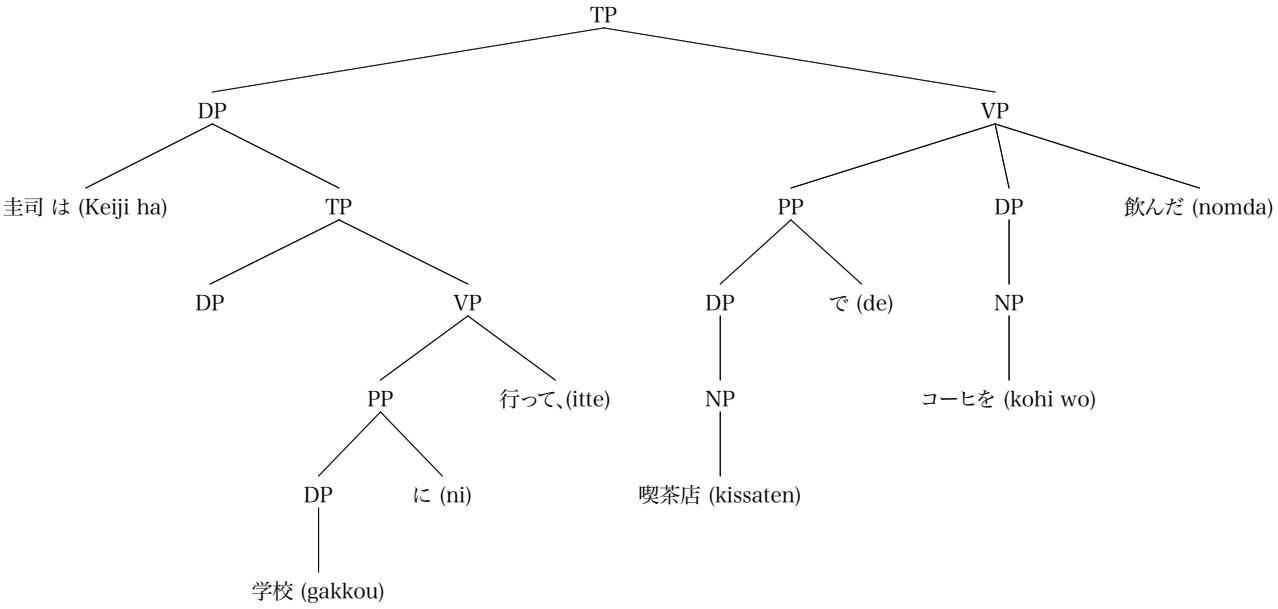
We must also consider cases where structural asymmetries are caused by the translation of a word that has no equivalent in the target language. In the following example (that was already presented in the introduction of this thesis), the coordination conjunction *and* has no identical equivalent for verbal clause coordination in Japanese. Its translation causes an asymmetry in syntactic structures and causes the selection of a gerundive form for the verb of the first clause in the Japanese sentence. We will come back to the translation of coordination in Chapter 6.

(3.3) Keiji went to school and drank a coffee at the café.

圭司	は	学校	に	行って、	喫茶店	で	コーヒ	を
keiji	ha	gakkou	ni	itte,	kisaten	de	kôhi	wo
keiji	[nominative]	school	to	go(gerund),	café	at	coffee	[object]

飲んだ。
nomda.
drank





In order to be able to generate asymmetrical translations with a transfer-based LBMT system such as Its-2, bilingual lexicons accept correspondences

between lexemes of different categories, when they are required.

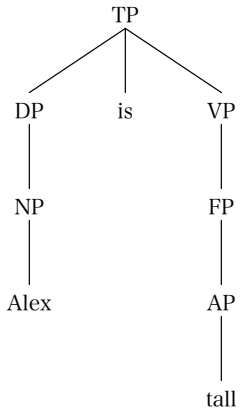
3.2.3 Asymmetries involving collocations or multi-word expressions

This phenomenon happens when a source word is translated into a collocation or a multi-word expression[2], causing an asymmetry between the source and target syntactic structure. It has been described as *conflational divergences* in Dorr's classification. In the following example, the multi-word expression 背 が 高い ("se ga takai": height (is) high) is used for the translation of the adjective *tall*.

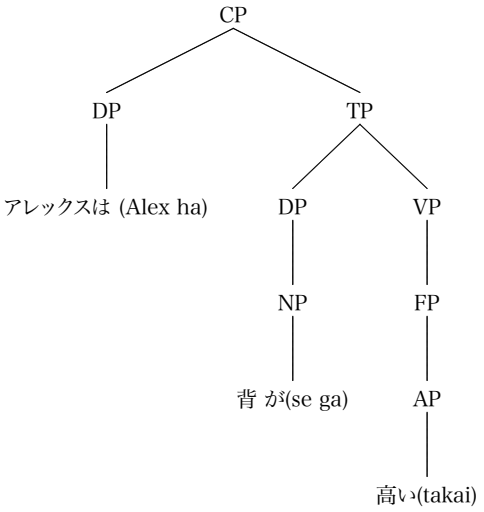
(3.4) Alex is tall.

アレックス	は	背	が	高い。
Alex	ha	se	ga	takai
Alex	[topic/subject]	height	[nominative]	high

literally: As for Alex, height is high.



[2]or when a collocation or multi-word expression is translated into a single word.



At the lexical level, large-scale monolingual lexicons of collocations and other multi-word expressions, and bilingual lexicons that record their translations are required for the treatment of conflational divergences. Such bilingual lexicons can be obtained making use of bilingual corpora, applying word alignment techniques that allow one word in one language to be aligned with several words in the other language (see Mochizuki et al. (2011)). Other techniques have shown good results in collocation-to-collocation translation detection, but have been less accurate in asymmetrical case detection (see Seretan and Wehrli (2007)).

3.2.4 Asymmetries in subcategorisation frame

Subcategorisation frames[3] define the syntactic behaviour of verbs with their arguments. They show if verbs are transitive or intransitive, if they take direct or indirect objects, if they can take sentential objects, etc.

Sometimes, the subcategorisations of the source and target verbs are different. This phenomenon has been classified by Dorr as *structural divergences* or *thematic divergences*. Thematic divergences happen when the argument that is in subject position in a language is moved to the object position in the other language, while structural divergences deal with other cases of argument structure asymmetries, as in example 3.5. In this example, the English *to ride* is a direct transitive verb, and it is translated into the indirect transitive verb ("noru") that requires the postposition に ("ni": on) after the object.

(3.5) Mr Shibata rides a bicycle.

[3]also referred as *verb valency frames* in the dependency grammar framework, or *predicate-argument structures*.

柴田	さん	は	自転車	に	乗る。
Shibata	san	ha	jitensha	ni	noru
Shibata	Mr	[topic/subject]	bicycle	at	rides

At the lexical level, a deep monolingual and bilingual knowledge about verb subcategorisation is required for a good translation of thematic and structural divergences. We will come back on this point in the last section of this chapter.

3.2.5 Asymmetries in semantic content

This phenomenon happens when words with different semantic meanings are found in the source and target sentences. It has been described by Dorr as *lexical divergences*. (Hutchins and Somers, 1992) described it as *translational ambiguities*, distinguishing *grammatical* and *conceptual* translational ambiguities.

For example, the Japanese verb 行く ("iku": to go) is used instead of 来る ("kuru": to come) when the subject is at the first person, in sentences where *to come* is used in English:

(3.6) I came to the university

私	は	大学	に	行った。
watashi	ha	daigaku	ni	itta
I	[topic/subject]	university	to	went

Asymmetries in semantic content do not necessarily cause structural asymmetries at the syntactic structure level, but they may have an impact on lexical or morphological selection. In the cases of underspecified input[4], they may lead to arbitrary or statistically-motivated lexical selection, especially when the pragmatic context of the source sentence is not known by the MT system.

In the next example, which is a typical case of conceptual translational ambiguity in Hutchin's description, we can see that two possible words exist in Japanese for the translation of *rice* : ご飯 ("gohan" : cooked rice) and 米 ("kome" : raw rice). Hence, in sentences 3.7.b or 3.7.c the Japanese translation will convey a nuance which may have been pragmatically implied but was not explicitly expressed in the source sentence.

(3.7) (a) I want to buy rice.

(b)

米	を	買いたい。
kome	wo	kaitai
rice (raw)	[object]	want to buy

[4]When the target sentence requires semantic knowledge which is not explicitly expressed in the source sentence, it is a case of *underspecified input*.

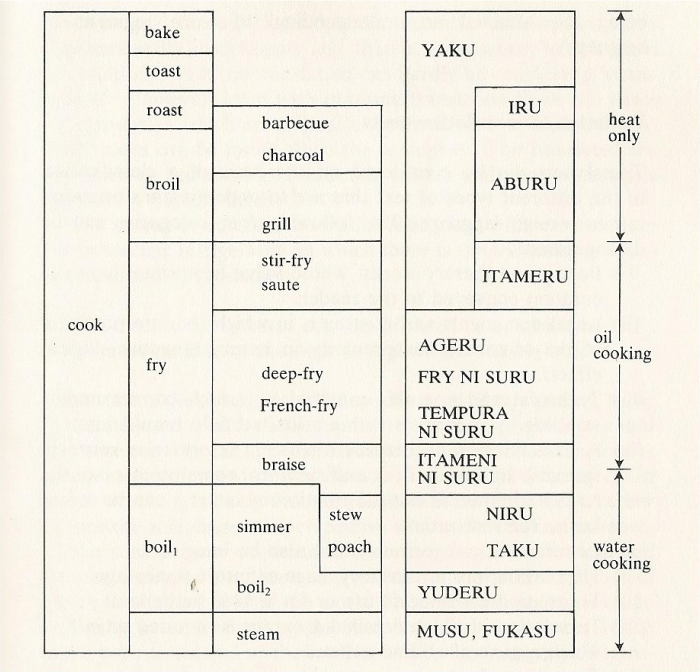


Figure 3.1: Possible translations of *cook* in Japanese, as cited in (Nagao, 1989)

(c)	ご飯	を	買いたい。
	gohan	wo	kaitai
	rice (cooked)	[object]	want to buy

Another example of the same kind, given in (Hutchins and Somers, 1992) and in (Blanchon et al., 2006), is the English noun *wall* (or French noun *mur*) which, in German, must be translated into *Wand* when it is inside a building, or *Mauer* when it is outside. Figure 3.1 was cited in (Nagao, 1989) and shows how complex semantic asymmetries can get, describing different possible translations of the English word *cook* into Japanese.

Semantic asymmetries can also be found at the morphological level. As shown in (Miyagawa, 2005), while agreement features are essential in Western languages, politeness expression is an essential feature of Japanese and Korean, influencing verbal morphology at all tense forms. Hence, the level of politeness often remains an underspecified parameter in English-to-Japanese MT. As (Ma-

hesh et al., 2005) mentioned it, this aspect of translation asymmetries has not been pointed out in Dorr's work. In the following example, we can see that the same English sentence can take three different Japanese translations, depending on the politeness level: simple, polite or honorific form. The verb at the simple form たべる ("taberu": to eat), becomes たべます ("tabemasu") at the polite form and 召し上がります ("meshi agarimasu") at the honorific form.

(3.8) (a) She eats an apple.

(b)

彼女	は	林檎	を	たべる。
kanojo	ha	ringo	wo	taberu
she	[topic/subject]	apple	[object]	eats (simple form)

(c)

彼女	は	林檎	を	たべます。
kanojo	ha	ringo	wo	tabemasu
she	[topic/subject]	apple	[object]	eats (polite form)

(c)

彼女	は	林檎	を	召し上がります。
kanojo	ha	ringo	wo	meshi agarimasu
she	[topic/subject]	apple	[object]	eats (honorific form)

At the lexical level, the indication of the politeness level in the classification of the verb forms, pronouns and politeness-related vocabulary is a required step for the generation of the translation in a specific politeness level, as in example 3.8. Asymmetries in semantic content such as example 3.6 can be learned from aligned corpora and recorded in lexicon. As much as possible, a semantic or statistical selection should be applied on ambiguous cases such as example 3.7 during the transfer phase, in order to reduce the frequency of incorrect translations.

3.3 Application: building large-scale lexicons for MT

In this section, we will describe the development of the English monolingual and English-Japanese bilingual lexicons for the Its-2 MT system.

3.3.1 Defining the Japanese lexicon classification

Several Japanese lexicon classifications have been defined. A historically important classification was presented in (Hashimoto, 1934). It has had a lot of impact on the lexical classification and grammar taught in Japanese schools nowadays (see Nakamura-Delloye (2003), p.8). The classification presented in (Nagara, 1990) was meant to be more modern and more accessible to foreign learners of Japanese (Nakamura-Delloye (2003), p.17). The one proposed in

(Masuoka and Takubo, 1992) has been used for natural language processing applications such as the morpho-lexical analyser Juman and the syntactic parser KNP (see Kurohashi and Nagao (2003)). The CJKI Japanese lexical database (Halpern (2008c); Halpern (2008a); Halpern (2008b)), which is based on its own lexical classification, has been used in this work.

We have defined a classification which contains the same word categories as classifications of other languages translated by Its-2. In this classification, the copula has been considered as a particular type of verb; postpositional particles have been considered as postpositions and other particles have been put in the particle category and classified by types; counter words (see Figure A.1 in Section A.0.1) have been considered as a particular type of noun. The classification is shown in Table 3.2. A more detailed description of the word categories and their properties is given in Appendix A.

3.3.2 Japanese monolingual lexicon

Related work

Among existing monolingual lexicons, the Japanese thesaurus "Bunrui Goihyou" by the National Institute for Japanese language (NIJL, 2009) contains a semantic classification of 101'070 Japanese lexemes. It was published for the first time in 1964 and has been updated since.

Nihongo Goi Taikei by NTT (Shirai et al., 1998) is a Japanese lexicon with semantic classification. It contains about 400'000 entries. This number includes orthographic variants and about 200'000 proper names. It was designed namely for machine translation and consequently, it has been used by MT systems.

The EDR electronic dictionaries (NICT, 2009) are a set of Japanese and English electronic dictionaries. Among their Japanese monolingual dictionaries, there are a word dictionary, a thesaurus and a technical dictionary. The word dictionary contains about 270'000 words.

WWWJDIC is a free dictionary website to which users can submit new entries (Breen, 2009). It is dedicated to the Japanese language. It contains a general vocabulary Japanese-English dictionary called EDICT, the proper names dictionary ENAMDICT, specialised Japanese-English dictionaries, and smaller bilingual dictionaries giving word translation from Japanese to other languages. The EDICT dictionary file contains about 120'000 Japanese lexemes and their possible translations. It also gives detailed information about the lexemes and in this way, it can be considered equivalent to a monolingual Japanese dictionary. The ENAMDICT contains 714'630 Japanese name entries: mainly place names, first names and last names, but also company and organisation names. Every entry is classified by its semantic category. The gender of first names is always given.

	category	**type**	**subtype**	**inflection**
words	noun	common proper pronoun counter	 demonstrative Wh personal	
	verb			group 1 ("godan") group 2 ("ichidan") group 3 ("kuru", "suru") copula
	adjective			"-i" "-na" "-no"
	adverb	standard sentential negative		
	determiner	demonstrative interrogative quantitative numeral		
	conjunction or conjunctive particle	 subordination coordination coordination	 clause coordination noun coordination	
	postposition			
	particle	case topic adverbial nominalisation explanation sentence final		
	interjection			
non-words	prefix suffix			

Figure 3.2: *Its-2 Japanese lexicon classification*

The CJKI's Japanese lexical database has been created at the CJK Dictionary Institute. It stores Japanese lexemes, prefixes, affixes and some expressions. It has been designed for natural language processing. Every entry is stored with information related to its pronunciation, morphological category, syntactic behaviour, etc. It is not a dictionary for human reading and does not contain any synonyms, definitions, explanations, example sentences, etc. The version used here includes 232'990 entries, counting one entry per word pronunciation (which makes a higher number of entries than counting one entry per lexeme and regrouping the several possible pronunciations).

Monolingual Japanese lexical information can also be found in bilingual Japanese-English electronic dictionaries, such as the Eijiro dictionary (Eijiro, 2012).

Method

Following the defined lexical classification, a Japanese monolingual full form lexicon has been built, using data extracted from two existing lexicon: the CJKI monolingual Japanese lexicon and the ENAMDICT. Proper nouns have been extracted from the ENAMDICT and other words have been extracted from the CJKI monolingual Japanese lexicon.

The work has been divided into five steps: comparison of the lexical classifications, data extraction (for each category), morphological generation of inflected word forms (for verbs and adjectives), addition of syntactic or semantic information, insertion of all the lexemes and inflected word forms.

The comparison of lexical classifications has shown that CJKI Japanese lexical database classification is quite similar to the one defined for Its-2. However, some differences remain: some lexemes have been given a category in the CJKI Japanese lexical database and another one in the Its-2 Japanese lexicon ; some lexical subcategories exist in the CJKI Japanese lexical database and do not in the Its-2 Japanese lexicon, where syntactic or semantic properties have been stored in additional flags.

The data extraction has been achieved using a series of Perl procedures. The differences in lexical classification have been taken into account and the extracted data has been converted into the Its-2 lexical database format.

In Japanese, only verbs and adjectives ending in -い ("-i") have inflected forms, while all the other word categories are invariable (Kauffmann (2008b)). Therefore, the morphological generation has consisted in generating 162 forms for each verb and 18 for each adjective of the lexicon[5]. This task has been

[5]Among the conjugated forms, only simple tense forms have been considered. Compound conjugations (such as verb+semi-auxiliary, adjective+copula, etc.) have not been included in the full form lexicon.

carried out with a Component Pascal program running under the Blackbox environment (Oberon, 2001).

Syntactic or semantic information has been added to some of the lexical entries. For example, person, number and gender of personal pronouns were specified when it was possible (see Table A.2); syntactic properties of conjunctive words have been specified (see Chapter 6); information has been given about verb subcategorisation (see Section 3.4). As we will see throughout the thesis, syntactic information stored in the lexicon is useful for the treatment of structural asymmetries, because they help to generate the target phrases independently from the source phrase structure. Verbs, postpositions, and nouns have been given semantic properties. For example, the *time* semantic feature to time nouns like 月曜日 ("getsuyoubi" : Monday). Semantic information can influence lexical selection and is then especially useful for the treatment of lexical divergences.

The insertion has consisted in inserting lexemes in the Its-2 Japanese lexeme database and inflected word forms in the Its-2 Japanese word database, using SQL queries.

Results

We have obtained a Japanese monolingual lexicon divided into two databases: the lexeme database and the word database. The lexeme database contains 922'157 lexemes, in which 714'640 are proper names and 207'517 are lexemes of other lexical categories. The word database contains 6'140'367 morphological forms, including canonical and inflected forms. This amount of data should be sufficient to ensure a good coverage of Japanese general vocabulary.

Furthermore, syntactic and semantic information about the lexeme properties has been added.

The structure of the lexical database also enables collocations and other multi-word expressions to be added, provided that the lexemes found in the expressions have previously been entered in the database. Further work in this direction would be very useful for the treatment of conflational and lexical divergences.

3.3.3 English-Japanese bilingual lexicon

Related work

The Eijiro dictionary (also called Electronic Dictionary Project) is an English-Japanese and Japanese-English electronic dictionary file that includes impressive data about coocurrence and collocation translation, with examples of aligned sentences showing in which context coocurrences are found. It contains a total

of about 7'000'000 entries (Eijiro, 2012). It has a free online version available on the ALC space website (ALC, 2012).

The EDR electronic dictionary contains both Japanese-English and English-Japanese dictionary files. The WWWJDIC contains Japanese-to-English word translations, but no information about English-to-Japanese translation.

The CJKI English-Japanese dictionary contains 80'676 entries. Each entry includes an English lexeme with its different possible Japanese translations and its lexical category. Japanese translations are grouped by senses. Some of the English lexemes are proper nouns. The CJKI Japanese-English dictionary of general vocabulary contains 108'964 entries. Each entry includes a Japanese lexeme with its different possible English translations, its default pronunciation, and its lexical category. There are no proper nouns in this electronic dictionary.

Transfer-based Japanese-English or English-Japanese LBMT systems (such as Yakuse Goma, Yakushite, The Honyaku, etc...) all contain large-scale bilingual lexicons. For example, The Honyaku V15 contains bilingual lexicons, with 4'850'000 entries, and translation memories, with 250'000 pre-translated sentences.

Atlas V14 interlingual MT system also contains bilingual lexicons, with 8'450'000 entries (2'880'000 entries of general vocabulary and 5'570'000 entries of technical vocabulary). Its translation memories contain 840'000 pre-translated sentences.

Method

An English-Japanese and Japanese-English lexicon has been built, using data extracted from the ENAMDICT for proper nouns, and from the CJKI Japanese-English and English-Japanese lexical databases for all the other word categories.

Most of the monolingual items found in these bilingual dictionary files can also be found in the Japanese and English monolingual lexicons of Its-2. Hence, bilingual correspondences between known monolingual items have been extracted and inserted into the Its-2 English-Japanese bilingual lexicon. The extraction and insertion has been achieved using Perl procedures and SQL queries.

Structure of the lexicon

The structure of the Its-2 English-Japanese lexicon is similar to the other bilingual lexicons of Its-2 : bilingual correspondences between monolingual items (lexemes, multi-word expressions or collocations) are stored and ranked, as in a bilingual dictionary from the most current to the most common use to the most unusual, for each translation direction.

Every entry of the English-Japanese lexicon contains an English monolingual item and its Japanese equivalent. The monolingual items that are found in the

bilingual lexicon must have previously been recorded the monolingual lexicons. Every entry also contains a score (between 1 and 6) indicating the preference of this correspondence for the English to Japanese translation, and another one indicating the preference for the Japanese to English translation. The highest score (6) has been given to the translations located in first position in the CJKI bilingual dictionaries, because these dictionaries have put the most common translations in first rank.

Moreover, for the monolingual items that take arguments, the translation and possible reordering of the argument structure is recorded.

Results

We have obtained a Japanese-English bilingual lexicon for Japanese-to-English and English-to-Japanese translation, containing 117'354 bilingual pairs of monolingual items.

The lexicon allows correspondences between words of different categories, hence allowing categorial, demotional and promonotional divergences. With the information argument structure translation, it also allows structural or thematic divergences (see next section about verb subcategorisation). Sometimes, translation asymmetries convey both structural and categorial divergences, such as verb-to-adjective translations with asymmetrical argument structure. Such cases can also be recorded in the lexicon.

3.4 Improving verb subcategorisation classification

In this section, we focus on verb subcategorisation in the lexicons, an essential question for the treatment of structural and thematic divergences, which was described in (Hino et al., 2011).

We will present a method for the detection of subcategorised verb bilingual correspondences, using monolingual data and basic bilingual correspondences. We will first describe the method, then we will show how information from a dedicated lexical file called Case Frames has been extracted, how it has been inserted into the Japanese monolingual lexicon and how this method has been applied to the Its-2 bilingual lexicon. We will finally describe the insertion of the subcategorised verb correspondences into the bilingual lexicon and show the results of the experiment.

3.4.1 Related work

The compilation of lexical data about Japanese verb subcategorisation has been an important research subject in MT and NLP. We notice the presence of such data in MT systems developed since the eighties (Wilks (2009), p. 128-130), as well as in more recent research (Komachi et al. (2006); Kawahara and Kurohashi (2010); Sasano and Kurohashi (2011); Hino et al. (2011)).

In this work, we have made use of the Case Frames file (see Kawahara and Kurohashi (2006)), an electronic dictionary file that stores Japanese verbs (or adjectives) and their subcategorisation frames. It currently contains 90'000 entries and has been computed from a text corpus extracted from the web. For each verb, it gives a list of subcategorisations (*case frames*[6]). For each case frame, it shows the number of occurrences in the corpus and the case particle or postpositional particle of every argument. For each argument, it also gives counts about the thematic roles, and a list of nouns frequently found in the argument positions. From this data, typical noun-verb coocurrences can be deduced.

For the use we would have of the file, the noun lists would not have been necessary. So, instead of using the original full version of this dictionary file, we have chosen a reduced version, without the noun lists. This reduced version contained about 31'000 verbs or adjectives, their different subcategorisations and their number of occurrences (see Japanese verb case structure in Table 3.3).

3.4.2 A method for bilingual verb subcategorisation detection

When lexical data is obtained for monolingual verb subcategorisation and for bilingual verb-to-verb translation, verb subcategorisation mapping between the two languages is a remaining problem.

As much as possible, correspondences between subcategorised verbs should always lead to equivalent verbal arguments (Table 3.3 shows some English verb subcategorisations and their Japanese equivalents, as they were recorded in the Its-2 English lexicon and in the Case Frames file.).

Equivalent arguments often have a similar function in both languages. However, in asymmetrical subcategorisation cases (structural divergences or thematic divergences), they have different syntactic values: for example, an indirect object in a language can become a direct object in the other language, as in example 3.5.

[6]The appellation *case frames*, used in Kawahara and Kurohashi (2006) is especially relevant for Japanese, where *case particles* are often affixed to the verb arguments.

	Verb	Argument 1 (Sujet)	Arg. 2	Arg. 3
English verb subcategorisation	go	NP		
	go	NP	PP(to)	
	go	NP	PP(from)	PP (to)
	write	NP	NP	
Japanese verb subcategorisation	行く ("iku")	が (ga)		
	行く (iku)	が (ga)	に (ni)	
	行く (iku)	が (ga)	から (kara)	まで (made)
	書く (kaku)	が (ga)	を (o)	

Figure 3.3: Verb subcategorisation in English and Japanese

In order to find the correspondences between subcategorised verbs, we have set up a method. Using monolingual subcategorisation data and a list of correspondences between non subcategorised verbs, this method enables to find most of the subcategorised verb correspondences automatically. It consists of three steps:

1- selection of candidate pairs by heuristical rules
2- division into two groups: safe and unsafe correspondences.
3- validation, correction or elimination of unsafe correspondences by a human corrector.

Figure 3.4: *detection and validation of bilingual correspondences of subcategorised verbs*

The first step consists in launching an automatic selection of exact correspondences between English and Japanese subcategorised verbs, using data from the two monolingual lexicons, for the verb pairs of the bilingual lexicon. This selection is based on a series of heuristic rules, implemented in SQL queries. For the verb pairs indicated in the bilingual lexicon, correspondences between subcategorisations whose arguments satisfy the conditions shown in Table3.5 are automatically validated. For example, correspondences between two direct transitive verbs are always validated; correspondences between a direct transitive verb A and an indirect transitive verb B are validated only in cases where there is no existing direct transitive subcategorisation of verb B; etc.

In the second step, two types are distinguished among the automatically de-

English verb arguments		Japanese verb arguments		Required properties
Arg. 2	Arg. 3	Arg. 2	Arg. 3	
NP		NP (を″ o″ or sometimes が″ga″)		
S		completive S (と″to″)		
S		nominalised S (のを″no o″ or のが″no ga″)		no completive S (と″to″) in Japanese arg2 position
PP		PP (に″ni″, で″de″, へ″he″...)		to be checked manually
NP		PP (に″ni″, で″de″, へ″he″...)		no direct object NP in Japanese arg2 position
PP		NP (を″o″ or sometimes が″ga″)		no direct object NP in English arg2 position
NP	PP	NP (を″o″ or sometimes が″ga″)	PP (に″ni″, で″de″, へ″he″...)	to be checked manually
PP	PP	PP (に″ni″, で″de″, へ″he″...)	PP (に″ni″, で″de″, へ″he″...)	to be checked manually
PP	S	PP (に″ni″, で″de″, へ″he″...)	completive S (と″to″)	
PP	S	PP (に″ni″, で″de″, へ″he″...)	nominalised S (のを″no o″ or のが″no ga″)	no completive S (と″to″) in Japanese arg3 position

Figure 3.5: Automatic validation of subcategorisation correspondences

tected pairs: the *safe* ones that are fully validated, and the *unsafe* ones, semantically ambiguous, that will need to be manually verified and maybe corrected for a full validation.

The last step of the selection process consists in correcting manually the unsafe pairs.

3.4.3 Implementation

Improving monolingual verb subcategorisation

The first step consisted in extracting the lexical data from the Case Frame file and inserting it into the Japanese monolingual lexicon.

In lexical databases of Its-2, each verbal lexeme corresponds to a single subcategorisation (see Wehrli et al. (2009a)). Thus, if a verb has, for example, five

different possible subcategorisations, then five verbal lexemes are recorded in the monolingual lexicon. The Its-2 monolingual English lexicon already contained a detailed classification of verb subcategorisation (see English verb subcategorisation in Table 3.3), with about 16'000 verbal lexeme entries recorded. However, such information was absent in the Japanese monolingual lexicon, which only contained incomplete subcategorisation data on verbs. Information extracted from CJKI dictionary file only indicated if a verb was *transitive*, *intransitive*, or both.

In order to store more information about Japanese verb subcategorisation, we have extracted data from the Case Frames file. By a series of SQL queries, we have assigned argument structures found in the Case Frames file to the verbs of the Its-2 Japanese monolingual lexicon and we formatted the data to the lexicon format. Most Japanese case frames could be directly converted into our subcategorisation framework, but some were slightly ambiguous. For example, the particle が ("ga") and に ("ni") can either correspond respectively to a subject and a prepositional object, or to a direct object and a subject[7]. So, we looked at the *agent* ratio given in the Case Frames file, that indicates the ratio of *agent* thematic roles among arguments with the same case particle for the same subcategorisation. In the cases where the agent ratio of the が ("ga") argument was high enough, we have given the *subject* function to the が ("ga") argument and the postpositional object function to the に ("ni") argument. Otherwise, we have given the *subject* function to the に ("ni") argument and the direct object function to the が ("ga") argument.

Finally, we have selected subcategorisations that had not been previously recorded in the lexicon and inserted them. Thus, we have added 5000 Japanese verb subcategorisations to the 2500 ones that were already stored in the lexicon, reaching a total of 7500 subcategorised verbs.

Bilingual verb subcategorisation detection

We have applied the method presented above to the English-Japanese bilingual lexicon of Its-2.

In the first step, the automatic selection of bilingual subcategorised verb correspondences has been launched, using the English and Japanese monolingual lexicons and the verb pairs from the English-Japanese bilingual lexicon.

In the second step, unsafe correspondences have been separated from safe correspondences. 70% of the detected subcategorised verb pairs were safe. The other 30% have been separated from the corpus before being manually corrected.

In the last step, manual corrections have been given to the unsafe bilingual correspondences. These were the ones containing at least one prepositional

[7]In those cases に ("ni") is often replaced by the topic particle は ("ha"). But, in order to avoid ambiguity, topic particles have not been included in the Case Frames file argumental structures.

argument for the English verb and one postpositional argument for the Japanese verb. A Japanese native speaker with a good knowledge of English has validated or corrected the 2403 unsafe bilingual correspondences. Finally, 1028 of these were deleted, 718 were corrected, and 657 were accepted without modification. The sum of the safe correspondences and the manually corrected ones resulted in a total of about 7000 pairs of subcategorised verbs.

Insertion of the new data into the bilingual lexicon

After the automatic detection of bilingual correspondences and manual correction of unsafe detected pairs, the bilingual lexicon has been updated with the new pairs of lexemes.

Once the list of bilingual subcategorisation correspondences was established, we have compared it with those already present in the lexicon, in order to find which ones among the list were already in the lexicon and which ones had to be added. Moreover, we had to determine which correspondences in the lexicon were wrong and needed to be deleted.

We also had to adjust the correspondence scores. When two Japanese subcategorisations were possible for the translation of the same English subcategorisation, the best score has been given to the one that had the largest number of occurrences in the Case Frames file.

Then, the data that would be added to the lexicon had to fit the required format. So, the data has been formatted, false entries previously present in the lexicon have been deleted and the new data has been inserted.

3.4.4 Results

We have been able to detect about 8000 subcategorised verb bilingual correspondences automatically, knowing that 30% were unsafe and would need a manual verification and considering that the other 70% were correct. On this basis, the verification of the unsafe ones showed that about 27% of them were indeed correct, 30% needed to be slightly modified and 43% were false.

Then, the revised collection of 7000 subcategorised verb bilingual correspondences has been used for the improvement the English-Japanese bilingual lexicon: the new correspondences have been inserted into the bilingual lexicon and the erroneous ones have been removed. This improvement has been useful. It has enabled a better translation of prepositions in the verb phrase (see chapter 4, section 5.2) and a more correct generation of verbal or sentential objects (see chapter 7).

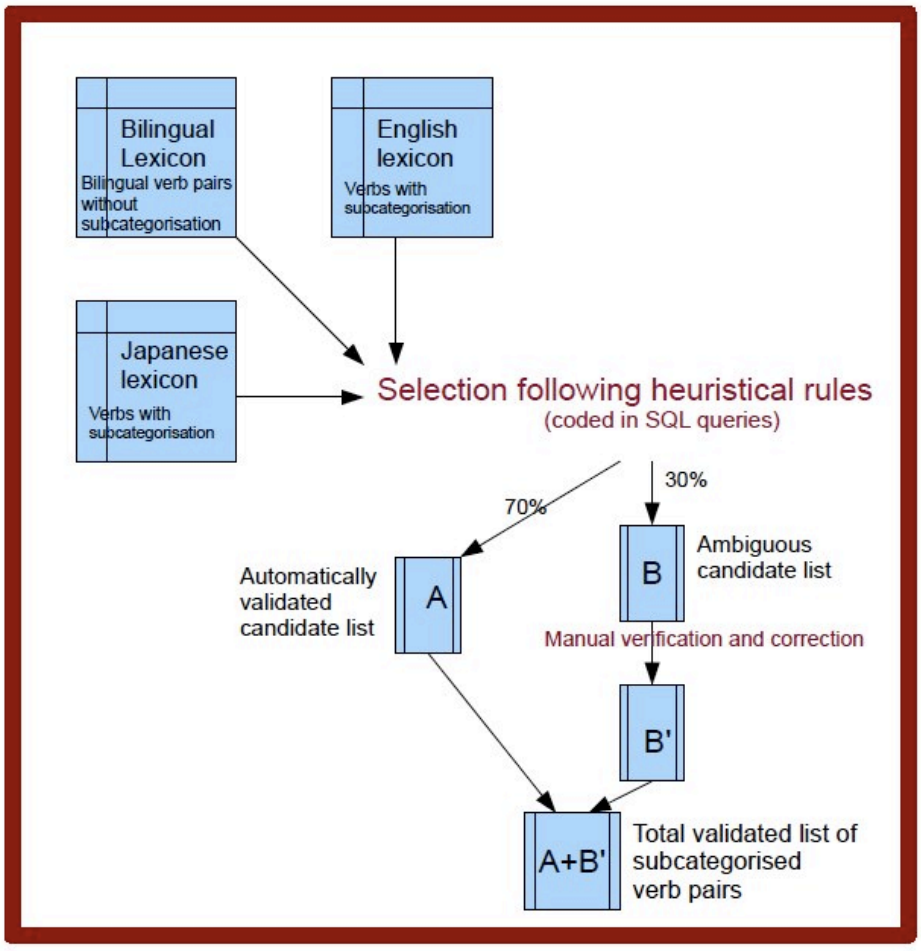


Figure 3.6: *Detection and validation of bilingual correspondences of subcategorised verbs*

3.5 Conclusion

We have succeeded in building large and robust lexicons for MT. After the improvement of the data about verb subcategorisation, the Its-2 Japanese lexicon contains 927'480 lexemes (714'650 proper nouns and 212'830 other lexemes) and the Its-2 English-Japanese lexicon contains 124'837 bilingual correspondence. They offer a good coverage of the Japanese general vocabulary and of the English-Japanese bilingual correspondences. They contain useful lexical data for the treatment of structural asymmetries, including categorial, demotional and promonotional divergences, structural and thematic divergences and lexical divergences.

However, improvements can be still be given to these lexicons in future work, especially for collocation and multi-word expression translation. This would require adding a significant number of Japanese collocations, passive collocations[8] and multi-word expressions, and their English translations in the lexical database. Further work in this direction would be very useful for the treatment of conflational and lexical divergences.

[8] A *passive collocation* is a collocation which is usually not used in the target language, but that makes sense for the translation of a word (or collocation) of the source language, when this word has no equivalent in the usual target language vocabulary.

Chapter 4

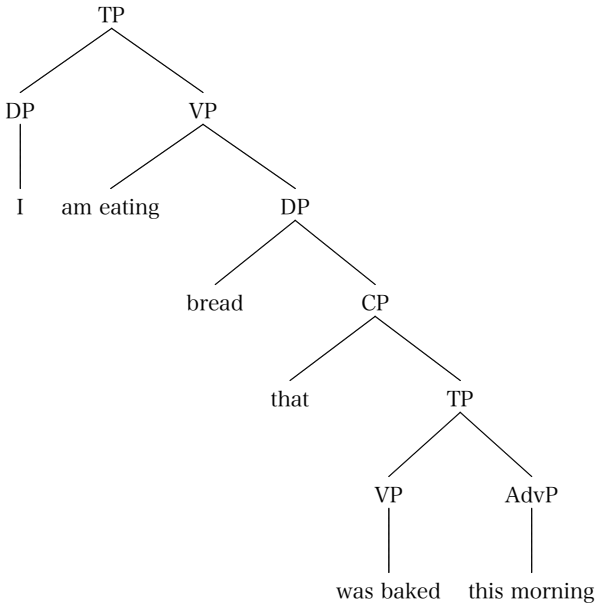
Word reordering and translation of simple sentences

4.1 Introduction

Every language has a typical word order and constituent[1] order. The difference in word and constituent order is one of the most evident structural differences between English and Japanese, and word and constituent reordering is a fundamental stage in the process of translation between the two languages. Example 4.1 illustrates the contrast of the SVO (Subject Verb Object) order in English with the SOV (Subject Object Verb) order in Japanese: in English, the verb comes before the object, while it comes at the end of the verbal phrase in Japanese. Moreover, we can see in example 4.1 that the relative clause comes before the noun in Japanese, and after the noun in English.

(4.1) (a) I am eating bread that was baked this morning

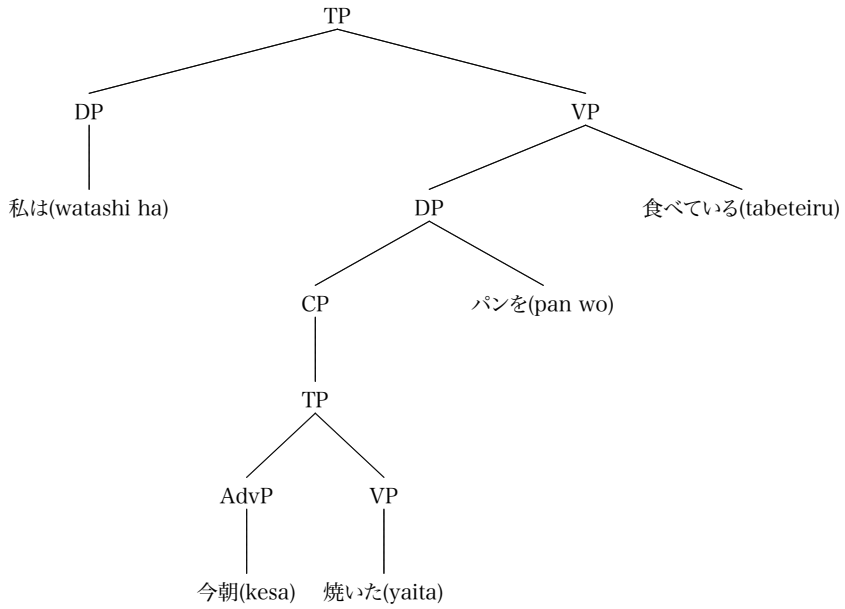
[1] A constituent is a group of words that forms a functional unit of a grammatical construction, as a verb phrase, determiner phrase, etc.



(b)

私	は	今朝	焼いた	パン	を	食べている。
watashi	ha	kesa	yaita	pan	wo	tabeteiru
I	[topic/subject]	this morning	baked	bread	[object]	am eating

I am eating bread that was baked this morning



Lexical selection is another issue that must be dealt with during the simple sentence translation process. It is a major issue, because any incorrect selection

in case of translational ambiguity may lead to a semantic or syntactic error in the output (see (Hutchins and Somers, 1992), pp. 99-102).

In this chapter, we will present the different steps of the reordering and translation of simple sentences. First, we will explain the reordering needed for the translation of a determiner phrase (DP), and for the translation of complementizer phrase (CP) or a prepositional phrase (PP). Then, we will describe the translation of a simple sentence. Finally, we will describe rules that have been implemented for reordering and for lexical selection in context.

4.2 Reordering and translation of the determiner phrase

The Japanese determiner phrase (DP)[2] is organised in three parts: the determiner, when there is one, comes first, adjectives or other noun modifiers are found in the middle part, and the noun comes in the end.

The determiner comes in first position in the DP in both English and Japanese. However, definite and indefinite articles do not exist in Japanese. Hence, DPs starting with definite or indefinite articles in English should be translated into null article DPs. Demonstrative, interrogative or possessive determiners are the only determiner that must actually be translated into Japanese. We also have to notice that second person pronouns[3] are seldom used in Japanese. They are often replaced by the name of the person who is addressed.

(4.2) Where is your bicycle?

鈴木さんの	自転車	は	どこ	です	か。
suzuki san no	jidensha	ha	doko	desu	ka
Mr Suzuki's	bicycle	[topic]	where	is it	[question]

(literally: Where is Mr Suzuki's bicycle?)

Because such replacement is impossible to achieve when the context remains unknown to the translation system, we have chosen to generate sentences with あなたの ("anata no" : your). Still, researches in context analysis have been conducted, especially in Japanese-English MT (see Bond (2005)) and Japanese analysis (see Sasano and Kurohashi (2011)). Replacing the second person possessive determiner by a name in the genitive case when it is possible would be a possible improvement to the system in future work.

[2]We will consider here that the determiner is the syntactic head of the noun phrase, even if the noun is the semantic head, following the *DP hypothesis* (Abney, 1987). So we can call it *DP* (Determiner Phrase), even for cases where there is no determiner before the noun, where we consider a *null determiner*.

[3]In Japanese, possessive determiners are personal pronouns followed by the genitive case particle の ("no") (see section A.6).

Then, in the Japanese DP, between the determiner and the noun, adjectives, qualifying prepositional phrases, relative clause or modifying noun can be found. This is a difference with English where the relative clause always comes after the noun (see examples 4.1.a and 4.1.b). Reordering must be achieved in the translation process in order to handle these differences.

Finally, at the right end of the DP, the noun comes. When the DP is an argument of a non-stative verb or the subject of a stative verb, it has to be given a case particle, a postpositional particle or a topic particle, which will be added as a suffix to the noun (see section 3.2.3). As case and topic particle do not exist in English (except the Saxon genitive marker *'s*), they have to be added in the output sentence when it is generated.

4.3 Prepositional phrase and complementizer phrase translation

English prepositions are translated into Japanese postpositions. They follow noun phrases at the end of postpositional phrases (PPs) (instead of preceding noun phrases in PPs, like English prepositions do).

(4.3) by bus

バス	で
basu	de
bus	by

The same phenomenon happens with complementizers[4] They come at the end of CPs. Complementizers appear in adverbial subordinate clause and completive clauses, as we will see in chapters 6 and 7.

In Japanese relative clauses, relativizers are not expressed. The English relative pronoun is translated into a null relativizer then (see examples 4.1.a and 4.1.b).

Reordering is not the only issue related to the translation of prepositions and complementizers. It can be noticed that English prepositions may also be used as complementizers:

(4.4) The time after the war was difficult.

(4.5) I will come after the war ends.

[4] *Complementizer* is the term used in generative grammar equivalent to subordination conjunction or a relative pronoun, in a CP (complementizer phrase). The term equivalent to a relative pronoun in a relative clause is *relativizer*.

In Japanese, prepositions and complementizers are clearly distinct. For example, *after* can be translated into 後で ("ato de"), 後に ("ato ni"),or 後 ("nochi") when it is a complementizer, and 後の ("go no"), の後の ("no ato no"), の後で ("no ato de"), の後に ("no ato ni") when it is a preposition.

In addition, English prepositional phrases can either be used as a noun complement (as in example 4.4) or as a verb complement (as in example 4.6).

(4.6) I will come after the war.

Most Japanese postpositional particles have a slightly different form depending on whether they are used in a verb complement or a noun qualifier. Among the translation of the preposition *after*, の後で ("no ato de") の後に ("no ato ni") are used in adverbial phrases, whereas 後の ("go no") or の後の ("no ato no") are used only in noun qualifiers.

The syntactic polysemy of words such as *where, when, after, to* or *in* English may lead to a grammatical translational ambiguity for the MT system and cause translation mistakes. However, this phenomenon seems to be handled well by current commercial systems such as Google Translate or Yakuse Goma, in cases of simple sentence or short complex sentence translation. Again, we will take *after* as an example. We can see here the outputs of the Google Translate MT system with September 2011 version, for the three sentences of examples 4.4, 4.5 and 4.6.

(4.7) The time after the war was difficult.

戦争	後の	時間	は	困難	でした。
sensou	go no	jikan	ha	konnan	deshita
war	after	time	[topic]	difficult	was

(4.8) I will come after the war ends.

戦争	が	終了	した	後、	私	は	来る
sensou	ga	shuuryou	shita	nochi	watashi	ha	kuru
war	[subject]	end	did	after	I	[topic/subject]	come

でしょう。
deshou
will

(4.9) I will come after the war.

私	は	戦争	の後に	来る。
watashi	ha	sensou	no ato ni	kuru
I	[topic/subject]	war	after	come

Google Translate has achieved a correct translation of the English word *after* in the three generated sentences: 後の ("go no") when it is preposition in a noun qualifier, 後 ("nochi") when it is a complementizer, and の後に ("no ato ni") when it is preposition in an adverbial phrase.

4.4 Simple sentence translation

In a Japanese simple sentence, the subject, when it is expressed, is most likely to come in first position. Then, in the middle, come the verb complements. Finally, the sentence always ends with the verb or verbal adjective, sometimes followed by a post-verbal particle.

Subjects are not necessarily expressed in Japanese and can be only implied in null subject sentences. However, in order to keep accurately the semantic content of the source sentence, we have chosen to express them in the output generated by Its-2. Other choices may have been possible too, as we know that other English-to-Japanese MT systems do generate null subject or other null anaphora when it is possible (see Nakazawa et al. (2006) and Ryu et al. (2004)).

Still, we have chosen to generate null subject sentences when the source subject is an impersonal third person pronoun, such as *on* or *c'* in French or *it* in English, or other syntactic expletives such as *there* when it is used in the expression *there is*. Syntactic expletives and impersonal pronouns do not exist in Japanese. For example, in the following sentence, *it* is replaced by a null subject in the Japanese sentence, and the meaning of the copula should be understood as *it is* instead of *is*:

(4.10) It is expensive

高い	です
takai	desu
expensive	(it) is

(4.11) (a)

On	peut	aller	au	musée	gratuitement	
One	can	go	to	the	museum	for free.

(b)

美術館	に	ただで	行ける
bijutsukan	ni	tada de	ikeru
museum	to	for free	(one) can go

One can go to the museum for free.

Japanese is a head-final language (SOV). Therefore, Japanese verbs are always located at the end of simple sentences, sometimes followed by postverbal particles. As English is a SVO language, the verb often needs to be moved to the end when a simple sentence is translated from English to Japanese, like in the following example:

(4.12) I am eating bread

私	は	パン	を	食べている
watashi	ha	pan	wo	tabeteiru
I	[subject/topic]	bread	[object]	am eating

Most complements can be located anywhere in the Japanese sentence in the part that precedes the verb. Only subjects have a fixed position. They must be attached to the copula, or to another stative verb. Moreover, the copula itself cannot exist without a subject predicate because it is not considered as an independent lexeme (see Saint-Jacques (1966)). In English and French, it is possible to find an adverbial phrase between a subject predicate and a stative verb. Hence, reordering is needed, in order to place the Japanese subject predicate at the right position: just before the copula or stative verb. In the two following examples, we can see that an error in the reordering can lead to a false or ungrammatical translation:

(4.13) (a) This is true today.

(b) *これ は 本当 今日 だ
kore ha hontou konnichi da
this [topic] reality today it is
* This is reality today

(c) これ は 今日 本当 だ
kore ha kyou hontou da
this [topic] today real it is
This is true today

(4.14) (a) C'est toujours un beau paysage
It's always a beautiful landscape

(b) ? 美しい 景色 いつも です
utsukushii keshiki itsumo desu
beautiful landcape always it is
? a beautiful landscape is always

(c) いつも 美しい 景色 です
itsumo utsukushii keshiki desu
always beautiful landcape it is
It's always a beautiful landscape

4.5 Implementation

4.5.1 Reordering and transfer rules

Rules for word reordering and translation have been implemented, following the linguistic principles described in Sections 4.2 to 4.3. These rules are applied

at the transfer phase, which is the phase in the translation process where the source words, or group of words, are translated. The default cases, where the words or constituents are put in the same order as in English, are not mentioned in the rules.

The following rules are applied to word reordering inside the DP:

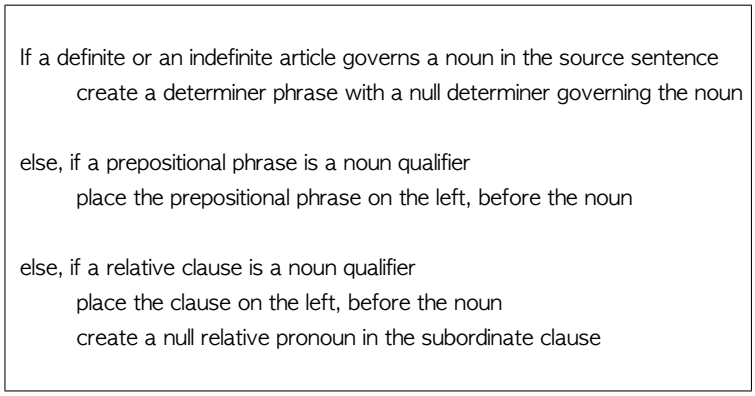


Figure 4.1: *Word reordering in the DP*

A rule has been added for the translation of subject pronouns:

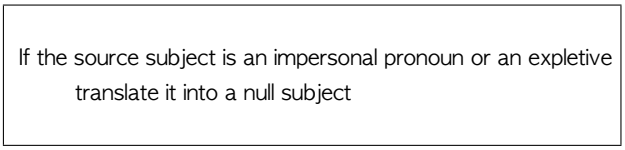


Figure 4.2: *Translation of subject pronouns*

This rule is applied to word reordering inside the PP:

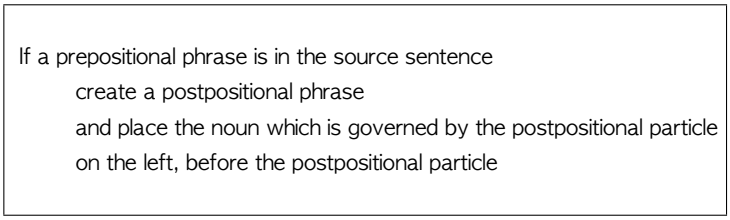


Figure 4.3: *Word reordering in the PP*

Transfer rules for word reordering at the sentence level have also been implemented:

```
In a simple sentence
        place the verb on the right, at the end of the sentence
if there is a functional projection with a stative verb and a subject predicate
        place the attribute (adjective or noun) just before the stative verb
```

Figure 4.4: *Word reordering in the simple sentence*

4.5.2 Generation procedures

Procedures for the generation of case particles or topic particles have been implemented. During the transfer phase the grammatical functions of the verb arguments are defined depending on the lexical information about the verb syntactic structure.

Then, the generation phase is the second phase of the translation process, where the inflected form of the output lexemes is selected and the output sentence generated. During this phase, case and topic particles are added following the grammatical functions given in the transfer phase. Indirect objects receive a postpositional particle which is the translation of the source preposition, according to the verb subcategorisation as it has been recorded in the lexicons (see Chapter 3). In the case where no postposition has been assigned, dative postpositional particle に ("ni") is generated.

```
if the argument is subject
        add topic particle は ("ha")
else if it is a direct object
        if the verb has a "passive structure"
                add case particle が ("ga")
        else add case particle を ("wo")
else if it is an indirect object and no postpositional particle has been added previously
        add postpositional particle に ("ni")
```

Figure 4.5: *Particle generation*

4.5.3 Context-sensitive lexical selection

When several translations are eligible for the same lexeme (cases described as *translational ambiguities* in (Hutchins and Somers (1992), p.99)), a lexical selection should be achieved. In the default cases, when no lexical selection procedure

is available, Its-2 (in its current state of development) choses the default translation. The default translation is the one which is given the highest score among the possible translations of the source lexeme in the bilingual lexicon. However, this ranking is a generic ranking which is not sensitive to context of the sentence. This can lead to translation mistakes, in cases where the default translation is not the most appropriate.

That is why specific lexical selection procedures have been written. The goal of these procedures is to select a translation candidate which is different from the default one, in a specific context. The context must be defined by syntactic, semantic, or lexical properties that can be deduced from the syntactic parsing of the source sentence.

To avoid errors that can be caused by the translation of words that can either be conjunctions or prepositions that have several possible translations, a lexical selection procedure has been written. The first step comes directly from the syntactic parsing result: if the word is analysed as a conjunction, the conjunctive translation is naturally chosen, and if the word is analysed as a preposition, the prepositional translation is chosen. Then, if a choice is needed between two possible translations of the same preposition, the selection procedure is used, applying the following rule:

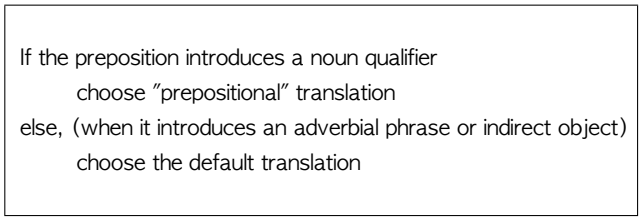


Figure 4.6: *Selection for preposition translation*

Another point where lexical selection is needed is the translation of the English verbs *to be* and *to have*, when they are used as verbs rather than auxiliaries. *To be* needs to be translated into the Japanese copula だ ("da": it is) when it introduces a subject predicate, and into a position verb such as ある ("aru") or いる ("iru") when it describe a location. These two verbs both mean either *to be (at)* or *there is*, and いる ("iru") needs to be used instead of ある ("aru") when the subject or subject predicate is a person.

```
subject +be + x -〉
    if x is an adverbial phrase of place and the subject is a person:
        translation of the subject + は (″ha″)
        + translation of the adverbial phrase of place + が (″ga″)
        + いる (″iru″)
    else if x is an adverbial phrase of place and the subject is not a person:
        translation of the subject + は (″ha″)
        + translation of the object + が (″ga″)
        + ある (″aru″)
    else (if x is a subject predicate):
        translation of the subject + は (″ha″)
        + translation of the subject predicate
        + copula だ (″da″)
```

Figure 4.7: Translation of *to be*

To have is also ambiguous in its translation because it can either actually express possession, and then be translated into 持っている ("motteiru") or express the presence of something or someone, and then be translated into ある "aru" or いる "iru" (which literally means *there is*). In those cases, the semantic asymmetry is clear, and represents a lexical divergence in Dorr's classification. We have chosen to always select ある "aru" or いる "iru", which have a more general meaning than 持っている ("motteiru"), in order to avoid the generation of false translations:

```
subject + ″have″ + object -〉
    if the object is a person:
        translation of the subject + は (″ha″)
        + translation of the object + が (″ga″)
        + いる (″iru″)
    else:
        translation of the subject + は (″ ha″)
        + translation of the object + が (″ga″)
        + ある (″aru″)
```

Figure 4.8: *Translation of "to have + object"*

4.5.4 Tests

We have performed a series of tests that showed very good results on word and constituent reordering in the simple sentence. However, the overall translation quality was not as high as the one of state-of-the-art MT systems.

This was mostly due to problems related to lexical selection, and sometimes to an incorrect handling of unknown words. The selection procedures that we have implemented have solved important *grammatical* translational ambiguities, but only a tiny part of existing *conceptual* translational ambiguities (Hutchins and Somers (1992), pp. 99-102). Hence, remaining errors in lexical selection often resulted in the generation of Japanese words that were syntactically acceptable, but with a meaning that seemed very improbable in the source sentence semantic context.

4.6 Conclusion

In this chapter, we have presented a set of rules that have enabled us to handle structural differences, achieving correctly word reordering, constituent reordering and case particle generation for the translation of simple sentence from English to Japanese. We have also shown lexical selection procedures that have solved some syntactical and lexical ambiguities encountered in the translation process.

The possible replacement of the second person pronoun by the name of the addressed person, and the possibility of a null subject sentence generation without a too heavy semantic loss remain two open questions that would need to be investigated further.

Handling politeness level generation would also be a possible improvement that could be added in simple sentence generation. A selection of the politeness level could be applied in lexical selection, especially for pronoun selection but also to some verbs and nouns. It would also imply a selection of the inflected forms at the sentence generation phase[5].

Beyond constituent reordering and lexical selection, other points such as multi-word expressions (as in Hino et al. (2011)) or structural asymmetries in translation need to be addressed. Thus, the next chapter will deal with Japanese adjectival sentence generation, which will show an aspect of structural asymmetries in simple sentence translation.

[5]The choice of a Japanese politeness level affects not only word selection (see section 3.2.5), but also verb tense form selection, as Japanese tense forms have a specific politeness value(see Kauffmann (2008b) and Kuwae (1984)).

Chapter 5

Treatment of structural asymmetries: the example of Japanese adjectival sentences

5.1 Introduction

Linguistic divergences or *translation asymmetries* (see Pause (1997)) arise when the structure of a sentence and of its translation are different. Linguistic divergences, in manual translation and in MT, have been studied and classified in six different types (see Dorr (1994) and Mahesh et al. (2005))

In this chapter, we will deal with *categorial divergences* (Dorr (1994); Mahesh et al. (2005)), which happen when the categories of the source and the target lexical entries are different, yielding *asymmetrical* correspondences in the lexicon. They lead to structural asymmetries between the original sentence and the resulting translated sentence. This phenomenon is a central topic in this thesis.

Asymmetrical correspondences can sometimes be found among Western languages. Here is an example with the French-English language pair:

(5.1) I'm hungry.

J'ai *faim.*
I have hunger

In this example, which was cited in (Dorr (1994), p.3), a collocation containing a verb and a noun is translated into a collocation containing a verb and

an adjective. Such examples are more frequent between East Asian languages (like Japanese or Korean) and Western languages (like French or English). For example:

(5.2) It hurts.

痛い。
itai
painful (it is)

In this example, the equivalent for the English verb *hurt* is the Japanese verbal adjective 痛い ("itai"). This phenomenon of predicative adjectives, often appearing without any auxiliary, is very typical in Japanese and Korean.

This chapter is primarily concerned with asymmetrical correspondences that include a Japanese adjective. Such correspondences were presented in detail in books such as (Kuwae (1984), pp. 60-77), and the fundamental principles for a rule-based translation of these asymmetrical cases were exposed in (Nagao (1989) pp.66-67). However, recent MT systems, in their English-Japanese or French-Japanese versions, have very unequal results in their way of treating them. We will describe here the implementation of transfer rules, on the basis of the ones that were presented by Nagao. We will compare our results with the ones of an existing MT systems and try to see the advantages and disadvantages of our method. Examples and experiments will also mention the Its-2 French-Japanese version, which is using smaller and less complete electronic lexicons than the English-Japanese version.

In the next section we will see how translations generate very different syntactic structures for the source and target sentences. Then, in the following section we will explain how we have treated this phenomenon in the Its-2 LBMT system. The last section will deal with the evaluation process and the results that have been obtained.

5.2 Description of the phenomenon

In this section we will first describe the typical features of Japanese adjectives. Then we will focus on the different kind of asymmetrical translation cases in which they can be found. Afterwards, we will study some different syntactic explanations related to these phenomena and finally, we will see what works have already been done about this subject.

5.2.1 Typical features of Japanese adjectives

In English or French, adjectives in predicative position always come after auxiliaries *to be* or *être*, or other stative verbs such as *seem* or *paraître*. The situation

is different in Japanese. In this language, the boundary between verbs and predicative adjectives is very thin (see Kuroda (1979a), p.236). Japanese adjectives can appear in predicative position without any auxiliary or stative verb:

(5.3) It is quiet here.

こちら	は	静か。
kochira	ha	shizuka
quiet	[topic]	(it is)

This phenomenon also exists in Korean, like in the two following examples:

(5.4) 맛있어요.
mashisseoyo
delicious (it is)
It is delicious. (polite way)

(5.5) 오늘은 추워요.
onurun chuweoyo
today〔topic〕 cold
Today it is cold.

Japanese adjectives are divided into three morphological categories: -い ("-i") adjectives, -な ("-na") adjectives and -の ("-no") adjectives.

"-i" adjectives are often called *verbal adjectives*. They are the inflectional ones. Their conjugation depends on tense and negation, like verb conjugation.[1] Their syntactic and semantic value, especially on attribute position, has been recently debated (see Namai (2002); Baker (2003); Nishiyama (2005)). They usually occur without any auxiliary verb (like in example 5.2). However, for predicative polite conjugation, out of relative or completive clauses, the polite present copula です ("desu") is added after "-i" adjectives:

(5.6) It is new.

新しい	です。
atarashii	desu
new	it is (polite way)

"-na" adjectives are sometimes called "nominal adjectives".[2]. They are invariable, but followed by the postposition な ("na") in attribute position. When they are used as a predicate, they are followed by the copula, which is conjugated in the right tense:

[1]However their conjugation is much more simple than verb conjugations. An inflectional adjective generates only about 14 inflected forms, instead of 162 for a verb (see Kauffmann (2008b)).

[2]Nouns and "-na" adjectives share some similarities. For example, nouns can be followed by the simple present copula だ ("da"), as "-na" adjectives can.

(5.7) It was quiet here.

こちら	は	静か	だった。
kochira	ha	shizuka	datta
here	[topic]	quiet	it was

However, the simple present copula だ ("da") is often omitted, resulting in in a purely adjectival sentence, similar to the ones with "-i" adjectives (see example 5.3).

"-no" adjectives are in fact nouns that take a particular adjectival meaning when they are used as an attribute, followed by the genitive postposition の "no" and a noun. When they are used in a predicative way, most of them behave like "-na" adjectives:

(5.8) He is rich.

彼	は	お金持ち	だ。
kare	ha	okane mochi	da
he	[topic]	rich	is

Still, this rule does not work all the time. For example, "パリの" ("Paris no": of Paris) can be analysed as a "-no" adjective which is a correct translation of the French "parisien" (Parisian") in attribute position:

(5.9)

un	café	parisien
a	café	Parisian

a Parisian café

パリ	の	喫茶店。
Paris	no	kisaten
Paris	of	café

a Parisian café

In predicative position, "パリの" ("Paris no") cannot be used without the genitive postposition の ("no"):

(5.10)

??この	喫茶店	は	パリ	だ。
kono	kisaten	ha	Paris	da
this	café	[topic]	Paris	is

?? This café is Paris

When the postposition "no" remains, "パリの" ("Paris no") can be used in predicative position. But it should not qualify a person. So we can say:

(5.11) この 喫茶店 は パリ の だ。
kono kisaten ha Paris no da
this café [topic] Paris of is
This café is Parisian

But we cannot say:

(5.12) *この 女の子 は パリ の だ。
kono onna no ko ha Paris no da
this girl [topic] Paris of is
*This girl is of Paris

Instead, we should use sentence 5.13 or 5.14. In a more global view, nominal suffixes 人 ("jin": person) or 出身 ("shusshin":origin) should be added to any geographical "-no" adjective in predicative position.

(5.13) この 女の子 は パリ人 だ。
kono onna no ko ha Parisjin da
this girl [topic] Parisian (person) is
This girl is Parisian

(5.14) この 女の子 は パリ出身 だ。
kono onna no ko ha Parisshusshin da
this girl [topic] Parisian origin is
This girl is Parisian

The examples of Japanese predicative adjectival sentences we have seen so far were all translating the English auxiliary verb *to be*. What about other stative verbs? Some stative verb + adjective group sequences in English are translated into a modification of the "-i" adjective termination. For example here, the termination -そう ("-sou"), instead of the regular -い ("-i") conveys a meaning of hypothesis, quite similar to the one obtained with the sequences *it looks* + adjective or *it seems* + adjective:

(5.15) It looks fun!

楽しそう。
tanoshisou
fun[hypothetical]

Other stative verb + adjective group sequence are translated in a more symmetrical way: with a modification of the -i adjective termination and the adjunction of a Japanese stative verb:

(5.16) It becomes fun.

楽しく	なります。
tanoshiku	narimasu
fun	it becomes (polite form)

In such situations, "-na" adjectives do not have any termination variation, but they are added a に ("ni") postposition:

(5.17) It became quiet.

静か	に	なった。
shizuka	ni	natta
quiet	to	it became

We have seen here that, depending on the inflexion category and the politeness level:

- an auxiliary + adjective group sequence in English is translated into an adjective, or into another adjective + auxiliary group sequence;
- a stative verb + adjective group sequence in English is translated into an adjective, or into another adjective + stative verb group sequence.

These translation cases are either partially asymmetrical or symmetrical, because the categories of the words in the target sentence are the ones of the words in the source sentence: the adjective is translated into an adjective, and the verb is translated into a verb, or influencing the termination of the adjective.

5.2.2 Asymmetrical translation cases involving Japanese adjectives

Other cases are fully asymmetrical. Unlike the ones we have just seen, they are based on bilingual correspondences between two words of a different lexical category.

Most of the cases we will consider here consist of English or French verbs that should be translated into a Japanese adjective (as in example 5.2), or adjective + copula group sequence, as in this example:

(5.18) Je t' aime.
I you love
I love you.

私	は	あなた	が	好き	だ。
watashi	ha	anata	ga	suki	da
I	[topic/subject]	you	[nom.]	loved	is

I love you.

A more elliptical form of this last example is the following:

(5.19) Je t'aime. (I love you.)

私	は	あなた	が	好き。	
watashi	ha	anata	ga	suki	
I	[topic/subject]	you	[nominative]	loved	(is)

Although even more elliptical, another typical version of this example is that one:

(5.20) Je t'aime. (I love you.)

	あなた	が	好き。	
	anata	ga	suki	
(I)	you	[nominative]	loved	(is)

Here, the topic 私 ("watashi": "je") is not expressed but understood in the context, and so is the copula だ ("da"). The analysis of these examples raises several questions, which we will see in detail in section 5.2.3.

Still, from a purely descriptive point of view, we can see that two major kinds of verb-to-adjective sentence translation exist:

- the most common one is the one we have just seen in examples 5.18 to 5.20. As described in (Nagao (1989), pp.66-67), it can be structured like this: The subject of the English sentence is translated into Japanese and followed by a は ("ha") particle, or omitted. The object of the English sentence is translated and followed by a が ("ga") particle. The verb of the English sentence is translated into an adjective or an adjective + copula group sequence.

- the second one is in reverse order: The translated object of the English sentence comes in first position, followed by a は ("ha") particle, or is omitted. Then comes the translated subject, followed by a が ("ga") particle. Again, the verb is translated into an adjective or an adjective + copula group sequence. Here is an example of this kind:

  (5.21) Dogs scare me.

私	は	犬	が	怖い。
watashi	ha	inu	ga	kowai
I	[topic/subject]	dogs	[nominative]	frightening

Another type of cases was evoked by S.Y. Kuroda (1979a, p. 236). In those cases the best appropriate translation of English predicative clauses with an adjective are Japanese clauses composed of an adjective followed by a specific stative verb. Kuroda presented this example:

(5.22) (a) John is hot.

(b) ジョン は 暑 がっている。
John ha atsu gatteiru
John [topic/subject] hot seems to be feeling

(c) *ジョン は 暑い。
*John ha atsui.
John [topic/subject] hot

In this example the adjectival sentence 5.22.c is usually considered ungrammatical, and the verbal sentence 5.22.b can replace it as a translation of the English sentence 5.22.a. These cases happen for Japanese predicative adjective expressing emotions or feelings, when the subject is a second or third person and the sentence is not a question. In Japanese, saying how someone else feels without using a stative verb seems too affirmative, because it is thought that other people's feelings cannot be publicly known.

Nevertheless, Kuroda claimed that sentence 5.22.c remains correct when the sentence is included in a narrative story in the third person, what he calls *non reported speech*.

In many cases of sentence translation, the context which sentences are taken from remains unknown. So, sentence 5.22.b, which is more widely accepted than sentence 5.22.c, would seem the most appropriate translation of sentence 5.22.a. All adjectival clauses reaching the same requirements (second or third person, affirmative clause, feeling adjective) should be translated this way (with the stative verb がっている ("gatteiru": seem to be feeling)).

Many other asymmetrical cases contain collocations, instead of simple verbs, in the English or French sentence. In Door's classification, those cases represent both categorial (different word categories) and conflational (different number of words) divergences, and sometimes lexical divergences (different literal meanings). For example:

(5.23) J'ai peur.
I have fear.
I am scared

怖い。
kowai
frightening (it is)
I am scared

In this example the French auxiliary + noun collocation *avoir peur* (be afraid, literally: *have fear*) is used when the adjective 怖い ("kowai") would be used in Japanese.

Also, In the French-to-Japanese version of example 5.15, instead of the English verb *to look*, we should find the verb + noun collocation *avoir l'air* (literally: *have the air*):

(5.24) Ça a l'air marrant!
It has the air fun!
It looks fun!

楽しそう。
tanoshisou
fun[hypothetical]
It looks fun!

Other asymmetrical cases contain specific types of group sequences. Ikeya (1991, p.64) showed this example:

(5.25) He is good at tennis.

彼	は	テニス	が	うまい。
kare	ha	tennis	ga	umai
he	[topic/subject]	tennis	[nominative]	good

Here, the sequence *be good at* + noun (*tennis*) is translated into the sequence: noun (here テニス"tennis") + が("ga") + うまい("umai": good). The noun seems to be a kind of object of the adjective, in English, and as well in Japanese. Such translations are mostly meaningful for structures such as *be good at, be bad at*, etc. This is also true in French with sequences such as *bon en*+ adjective:

(5.26) Il est très bon en maths.
He is very good at math.

彼	は	数学	が	とても	得意	だ。
kare	ha	suugaku	ga	totemo	tokui	da
he	[topic/subject]	math	[nominative]	very	skilful	is

He is very good at math.

Kuwae, in (1984, p.76) talked about quite similar cases. She insisted on the fact that sentences expressing capacity in French, instead of sequence *être bon en* (be good at) + noun, tend to count more descriptive verbs:

(5.27) Il joue bien de la guitare.
He plays well some guitar
He can play the guitar well.

The Japanese equivalent she presented has an adjectival structure, without any literal translation of the verb *jouer* (play).

(5.28) 彼 は ギター が 上手 です よ。
kare ha guitar ga jouzu desu yo
he [topic/subject] guitar [nominative] good is [affirmation]
He's really good at the guitar.

She also cited examples without any direct object in the French sentence:

(5.29) *Vous conduisez bien!*
You drive well!
You can drive well!

The Japanese equivalent has the same adjectival structure, with the substantive 運転 ("unten": driving):

(5.30) 運転 が じょうず です ね。
unten ga jouzu desu ne
driving [nominative] good is isn't it
You drive well!

Another kind of examples was showed in (Ikeya (1991), p.66):

(5.31) Elephants have long trunks.

象 は 鼻 が 長い。
zoo ha hana ga nagai
elephant [topic] nose [nominative] long

Like in this example, some clauses with the group sequence *to have* + adjective + noun should be translated into Japanese adjectival clauses in which the English direct object noun is followed by が ("ga") + the adjective (see Kinoshita et al. (1992)).

However, the literal meaning of Ikeya's Japanese sentence 象は鼻が長い。 ("zoo ha hana ga nagai") is "As for elephants, noses are long". This implicitly means that we already suppose that elephants have a 鼻 ("hana": nose or trunk).

Thus, when we have no idea about the context, only clauses where the object is clearly a part of the subject should be translated into an adjectival clause. In other cases, it will be better[3] to use either the Japanese semi-auxiliary ある ("aru") or いる ("iru", see example 5.32b), or the possesion verb 持つ ("motsu", see example 5.33b).

[3]You should notice here that sentences 5.32c and 5.33c are not ungrammatical, but they might be inappropriate or rather surprising translations for sentences 5.32a and 5.33a, depending on the context.

(5.32) (a) I have a beautiful lover.

(b) 私 には 美しい 恋人 が いる。
watashi ni ha utsukushii koibito ga iru
I [topic] beautiful lover [nominative] (there) is
I have a beautiful lover.

(c) ???私 は 恋人 が 美しい。
watashi ha koibito ga utsukushii
I [topic] lover [nominative] beautiful (is)
??? My lover is beautiful.

(5.33) (a) Anthony a une voiture rouge.
Anthony has a car red.
Anthony has got a red car.

(b) アンソニー は 赤い 車 を 持っている。
Anthony ha akai kuruma wo motteiru
Anthony [topic/subject] red car [object] is possessing
Anthony has got a red car.

(c) ???アンソニー は 車 が 赤い
Anthony ha kuruma ga akai
Anthony [topic] car [nominative] red (is)
??? Anthony, his car is red.

5.2.3 Syntactic explanations

A crucial point of difficulty is the question of how case structures of Japanese adjectival sentences which contain a syntagm that is the translation of a French direct object should be analysed. Let us take another look at example 5.20:

(5.20) Je t'aime. (I love you.)

あなた が 好き。
anata ga suki
(I) you [nominative] loved

Here, the French verb *aime* is translated into the adjective 好き ("suki") (or, in example 5.34, by the non elliptical adjective + copula group sequence "好きです" ("suki desu"). The French t' is the direct object of the French sentence. あなた ("anata") is a possible translation of the second person object pronoun, and it is followed by the case particle が ("ga") which is usually identified as the nominative case marker, which should intend that あなたが ("anata ga") is the subject of the adjectival sentence. This raises several questions: Is あなたが ("anata ga") the subject or object of the sentence? Can Japanese adjectives take objects? If あなたが ("anata ga") is really the subject, then what is 私

は("watashi ha") ?

In traditional Japanese grammar, it is said that, in such Japanese sentences, あなたが ("anata ga") is actually the subject. Kunio Kuwae maintains this judgement in (1984, p. 76):

(5.34) J' aime la musique
I like the music
I like music.

私	は	音楽	が	好き	です。
watashi	ha	ongaku	ga	suki	desu
I	[topic]	music	[nominative]	liked	it is (polite way)

I like music.

In this example, he explains that the literal French translation of the sentence is "Quant à moi, la musique est préférée." (As for me, music is preferred). In this traditional analysis, 私は ("watashi ha") is the topic of the sentence, whereas 音楽が ("ongaku ga") is the grammatical subject. Kuwae's analysis seems to fit well the literal meaning of words, and to explain why and how adjectives are used in a position where verbs could have been expected.

Others have claimed that, even if Japanese adjectives used in these sentences are adjectives, they really behave like verbs. Thus, in a sentence like 5.20, native speakers feel that 好き ("suki") is the verb and あなたが ("anata ga") is the direct object[4].

(Hinds (1986), p. 78-79) wrote that "There are (...) adjectives which allow two noun phrase arguments, the second of which may be termed a direct object.". This analyse makes us understand why such sentences sometimes appear:

(5.35) (a) Do you like spring?

(b)

春	を	好き	です	か。
haru	wo	suki	desu	ka
spring	[object]	liked	is it	[question]

Here, the usual nominative case marker が ("ga") is replaced by the accusative case marker を ("wo"). This tends to show that the speaker feels that 春 ("haru": spring) is the object of the adjective 好き ("suki"). This case frame is considered ungrammatical in traditional grammar, but sometimes appears in colloquial speech. This tends to show that some adjectives are seen as transitive verbs by the speakers.

[4]Izumi Tahara, personal communication

Kuno (1973, pp.62-93) also stated that some adjectives could take an object. He wrote a list of 27 of these, among which 好き ("suki") was included. He also pointed out that most adjectives are able to generate what he called "double subject" or "multiple subject" clauses. In those clauses, the noun which is followed by a が ("ga") particle as well as the one which is usually followed by a は ("ha") can stand for the subject. The distinction that Kuno made between transitive and intransitive adjectives is based on semantic criteria.

Some analyses reject the traditional view without arguing that adjectives can take objects. For example, (Ikeya, 1991) showed that the scope of adjective changes depending on several semantic criterions.

Tsujimura (1996, p.229-231) explained that the が ("ga") particle cannot always identify the subject. She refuted the double subject hypothesis, but without attempting to identify the exact function of が "ga" when it was not, from her point of view, indicating the subject.

Finally, Nagao's point of view was accepting the idea of adjectival objects, but without denying the fundamental difference between verbs and adjectives (see Nagao (1989), pp. 64-67). He wrote: *In Japanese, situational descriptions (centrally related to 'being') are frequently employed, whereas in English behavioural expressions (related to 'doing') are more frequently used* (Nagao (1989), pp. 114-118). I will adopt a similar point of view, considering that Japanese adjectives practically accept objects, which are promoted to a nominative case marking, because of the "passive" structure of adjectival clauses. This also refers to Kuroda's idea that, speaking of the role of が ("ga") and を ("wo") particles in the Japanese clause, the distinction between nominative and accusative was "formal rather than substantial" (see Kuroda (1979b), pp. 164-200).

5.2.4 Related Work

Tests have shown that recent versions of online MT systems such as Google Translate (Google, 2012a), Babel Fish[5] and Yakuse!! GOMA (MT Labs, 2012) have mixed results in handling asymmetrical translation cases involving Japanese adjectives. We will come back to this point in section 5.4.2.

The idea of linguistic divergences in MT and their classification in MT have been described in detail in (Dorr (1994)). This classification has been used later in other works (see Lin et al. (2005); Mahesh et al. (2005)). (Pause, 1997) also commented on the phenomenon of translation asymmetries in MT.

[5]This experiment was carried out when the free online version of Systran was still named Babelfish.

More specifically, asymmetries in translation of Japanese adjectives have been commented on in the literature by Makoto Nagao (1989). He described both the linguistic phenomenon and its implications in rule-based MT. Tsujii also worked on translation divergences in English-Japanese MT, which included rules about the translation of Japanese adjectival sentences (see Tsujii and Fujita (1991); Kinoshita et al. (1992)).

As we have seen above, other authors described the phenomenon of Japanese adjectival sentences which would have an asymmetrical translation, especially in books about Japanese Grammar or Japanese linguistics. S.Y. Kuroda (1979b; 1979a), Kuno (1973), Kuwae (1984) and Ikeya (1991) described the phenomenon in detail. Hinds (1986) and Tsujimura (1996) and K. Kuroda (2007) mentioned it more briefly.

Some publications about Japanese MT also mentioned the problem of asymmetrical translation and the behaviour of Japanese predicative adjectives. Nishida (1980) already dealt with differences of word categories and differences of sentence structures, in English-Japanese MT. So did Mangeot and K. Kuroda (2003), focusing on French-Japanese translation, in order to improve the multilingual electronic dictionary Papillon. In an article dealing with Japanese verbs and English-Japanese verb translation, Nakaiwa (1994) underlined the verbal behaviour of Japanese adjectives. Matsumoto and Komachi (2006) had a similar point of view, in a paper about sentence structure translation for Japanese-English statistical MT.

5.3 A proposed solution to the problem

In this section, we will present possible solutions to the treatments of the phenomena described above. First, we will describe our syntactical choices for the generated Japanese sentences. Then we will deal with the transfer rules, which have been directly deduced from the translation cases shown in sections 5.2.1 and 5.2.2. Finally, we will present the process of implementation of these rules.

5.3.1 Chosen syntactic approach

5.3.2 Specification of transfer rules

The transfer rules dealing with Japanese predicative adjectives that have been written will be presented here. They have been deduced from the linguistic phenomena mentioned before. Most examples will be taken from the English-to-Japanese version, and some examples from the French-to-Japanese one.

Every rule will be presented as the one shown in Figure 5.1. For the cases where the structure of the translation is not always the same but changes de-

pending on syntactical, morphological or semantic features, the rule will contain an algorithm.

```
English (or French) typical construction -〉
        typical Japanese translation
```

Figure 5.1: *A transfer rule*

First, rules have been written for the translation of simple or semi-asymmetrical adjective translation cases (see figure 5.2).

```
″be″ + adjective -〉
        if the Japanese adjective is a ″-i″ adjective:
                ″-i″ adjective (conjugated)
        else, if the Japanese adjective is a ″-na″ adjective:
                ″-na″ adjective + copula だ(″da″) (conjugated)
        else, if the Japanese adjective is
        a ″-no″ adjectival noun:
                if the adjectival noun requires
                the の(″no″) postposition:
                        ″-no″ adjective + 出身(″shusshin″)
                        + copula だ(″da″) (conjugated)
                else, if the adjectival noun is geographical:
                        ″-no″ adjective + の(″no″)
                        + copula だ(″da″) (conjugated)
                else: ″-no″ adjective + copula だ(″da″) (conjugated)
```

Figure 5.2: *Predicative adjective translation*

These rules, as well as the following ones, have been written in order to be implemented in Its-2 (see Wehrli and Nerima (2008)). We will give more details about implementation in the next subsection.

A possible refinement of the rule described in Figure 5.2 would consist in omitting the copula at the simple present form:

Other rules[6] have been written for the translation of adjectives with stative verbs (see Figure 5.4). They are lexicalised rules that depend on the verb. This

[6]We have seen that the copula can be present after a predicative adjective and that some adjectives may have morphological variation, depending on morphological class, tense and politeness level. In the rules described in Figures 5.4 to 5.8, these phenomena are not explicitly described, and *adjective* may refer to a conjugated adjective or an [adjective+copula sequence].

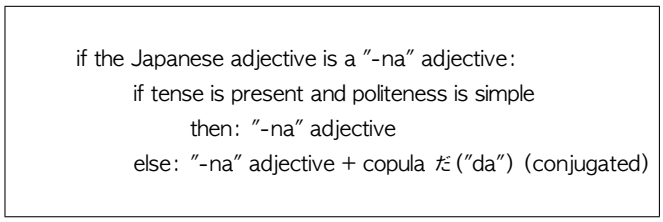


Figure 5.3: *Improvement of the first rule*

rule list is not exhaustive and other stative verbs can be added in a future work.

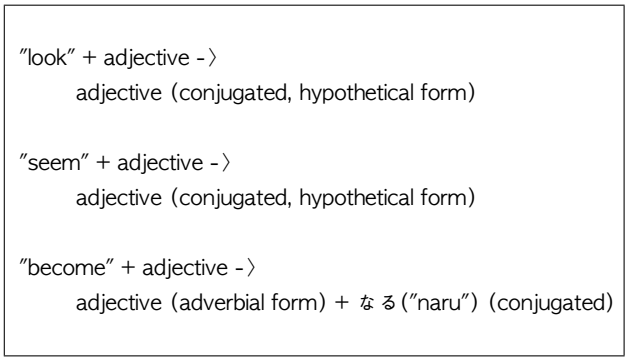


Figure 5.4: *Stative verbs + adjective translation*

The next rules we have adopted are for asymmetrical translation cases. The first ones deal with verb-to-adjective translation (see Figure 5.5). These rules are quite similar to the one proposed in (Nagao (1989), pp.66-67), but they also include the possibility of inversion of the argument positions, as in example 5.21. As in (Nagao (1989), pp. 66-67), the subject is translated, except for impersonal third person subjects (such as *it* in English or *on* in French), which are translated into null-subject clauses.

A specific rule explains the translation of emotional adjectives (see Figure 5.6). It is in fact another refinement to the rule 5.2. The generated Japanese clause will contain the stative verb がっている ("gatteiru": seem to be feeling).

The translation of some verb + noun collocations (such as example 5.23) into adjectives can usually be achieved directly from the lexical correspondence between the collocation and the adjective in the bilingual lexicon (see Seretan and Wehrli (2007)). However, some collocations, like the French *avoir l'air* (see

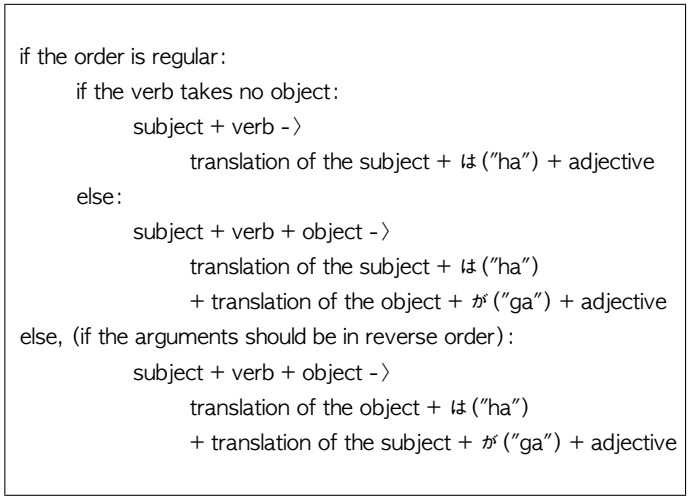


Figure 5.5: *Verb-to-adjective translation*

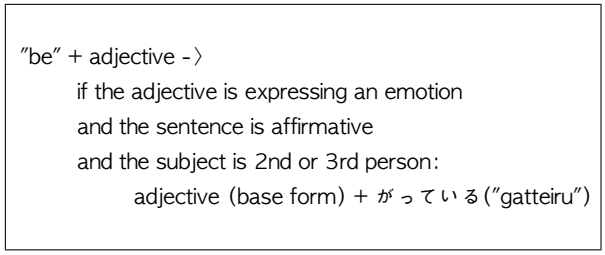


Figure 5.6: *Translation of emotional adjectives*

example 5.24) need specific lexicalised rules (see Figure 5.7).

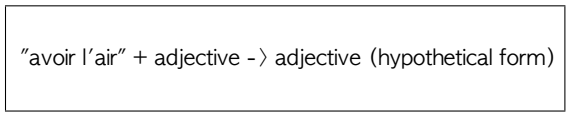


Figure 5.7: *Collocation+adjective-to-adjective translation*

The next rules have been written for the translation of specific group sequences. The first of these (see Figure 5.8) describe the translation of capacity expression when adjectival phrases like *bon en* ("good at", as in example 5.26)[7]

[7]This rule works with what we will call *level* adjective. We mean here adjectives such as *good*, or *average*, or *great*, etc., which describe someone's level in a given domain.

or *nul en* ("really bad at") are used, or when a verb and an adverb are used, as in example 5.30.

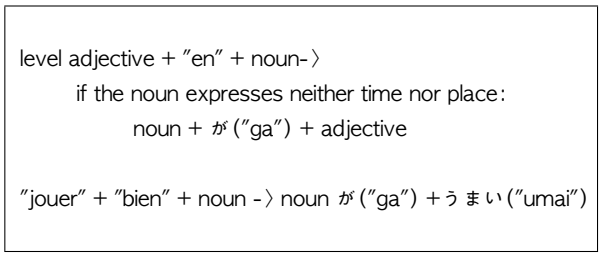


Figure 5.8: *Translation of capacity expression*

The last one (see Figure 5.9) deals with the translation of clauses containing the French group sequence: *avoir* + noun + adjective or *avoir* + adjective + noun, or the English one: *have* + adjective + noun, as in (Kinoshita et al., 1992). They require a semantic tagging of nouns and pronouns. Adjectival clauses should be generated when the object represents an entity which is a part of the entity represented by the subject (For example, in *The computer has a wide screen.*, the *screen* is a part of the *computer*; in Example 5.31, where the subject is *elephants*, the object is *trunk*, and a *trunk* is a part of the *elephant*'s body.). Because this constraint would be too complex to verify in all the possible cases, we will choose to produce this kind of adjectival clauses only when the object is a noun that designates a part of the body, such as *head* or *hair* or *trunk*, etc. In those cases, we will assume that the object refers to a part of the body of the person or animal mentioned in subject position. In the other cases, adjectives remain in attribute position, and verbs ある ("aru"), いる ("iru") or 持っている ("motteiru") should be used instead.

5.3.3 Implementation in the MT System

First, the rules described in Table 5.2 have been implemented, in order to translate English sentences containing adjectives and the *be* auxiliary. Then, the rules of Table 5.4, for the translation of adjectives with other stative verbs have been added. Then, the rules defined in Figure 5.5 have been implemented, enabling the program to translate verbal clauses into adjectival clauses. Finally, the lexicon has been updated and the implementation of the rules described in Figures 5.7 has been carried out.

```
subject + ″have″ + adjective + object -〉
    if the object is a part of the body:
        translation of the subject + は (″ha″)
        + translation of the object + が (″ga″)
        + adjective
    else if the object is a person:
        translation of the subject + には (″ni ha″)
        + adjective + translation of the object + が (″ga″)
        + いる (″iru″)
    else if the object is a physical object and the subject is a person:
        translation of the subject + には (″ni ha″)
        + adjective + translation of the object + を (″wo″)
        + 持っている (″motteiru″)
    else:
        translation of the subject + には (″ni ha″)
        + adjective + translation of the object + が (″ga″)
        + ある (″aru″)
```

Figure 5.9: *Translation of "have" + adjective + noun*

5.4 Evaluation and results

In this section, we will try to evaluate the quality of the implemented transfer rules. First, some tests and the obtained Japanese sentences will be shown and compared to the ones of other MT systems. Then, we will do an evaluation of Its-2 and of Google Translate on the generation of simple Japanese adjectival sentences. Finally, we will examine the results of the evaluation.

5.4.1 First tests

The first tests showed that Its-2 was able to generate adjectival sentences when the source sentence only contained a verb. For example, in Figure 5.10, the output sentence 私は犬が好きだ ("watashi ha inu ga suki da") is a correct possible translation. They also showed that collocations or multi-words expressions still missing in the lexicons were not translated correctly. So, some of them were inserted into the lexicon, and their translation has been corrected.

Then, we compared the output of Its-2 on possibly asymmetrical cases involving adjectival clauses with the ones of three other MT systems: Google

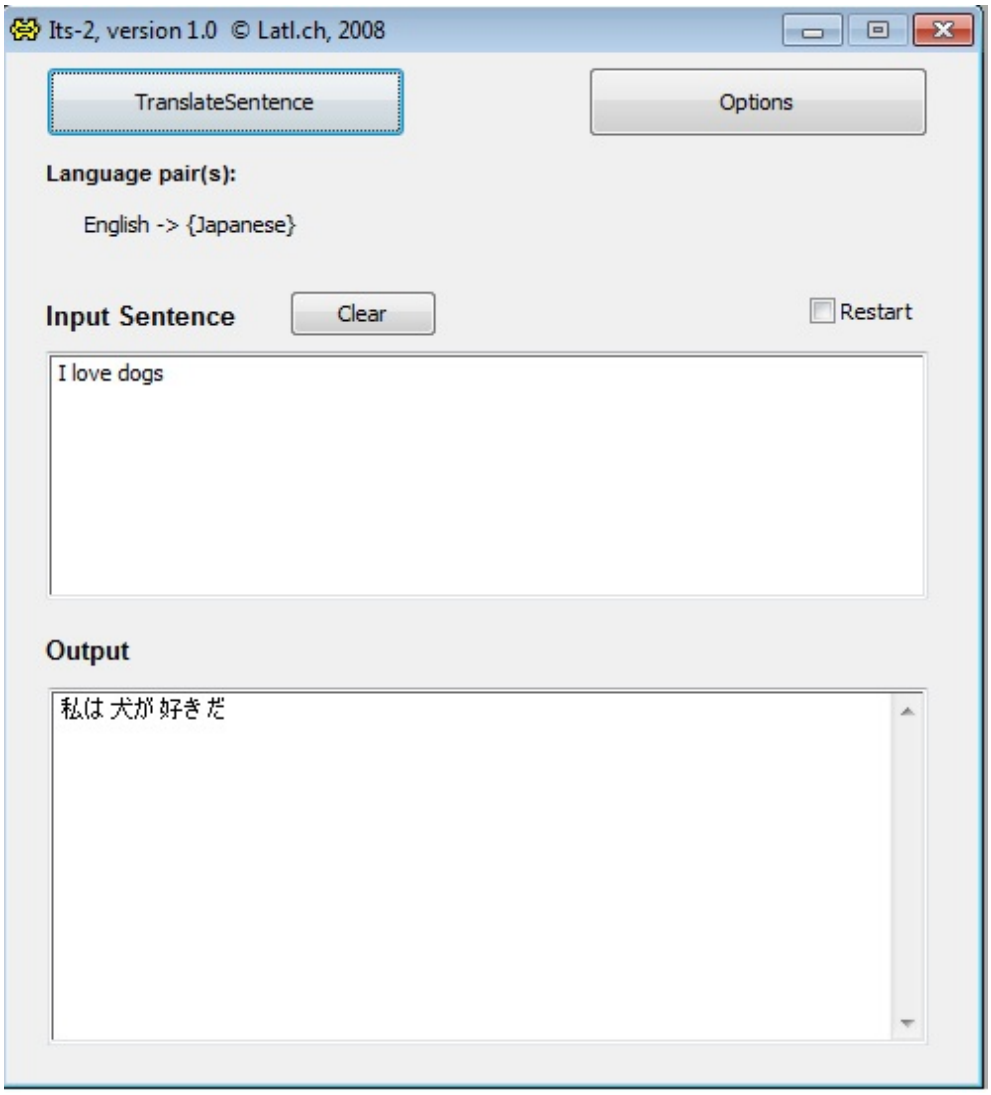


Figure 5.10: *Correct asymmetrical translation generated by Its-2*

Translate, Babelfish and Yakuse Goma, with the versions available on the internet in December 2009. The tests showed that Babelfish and, to a lesser extent, Google Translate and Yakuse Goma, tended to favour the use of verbs instead of adjectives when the source sentence contains a verb, whereas Its-2 was generating a larger amount of asymmetrical translations.

As an example, we can see the translation of the sentence *I enjoyed the holiday* by the four MT systems. The output of Its-2 (Figure 5.11) 私は休日が楽しかった ("watashi ha kyuujitsu ga tanoshikatta") which literally means *To me, the holiday was fun* The output of "Yakuse!! GOMA" (Figure 5.12) 私は休暇を楽しみました ("watashi ha kyuuka wo tanoshimimashita"). The literal meaning of this sentence is exactly the same as the original sentence: *I enjoyed the holiday.* The output of Google Translate (Figure 5.13) and Babelfish was 私は、休日を楽しんだ ("watashi ha kyuujitsu wo tanoshinda"), which is another literal translation of the source sentence. This translation, as well as the previous ones, are grammatically and semantically correct. There has been a debate among native speakers to determine which of the adjectival or the verbal clause was the more appropriate translation. Some claimed that the verbal translation had a better grammatical form while others claimed that the adjectival one seemed more natural in Japanese.

5.4.2 Evaluation process

This evaluation task has been realised with the help of Toshiaki Nakazawa and Hiroshi Manabe at Kurohashi-Kawahara laboratory, Kyoto University, in April 2010.

In order to get a more precise overview of the quality of the translations produced by Its-2, we set up a manual evaluation of the system, compared with Google Translate, on the generation of simple Japanese adjectival sentences, for English-Japanese and French-Japanese translation. The evaluation process was the following: a native speaker of Japanese defined 20 simple sentences of the source language. Those sentences were possibly translatable into adjectival sentences in Japanese. These two sets of 20 sentences were automatically translated by the systems. Then, a Japanese native speaker with a high knowledge of the source languages gave two scores between 0 and 5 to each output sentence: the first score was evaluating understandability and semantic accurateness and the second one was evaluating grammaticality and fluency.

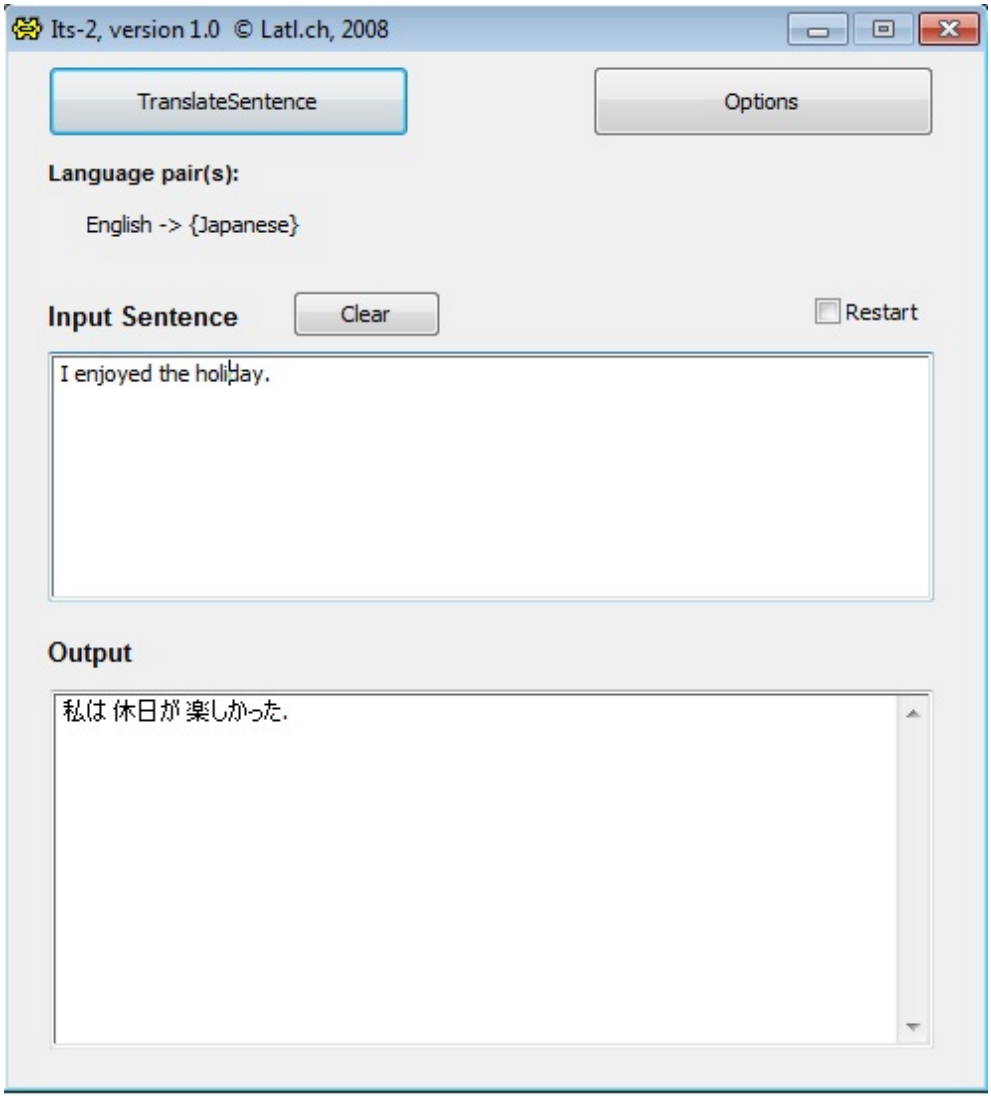


Figure 5.11: *Other correct asymmetrical translation generated by Its-2*

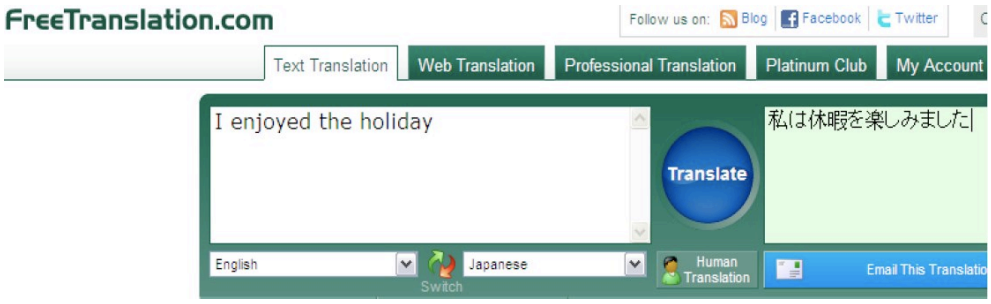


Figure 5.12: *translation generated by Yakuse Goma*

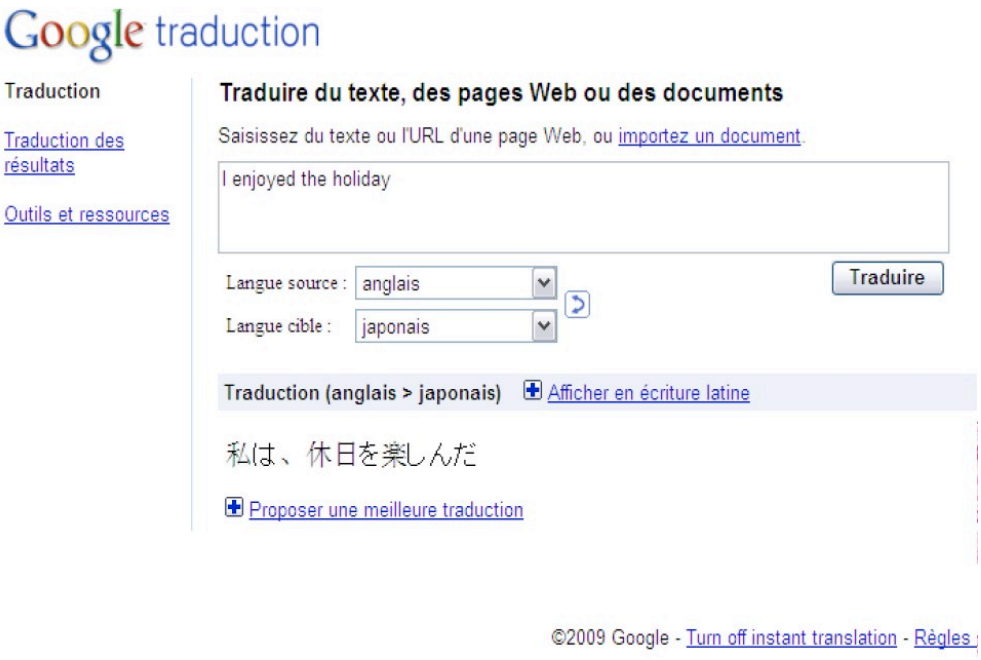


Figure 5.13: *translation generated by Google Translate*

5.4.3 Results

	Understandability and semantic acurateness	Grammaticality and fluency
Google Translate (April 2010)	4.35/5	4.5/5
Its-2 (April 2010)	2.9/5	3.45/5
Its-2 (June 2009)	2.25/5	2.75/5

Figure 5.14: *Compared evaluation of English-Japanese translations*

	Understandability and semantic accurateness	Grammaticality and fluency
Google Translate (April 2010)	3.68/5	4.21/5
Its-2 (April 2010)	2.21/5	2.37/5

Figure 5.15: *Compared evaluation of French-Japanese translations*

The results of the evaluation showed that the score of Its-2 remained very low for the French-Japanese language pair. This was mainly due to the incompleteness of the French-Japanese bilingual lexicon.

For the English-Japanese pair, the results were better in the April 2010 version of Its-2, that included adjectival clause generation procedures, than in the older version, that did not include any (+0.65/5 on understandability and semantic accurateness and +0.7/5 on grammaticality and fluency). This improvement highlights the relevance of our method. However, our scores remained lower than those of Google Translate, especially at the semantic level (1.45/5 less than Google Translate on understandability and semantic accurateness and 0.75/5 less than Google Translate on grammaticality and fluency). The main reasons for this failure to reach higher scores are the weakness of our lexical selection and the lack of bilingual data about collocations and other multi-word expressions (we will come back to these points in Section 8.5). Errors in lexical selection led to semantic mistakes while errors in collocation translation have been the source of grammatical mistakes.

5.5 Conclusion

In this chapter we have addressed the phenomenon of categorial translation divergences, among structural asymmetries in MT. We have focused on the example of Japanese adjectival sentences generation and proposed a set of transfer

rules, enabling the linguistics-based translation system Its-2 to generate asymmetrical translations, without using any pivot language in the transfer phase.

The results have been promising, but not as good as other current MT systems. This underlines the need for a better lexical selection in Its-2, and for the need to add English-Japanese bilingual data about multi-word expressions and collocations in the system.

Chapter 6

Treatment of complex sentences

6.1 Introduction

In this chapter, we will deal with structural differences between English and Japanese complex sentences. As words have a specific order in a language, clauses also have a typical order[1], which is sometimes different in English and in Japanese. For example, it is possible to find a subordinate clause after the main clause verb in English, whereas it is impossible in Japanese (except in colloquial speech, as a precision added after the sentence) :

(6.1) (a) He came even if the weather was bad.

(b) *彼 は 来ました、天気 が 悪くても。
kare ha kimashita tenki ga warukutemo
he [topic] came weather [nominative] bad even if
He came, even if the weather was bad. (incorrect sentence, except in colloquial speech)

(c) 天気 が 悪くても、 彼 は 来ました。
tenki ga warukute mo kare ha kimashita.
weather [nominative] bad even if he [topic] came
Even if the weather was bad, he came. (correct)

Consequently, the translation process of a complex sentence from English to Japanese often requires clause reordering, as well as it sometimes leads to a specific verb form selection in the target sentence. Clause reordering has always

[1]A clause is grammatical unit next below the sentence in rank, made of a verb (or a verbal adjective) and of its complements. A simple sentence is made of one clause, whereas a complex sentence contains at least two clauses.

been a fundamental issue in MT between English and Japanese (or French and Japanese) (Makinouchi (1970), p.36; Nishida et al. (1980); Nagao (1989), pp. 120-121). All English-Japanese or Japanese-English MT systems with a syntactic parsing phase have treated that point (see for example Kinoshita et al. (1992); Brockett et al. (2002); Kurohashi et al. (2005); or more recently Lee et al. (2010)). Some have also proposed to cut long sentences in order to reduce the risks of errors in clause reordering (see Goh and Sumita (2011)). Our approach is based on a multilingual system able to handle long distance relationships in the sentence. We will try to see here how it can apply to English-Japanese complex sentence translation and if it can achieve clause reordering efficiently.

The type of clause reordering varies depending on the type of complex sentence structure, such as coordination or subordination. The study of these phenomenon highlighted the need for a good classification of conjunctive words.

Moreover, translation of conjunctive words from a language to another is sometimes ambiguous. In order to solve this problem, we created lexical selection procedures that depend on syntactic context.

The first section will be dedicated to clause reordering, the second section will deal with the classification of Japanese conjunctive words and the third section will describe the lexical selection procedures of Its-2. The last part will describe and analyse the results of complex sentence translation with Its-2.

6.2 Clause reordering

In this section, we will first see the different kinds of English or French complex sentence structures and their respective Japanese translations. Then, we will focus on verb conjugation in Japanese complex sentences. In a third part, we will describe the transfer rules needed for the different complex sentence structures. Finally, we will briefly describe how the rules have been implemented in the system.

6.2.1 Complex sentence structures

Various cases of sentence structure exist. We will describe here examples of adverbial subordinations and coordinations or juxtapositions. The translation of English subordination into Japanese requires reordering, whereas the translation of an English coordination into Japanese does not.

When there is an adverbial subordination[2] in Japanese, the subordinate clause usually comes before the main clause. The conjunctive particle is then at the end of the subordinate clause. This reminds us that Japanese is a head

[2]Completive subordination will be studied in the next chapter and relative subordination was described in in chapter four.

final language, in which the verb of the main clause comes at the end of the sentence.

We find this pattern in example 6.2, where the subject of the subordinate clause is omitted.

(6.2) Yesterday, I stayed at home because it was cold.

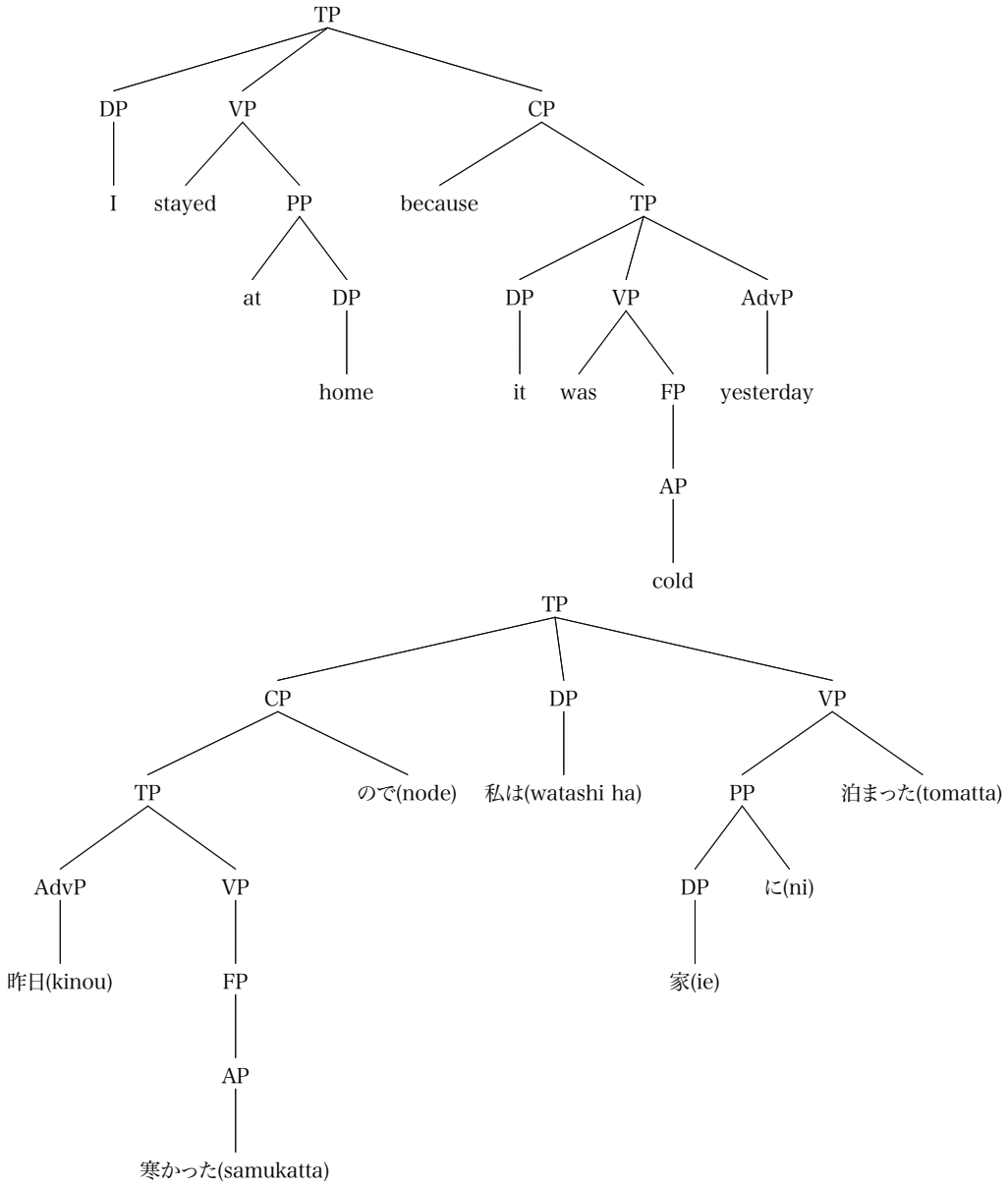
昨日	寒かった	ので、	家	に	留まった。
kinou	samukatta	node,	ie	ni	todomatta
yesterday	cold (past form)	because,	home	at	stayed

When the subject of the main clause is explicitly expressed, two patterns are possible. The first one, as in examples 6.2, 6.3 and 6.4 , starts with the subordinate clause and ends with the main clause. We can see an almost complete symmetry between the structures of the trees of the English and Japanese sentences in example 6.3[3], and a partial symmetry between those of example 6.2.

(6.3) I stayed at home because it was cold yesterday.

昨日	寒かった	ので、	私	は	家	に
kinou	samukatta	node,	watashi	ha	ie	ni
yesterday	cold (past form)	because,	I	[topic/subject]	home	at
留まった。						
todomatta						
stayed						

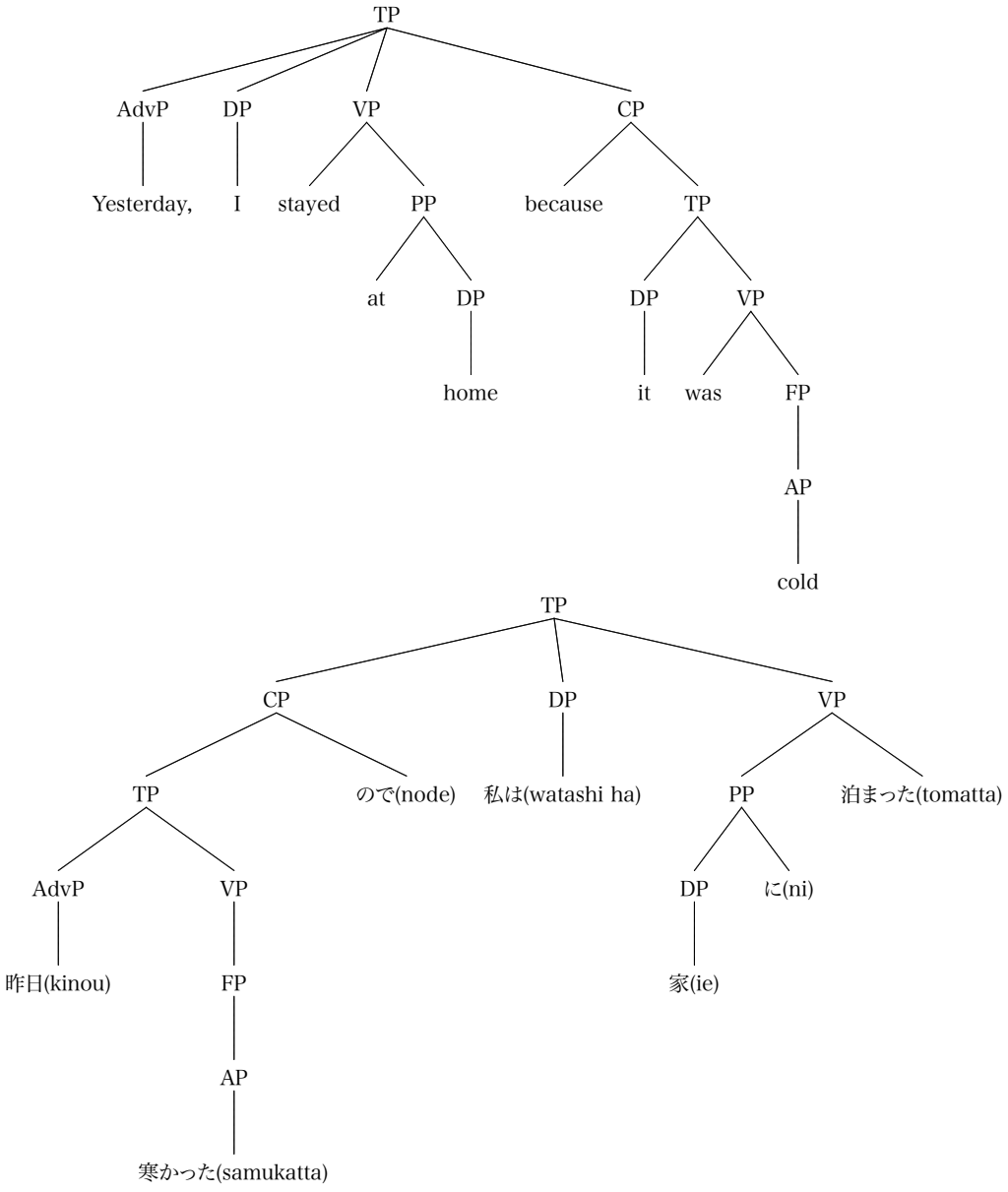
[3]In order to maintain a homogenous presentation between both languages, subjects have been put outside the VP in Japanese syntactic tree structures. However, as the expression of the subject is optional and as its position is not fixed in Japanese, some linguists have argued for a subject position within the VP.



(6.4) Yesterday, I stayed at home because it was cold.

昨日	寒かった	ので、	私	は	家	に
kinou	samukatta	node,	watashi	ha	ie	ni
yesterday	cold (past form)	because,	I	[topic/subject]	home	at

留まった。
todomatta
stayed



The second possible pattern, as in example 6.5, begins with the subject of the main clause, followed by the subordinate clause, and ends with the rest of the main clause. We have chosen to follow this pattern for the Japanese adverbial subordinate clauses generated by the translation system.

(6.5) Yesterday, I stayed at home because it was cold.

私	は	昨日	寒かった	ので、	家	に
watashi	ha	kinou	samukatta	node,	ie	ni
I	[topic/subject]	yesterday	cold (past form)	because,	home	at

留まった。
todomatta
stayed

Some English subordination conjunctions require the use of a Japanese auxiliary or a specific tense form when they are translated into Japanese. For example, it is possible to translate *if* using auxiliary なら ("nara") and to translate *while* conjugating the verb at ながら ("nagara") gerundive form (examples 6.6 and 6.7).

(6.6) If he comes, I will stay.

彼	が	来る	なら、	俺	は	留まる。
kare	ga	kuru	nara,	ore	ha	todomaru.
He	[nominative]	come	will,	I	[topic/subject]	stay

(6.7) She eats while she's watching TV.

彼女	は	テレビ	を	見ながら	食べます。
kanojo	ha	televi	wo	minagara	tabemasu
She	[topic/subject]	television	[accusative]	watching	eats

There are some less frequent cases where a Japanese conjunction of subordination can be put at the beginning of the subordinate clause, instead of the end. As those cases express condition, the verb must be conjugated in a Japanese conditional form. The conjunction can then either be explicitly expressed (see example 6.8, 6.9), or omitted (see example 6.10).

(6.8) Provided that he pays me a lot, I will work for him.

ただし	お金	を	いっぱい	もらえば、	彼	と
tadashi	okane	wo	ippai	moraeba,	kare	to
provided that	money	[accusative]	a lot	(I) would get,	him	with

働く。
hattaraku
work

(6.9) If it rains, I will go to the cinema.

もし	雨	が	降れば、	映画館	に	行きます。
moshi	ame	ga	fureba	eigakan	ni	ikimasu
if	rain	[nominative]	fall(conditional)	cinema	to	go

(6.10) If it rains, I will go to the cinema.

雨	が	降れば、	映画館	に	行きます。
ame	ga	fureba	eigakan	ni	ikimasu
rain	[nominative]	fall(conditional)	cinema	to	go

Coordination and juxtaposition do not require any reordering when they are translated by equivalent structures into Japanese. The clauses are expressed in the same order as in English. The translation of an English coordination can take different forms, depending on the conjunction in the source sentence and on the choice of the conjunctive word that will be used in the target sentence.

In most cases the verb at the end of the first clause is conjugated in the suspensive gerund (or て ("Te") form), followed by a Japanese comma. The English conjunction is then translated into a Japanese adverb that goes in the beginning of the second clause of the sentence.

(6.11) I went to school and then I came back home.

学校	に	行って、	そして	家	に	帰りました。
gakkou	ni	itte,	soshite	ie	ni	kaerimashita
school	to	go(gerund),	and then	house	to	came back

In other cases, a Japanese conjunctive particle is added at the end of the verb of the first clause. In these cases, a comma can be added after the conjunctive particle.

(6.12) I want to go skiing but there is no snow.

スキー	に	行きたい	けど、	雪	が	ない。
ski	ni	ikitai	kedo,	yuki	ga	nai
skiing	to	(I) want to go	but,	snow	[nominative]	there is not

Sometimes the English sentence with two coordinated clauses can be translated into two short Japanese sentences. The English conjunction is again translated into an adverb, that comes in the beginning of the second sentence.

(6.13) I want to go skiing but there is no snow.

スキー	に	行きたい。	でも	雪	が	ない。
ski	ni	ikitai.	demo	yuki	ga	nai
skiing	to	(I) want to go.	But,	snow	[nominative]	there is not

(6.14) It is late, so I'm hungry.

遅そい	よ。	だから	おなか	が	減った。
osoi	yo.	dakara	onaka	ga	hetta
late	[reaffirmation].	So	stomach	[nominative]	empty

The example of the coordination conjunction *and* is more complex, because there is no such coordination that could be used between clauses in Japanese. Hence, generating a Japanese juxtaposition seems the more adequate translation for an English coordination with *and*.

(6.15) It is late and I'm hungry.

遅くて、	おなか	が	減った。
osokute,	onaka	ga	hetta
late(gerundive),	stomach	[nominative]	empty

Juxtaposition, like coordination, keeps the same clause order as English. The Japanese comma is used to separate clauses and the verbs of the first clauses must be in gerundive form.

(6.16) He turned off the light, left the room and came back home.

彼	は	電気	を	消して、	部屋	を	出て、
kare	ha	denki	wo	keshite,	heya	wo	dete,
he	[topic/subject]	power	[accusative]	close,	room	[accusative]	leave,

帰りました。
kaerimashita.
came back home

6.2.2 Consequences on conjugated verbs

When juxtapositions and most cases of coordination are translated into Japanese, the verbs of the first clause must be in a gerundive form (see examples 6.11, 6.15, 6.16). The most frequent one is the suspensive gerund (sometimes called "Te" gerund or "Te" form).

Hence, in the generation module of the Its-2 MT system, verb form selection has been modified for those cases: instead of the regular tensed forms (such as simple past, as in incorrect example 6.17), verbs should be conjugated in the "Te" gerundive (as in correct example 6.18).

(6.17) He turned off the light, left the room and came back home.

*彼	は	電気	を	消した、	部屋	を
kare	ha	denki	wo	keshita,	heya	wo
he	[topic/subject]	power	[accusative]	close(past),	room	[accusative]

出た、	帰りました。
deta,	kaerimashita.
leave(past),	came back home

(6.18) He turned off the light, left the room and came back home.

彼	は	電気	を	消して、	部屋	を
kare	ha	denki	wo	keshite,	heya	wo
he	[topic/subject]	power	[accusative]	close(gerund),	room	[accusative]

出て、	帰りました。
dete,	kaerimashita.
leave(gerund),	came back home

6.2.3 Classification and rules

Following the linguistic description of the different complex sentence structures, a series of transfer rules and generation rules has been written. We will first present the transfer rules.

Transfer rules for adverbial subordination:

```
if the target clause is an adverbial subordinate clause
        if the verb must receive the auxiliary なら ("nara")
                remove copula (if any); add なら
        else, if the verb must be accorded at ながら "nagara" gerundive form
                adjust the verb tense; if the verb is a copula, adjust at "Te" form

        insert the clause after the subject of the principal clause
```

For a subordinate clause expressing condition (as in example 6.8), that takes its conjunction in the beginning and comes before the main clause in both English and Japanese:

```
if the target clause is a subordinate clause
with a "provided that" type conjunction
and if the conjunction expresses condition
        adjust the verb tense
```

For coordination:

```
if the target clause is a coordinate clause
        if it starts with a sentential adverb type だから ( "dakara")
                insert a 。(Japanese dot) before the sentential adverb.
                (This separates the sentence into 2 shorter sentences)
        else, if it starts with a conjunctive particle like けど ("kedo")
                do nothing
        else, if the conjunction of the source sentence was "and"
                adjust the tense form of the verb of the preceding clause;
                put a Japanese comma; don't translate the conjunction
        else,
        (Default case, for sentential adverbs like そして ("soshite")
                adjust the tense form of the verb of the preceding clause;
                put a Japanese comma before the sentential adverb
```

For juxtaposition:

```
if the target clause is juxtaposed to the preceding one
        adjust the tense form of the verb of the preceding clause;
```

The implemented generation rules allows to assign the right tense forms to the verb. For example, in case of coordination or juxtaposition:

```
if the target verb must be put at "Te" form
        if it is already a "Te" form or another kind of gerund
                do nothing
        else
                select "Te" form
```

A heuristic rule that cancels the generation of topic particles は ("ha") in subordinate clauses has also been added, in order not to have too many topics in the generated sentence:

```
if the subject is in a subordinate clause
        add nominative particle が ("ga")
else (in main clause)
        add nominative particle は ("ha")
```

6.2.4 Implementation in the system

The rules have been implemented in the system at the English-Japanese transfer level and at the Japanese generation level. We have encountered problems related to syntactic parsing of juxtaposition and coordination. Incorrect syntactic parsing has a negative effect on the quality of the generated translations, causing errors in clause reordering, even if the reordering rules have been implemented correctly. These problems should be solved in a future version of the syntactic parser Fips for English.

The implemented rules, especially the ones dealing with coordination, underline the need for an accurate classification of conjunctive words. We will focus on that point in the next section.

6.3 Classification of Japanese conjunctive words

Japanese conjunctive words, that can either be conjunctive particles, conjunctions or sentential adverbs, play an important role in determining the structure of the sentence. In this section, we will first show how we have built a classification of conjunctive words. Then, we will explain how we have inserted the results of this classification into the lexical database.

6.3.1 A classification based on empirical data

The task of classification of Japanese conjunctive words has been realised in collaboration with Nobuhito Tamaki at Kurohashi-Kawahara laboratory, Kyoto University.

We wanted to use empirical data in order to determine the usual position of each conjunctive word in the sentence. Hence, we used a list of 109 Japanese conjunctive words or expressions, and a corpus of 90'000 sentences that had been extracted from the web. As words are not delimited by spaces in Japanese, we used the morpho-lexical analyser KNP (see Kurohashi and Nagao (2003)) to perform a word segmentation of the corpus. Then, we launched an automatic analysis of the segmented corpus, that counted the number of occurrences of each

conjunctive word and, among those occurrences, the percentage of sentence-initial uses. We could then deduce a classification of the conjunctive words ordered by the ratio of sentence-initial occurrences (see fig 6.1, 6.1, 6.3) or by the total number of occurrences in the corpus.

The second step of this experiment consisted in defining a manual classification based on the results of the analysis and on linguistic knowledge. The classification shows the position of the conjunctive words in the sentence and their syntactic properties. The results have shown that three groups of conjunctive words (or compounds) can be distinguished:

- The first 30% of conjunctive words were found in more than 90% of their occurrences in sentence-initial position and in most cases could not occur in another position;
- The second 61% could either be found in sentence-initial position or inside the sentence. When these conjunctive words are not in sentence-initial position, three subcategories of conjunctive words or compounds appear: post-verbal conjunctive particles, post-nominal conjunctive particles, and sentential adverbs that can appear after a Japanese comma, often at the beginning of a clause;
- The last 9%, that were found in less than 10% of their occurrences in sentence-initial position, are supposed to be in the middle of the sentence. More particularly, coordination particles が ("ga") and と ("to") are almost never found in sentence-initial position.

6.3.2 Annotation of the lexicon

The results of the classification have been added into the Japanese monolingual lexicon, classifying the conjunctive words into the appropriate subcategories. The final step was the update of the English-Japanese bilingual lexicon, in order to obtain a correct translation of English conjunctions. The Its-2 bilingual lexicon stores bilingual correspondences between lexemes, and some words have several possible lexemes. For example, *joke* can either be a verb or a noun, and those two lexemes are defined in the lexicon. In the same way, *so* can either be a conjunction or an adverb. We have chosen to translate *so* into だから ("dakara") when it is used as an adverb (example 6.20), and into それで ("sore de") when it is a conjunction[4] (example 6.19).

(6.19) Taro got hurt, so he cannot go to school.

[4]This choice was arbitrary. Other kinds of translations such as using だから ("dakara") and cutting sentences into two parts (as in example 6.14), would have made sense too.

太郎	は	けがして、	それで	学校	に	行けない。
Tarou	ha	kega	shite,	sore de	gakkou	ni
Taro	[topic/subject]	injury	do(gerund),	so	school	to

ikenai
cannot go

(6.20) So, he is late.

だから、	彼	は	遅い。
dakara,	kare	ha	osoi
so,	he	[topic/subject]	late

For about 100 English conjunctions, we checked the adequacy of the bilingual correspondences between conjunctive words, taking care of the choice of translations of the different possible lexemes of every conjunctive word. This approach gave good results on the generated output sentences, especially for the translation of coordination, which is highly dependent on the properties of the conjunctive word. We will come back to that point in the last section of the chapter.

Even a rich lexical knowledge in the database is not sufficient to solve problems related to lexical selection in ambiguous cases. We will deal with this question in the next section.

6.4 Lexical selection procedures

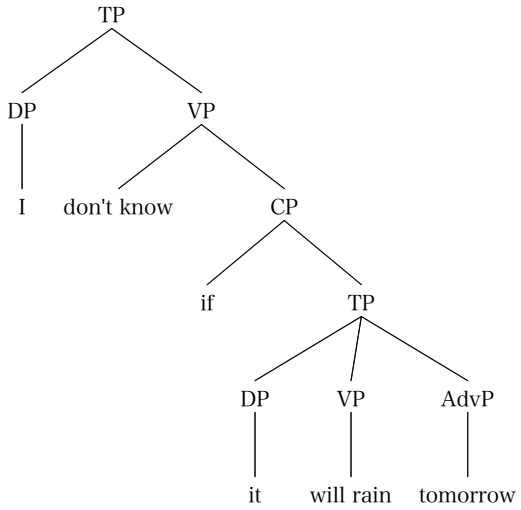
Some English conjunctions have several possible translations in Japanese, that differ depending on the context. These cases of grammatical translational ambiguities (Hutchins and Somers, 1992) may lead to translation mistakes. As we did for the translation of prepositions, we have added lexical selection procedures for a better selection among conjunctions.

6.4.1 Example

if is an English conjunction that is ambiguous in its translation into Japanese. *if* can be translated by か ("ka"), when it introduces an indirect interrogative clause. In example 6.21 , we can see in the sentence tree structure the CP introduced by *if* is attached directly to the VP of the main clause verb. It can also be translated by れば ("reba") conditional or なら ("nara") auxiliary when it expresses a condition (see examples 6.22, 6.23). In the tree structure of example 6.23, the CP is attached directly to the root TP.

(6.21) I don't know if it will rain tomorrow.

明日	雨	が	降る	か	知らない。
ashita	ame	ga	furu	ka	shiranai
tomorrow	rain	[nominative]	fall	[question]	(I) don't know

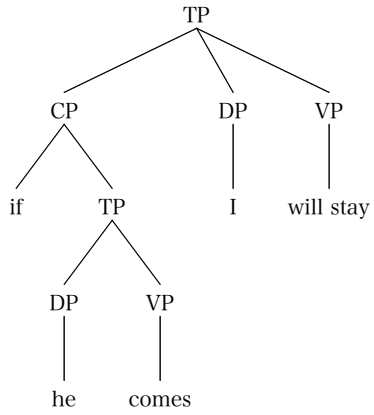


(6.22) If it rains, I will go to the cinema.

雨	が	降れば、	映画館	に	行きます。
ame	ga	fureba	eigakan	ni	ikimasu
rain	[nominative]	fall(conditional)	cinema	to	go

(6.23) If he comes, I will stay.

彼	が	来る	なら、	俺	は	留まる。
kare	ga	kuru	nara,	ore	ha	todomaru.
He	[nominative]	come	will,	I	[topic/subject]	stay



We have chosen to select なら ("nara") as default translation of *if*, and wrote a selection procedure that selects か ("ka") instead of なら ("nara") when *if* introduces a completive clause in the source sentence.

6.4.2 Results

The use of lexical selection procedures enabled us to overcome ambiguities in translation and to improve the quality of translation, in the respect of grammatical properties. However, when several synonyms are available for the same source lexeme, Its-2 may not always choose the one which will sound the more natural from a human speaker's point of view. Combining our method with language models or other statistical methods would add more precision in the quality of the lexical selection and should be considered in a future work.

6.5 Evaluation and results

In this section we will try to evaluate the improvement of translation quality after the work presented in this chapter.

6.5.1 First tests

The first tests showed good results on a test set of complex sentences. For example, Sentences 6.24, 6.25 could be translated in a grammatically correct way (even if the repetition of the pronoun 私たち ("watashitachi": we) in the second clause of example 6.25 seems rather unnatural in Japanese).

(6.24) Input: We made a cake because you had time.

Output:	私たち	は		あなた	が	時間	が
	watashitachi	ha		anata	ga	jikan	ga
	we	[topic/subject]		you	[nominative]	time	[nominative]

あった	から	ケーキ	を	作った。
atta	kara	keki	wo	tsukutta.
there was	because	cake	[accusative]	made

(6.25) Input: We made a cake, and we ate it.

Output:	私たち	が	ケーキ	を	作って、
	watashitachi	ga	keki	wo	tsukutte,
	we	[nominative]	cake	[accusative]	make(gerund),

私たち	は	食べた。
watashitachi	ha	tabeta.
we	[topic/subject]	ate

However, the sentences included in the first test set remain of a low complexity level : they all had at least two verbal clauses, but contained no interpolated clauses, no sentential objects and no elaborate combination of subordinate, coordinate and juxtaposed clauses.

6.5.2 Evaluation process

A small manual evaluation has been carried out. We made the MT system translate a text of 48 sentences extracted from a travel blog page. We compared the quality of the generated translations with the ones obtained with an older version of the MT system, that could not handle complex sentences. 73 percent of the extracted sentences (35 out of 48) were complex sentences and the text of the blog page was written in good English.

6.5.3 Results

Among the 48 translations of sentences from the travel blog page, 4 were improved and one was worsened. This makes an improvement on 6 percent of the sentences.

Analysing the overall results, the improvements made on complex sentence translations were not as good as expected. Even if there were some improvements, the system failed to translate correctly most of the complex sentences. This can be explained in part by the fact that these sentences had a more specific vocabulary and a more complicated structure than the ones used in the first test set, including for most of them elaborate combination of subordinate and coordinate clauses, interpolated clauses or sentential objects. So, three kinds of problems have arisen: problems with syntactic parsing, with words missing in the lexicon and with the treatment of sentential complements.

Non translated words or incorrectly parsed sentences have led to errors in word reordering or clause reordering, generating broken translations. We should remember here that Its-2 has been originally used for translations between European languages, that need less clause reordering than English-Japanese MT. Errors that are not very harmful in MT of pairs of languages with similar or close word order, like French-Italian or French-English (see Russo and Wehrli (2011)), can lead critically to wrong reordering in English-to-Japanese MT.

Errors in sentential objects translations have also led to incorrect reordering and incorrect tense form selection. We will go back to this point in Chapter 7.

6.6 Conclusion

We have shown a set of rules for the translation of complex sentence structures, lexical improvements based on empirical data and linguistic knowledge for the

generation of Japanese conjunction, and lexical selection procedures depending on the syntactic context.

We obtained good results on the translation of examples of moderately complex sentences. On more complicated sentences, extracted from a well-written blog page, the translations were sometimes improved but globally remained of poor quality, mainly because of problems related to syntactic parsing or caused by words missing in lexicons. We can conclude that the implemented transfer rules and lexical improvements were relevant for the translation of complex sentences, but would require other improvements to the MT system to work efficiently.

Some of these improvements would not require any change to the MT system structure: for example, punctuation should be handled better. Other improvements would be possible, but requiring deeper modifications in the system : statistical corrections could add more precision for syntactic parsing selection and for lexical selection.

It would also be good if the MT system was more resistant to problems related to unknown words (such as, for instance, some proper nouns[5] or domain-specific vocabulary), at both parsing and transfer levels. Currently, the system infers a category when an unknown word is found, but this often leads to a wrong parsing and then an error in word reordering at the transfer phase. If the system was able to infer the good category and then reorder correctly the unknown word, the translations would be less damaged.

[5]Among unknown words and expressions, proper nouns and named entities represent a specific problem. Most of them should be either translated into kanji or transliterated into katakana (as in Wu and Tsujii (2011)). But many Japanese proper nouns have several possible orthographic variants, which can often lead to a wrong selection of the translation. Named entities such as organisation names often require a specific translation, and thus need not only to be entered in the bilingual lexicon, but also to be recognised even if their form may vary or be abbreviated. Moreover, proper nouns often have homographs among common nouns an in the rest of the lexicon, which increases the risk of error in translation.

Ratio	Count in initial position	Total count	Conjunctive word or expression	Meaning
0.995	14627	14707	ついで (tsuide)	By the way
0.993	695701	700939	一方 (ippou)	Meanwhile
0.991	1453319	1466727	ただし (tadashi)	However
0.987	207436	210071	よって (yotte)	Therefore
0.976	740023	758223	ところが (tokoro ga)	However
0.969	63633	65645	そのほか (sono hoka)	Furthermore
0.969	371608	383531	なので (nanode)	So
0.967	358812	371211	しかしながら (shikashi nagara)	However
0.960	647513	674718	ところで (tokoro de)	By the way
0.960	213881	222860	とはいえ (to haie)	Nevertheless
0.959	472326	492654	すると (suru to)	Then
0.957	543198	567778	ですから(desu kara)	So
0.957	2118903	2215006	で (de)	And
0.957	13576	14180	とはいうものの (to haiu mono no))	Although having said that
0.953	27997	29393	なぜならば (naze naraba)	The reason why that is so
0.952	1641396	1724997	ちなみに (chinami ni)	Incidentally
0.951	1479886	1556056	さて (sate)	And now
0.944	224276	237664	なぜなら (naze nara)	The reason why
0.943	7130743	7558327	しかし (shikashi)	But
0.930	804111	864283	だが (da ga)	However
0.927	10626	11457	いっぽう (ippou)	Meanwhile
0.926	434081	468921	したがって (shitagatte)	Thus
0.926	238054	256940	ですが (desu ga)	However
0.921	115080	124960	けれども (keredomo)	However
0.920	342986	372734	その他 (sono hoka)	In addition
0.918	6505	7083	それというのも (sore to iu no mo)	Even that
0.910	277142	304395	但し (tadashi)	However
0.908	1371	1510	けども (kedomo)	But
0.907	240725	265409	従って (shitagatte)	Consequently
0.906	2600	2871	ですけれども (desu keredomo)	However
0.903	2325	2574	それにひきかえ (sore ni hikikae)	In contrast
0.901	18493	20527	とすると (to suru to)	Then
0.901	14156	15710	とすれば (to sureba))	Consequently
0.893	514372	575868	それにしても (sore ni shite mo)	Even so
0.890	140663	158122	けれど (keredo)	But
0.889	169199	190421	けど (kedo)	But
0.885	939	1061	なんとなれば (nan to nareba)	
0.882	58957	66879	ならば (naraba)	If that is so
0.881	87425	99283	なのに (nanoni)	Then
0.881	5547111	6299085	でも (demo)	But

Figure 6.1: Japanese conjunctive words and expressions classified by ratio of apparition in sentence-initial position (part 1/3)

0.880	40896	46457	にもかかわらず (ni mo kakawarazu)	Despite that
0.878	66715	75984	それなのに (sore nanoni)	Despite that
0.876	85	97	と同じに (to onaji ni)	Like
0.869	5014	5770	だけども (dakedo mo)	Though
0.866	820	947	何となれば (nani to nareba)	
0.863	3684	4268	それにつけても (sore ni tsukete mo)	Be that as it may
0.862	30665	35567	と同時に (to nan ji ni)	At the same time
0.858	265885	309719	だけど (dakedo)	However
0.847	7879	9306	そればかりか (sore bacari ka)	In addition
0.847	27253	32170	ゆえに (yue ni)	Thus
0.845	1549747	1833479	だから (dakara)	So
0.839	582255	693775	では (deha)	Well
0.825	32564	39460	故に (yue ni)	Thus
0.821	23	28	其の代わり (sono kawari)	Instead
0.818	10447	12778	ついては (tsuite ha)	
0.814	14476	17786	でないと (denai to)	Otherwise
0.813	368306	453143	だって (datte)	Then again
0.802	12968	16169	それ故 (sore yue)	Thus
0.786	11	14	とどうじに (todoujini)	
0.785	61216	77965	なにしろ(nani shiro)	At any rate
0.781	36653	46951	それゆえ(sore yue)	Thus
0.779	120365	154539	というのも(toiunomo)	
0.775	8080	10421	でなければ (denakereba)	Or else
0.764	137079	179465	だからこそ (dakara koso)	For this reason
0.762	282	370	ですけども (desu keredomo)	However
0.757	24801	32777	そのかわり (sono kawari)	Instead
0.752	128098	170345	おまけに (omake ni)	On top of that
0.749	54621	72906	何しろ (nani shiro)	At any rate
0.745	38104	51155	その代わり (sono kawari)	Instead
0.743	12189	16407	それにもかかわらず (sore ni mo kakarawazu)	Nevertheless
0.742	252041	339553	それでは (sore de ha)	Well then
0.738	14865	20153	そのくせ (sono kuse)	And yet
0.711	5810655	8170338	そして (soshite)	And then
0.707	1796	2542	その代り (sono kawari)	Instead
0.691	55977	81006	それだけに (sore dake ni)	For that reason alone
0.678	1371624	2023412	しかも (shikamo)	Furthermore
0.656	68507	104473	かといって (ka to itte)	While at the same time
0.650	80	123	其れ故 (sore yue)	Thus
0.647	33	51	其れゆえ (sore yue)	Thus
0.646	396	613	ですけれど (desu keredo)	However

Figure 6.2: Japanese conjunctive words and expressions classified by ratio of apparition in sentence-initial position (part 2/3)

0.625	16517	26444	だからと言って (dakara to itte)	
0.617	259072	420168	それとも (sore to mo)	Or
0.615	16	26	其の代り (sono kawari)	Instead
0.599	14528	24257	それゆえに (sore yue ni)	Thus
0.588	43601	74114	だからといって (dakara to itte)	
0.568	3984	7008	さては (sate ha)	
0.556	3497	6291	さもなければ (samonakereba)	Otherwise
0.488	5311	10878	されど (saredo)	
0.438	3619	8263	さりとて (saritote)	
0.432	31061	71960	それでいて (soredeite)	But
0.429	2081	4849	さもなくば (samonaku)	If not
0.398	24939	62717	次いで (tsuide)	After that
0.255	70802	278069	もしくは (moshiku ha)	Or
0.251	15318	61013	はたまた (hatamata)	Or
0.224	1488	6629	さらば (saraba)	
0.138	78001	564755	または (mata ha)	Or
0.085	15169	177866	及び (oyobi)	And
0.081	3064	37627	且つ (katsu)	And
0.081	2888	35839	並びに (narabi ni)	And
0.070	3430	48913	若しくは (moshiku ha)	Or
0.055	1734	31440	ならびに (narabi ni)	And
0.052	17044	329389	又は (mata ha)	Or
0.048	11091	231756	および (oyobi)	And
0.033	11807	363275	かつ (katsu)	And
0.001	45	39059	が (ga)	But
0.000	0	79590	と(to)	And
0.000	0	6	もっとも (mottomo)	However
0.000	0	2	其のかわり (sono kawari)	Instead
0.000	0	1	いっ方 (ippou)	Meanwhile

Figure 6.3: Japanese conjunctive words and expressions classified by ratio of apparition in sentence-initial (head) position (part 3/3)

Chapter 7

Treatment of modality and complex verbal structures

Part of the material contained in this chapter has been already published by the author in (Kauffmann et al., 2011).

7.1 Introduction

In chapter 4, we studied the translation of simple sentences. In chapter 6, we focused on complex sentences containing at least two verbal clauses, but did not give any comment on complex verbal structures. Unlike simple verbal forms, complex verbal structures often require to go beyond the word-by-word translation level and select the appropriate verbal structure in the target language. This will often lead to a translation asymmetry between the source and target verbal structures. We will provide information about this point in this chapter, describing several complex verbal structures in English and Japanese, comparing them and proposing dedicated methods for their translation.

First, in section 7.2, we will describe possible translations of modals and other expressions of modality. Then, in section 7.3, we will focus on the treatment of passives and causatives, and in section 7.4, we will describe other complex verbal structures such as verbal objects and sentential objects[1]. Finally, in section 7.5, we will describe the evaluation process that was set up and look at the results.

[1]We mention here *sentential objects*, where the object of a governing verb is a verbal clause or a sentence, as in *I remember that it was a dark night.*. *Verbal objects* correspond to the simplest case of sentential object, where the object of a governing verb is a verb, often at the gerundive form, as in *I love skiing.*

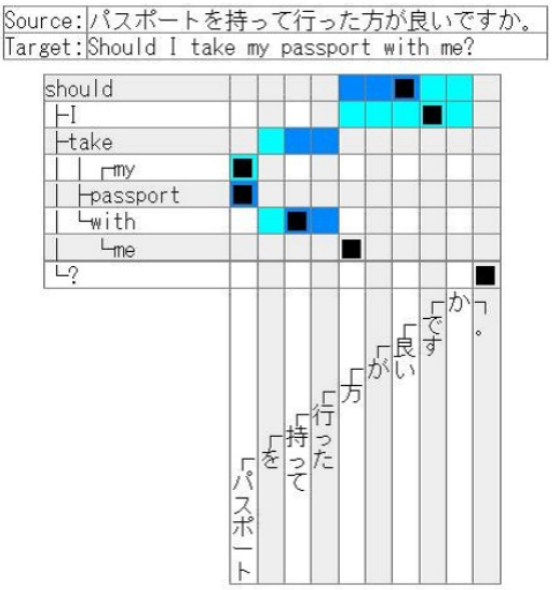


Figure 7.1: *Alignment between sentences with modality expression (Mochizuki et al., 2011)*

7.2 Modality

7.2.1 An overview of modality translation based on empirical data

This overview is based on Michiaki Mochizuki and Toshiaki Nakazawa's work on Japanese-English translation of modality. As they did in other studies (see Mochizuki et al. (2011)), Mochizuki and Nakazawa launched syntactic parsing tasks on aligned Japanese-English bilingual corpora. They obtained dependancy grammar trees of aligned sentences and therefore could deduce word and collocation translation from aligned trees.

Figure 7.1, which was published in (Mochizuki et al., 2011), shows an example of aligned sentences represented in dependency grammar trees. The authors explain that some mistakes in word alignment (represented by black squares) have been corrected using the multi-word expression alignment (blue and turquoise squares).

They applied this method on a 1'300'000 sentence-long travel domain aligned

corpus and performed a statistical study on the translation of modality. For a series of English and Japanese expressions related to modality, they recorded the different possible translations found in the corpus, ranking them by the percentage of occurrences.

Studying the results of this statistical analysis, and manually collecting more evidence in aligned bilingual corpora of newspaper articles depending on the syntactic context and on semantic constraints, the most frequent translations for every English modal or semi-modal have been set up.

7.2.2 Modality in English and Japanese

In English, modality is mainly expressed by the use of the English modals: *can, may, might, could, shall, should, will, would, must and ought.* They are a specific kind of auxiliaries with a very limited inflexion and have no finite form:

(7.1) He can speak!

Other English verbs or verbal expressions are also expressing modality, without being classified as *modals*. For example, *have to* or *be able to* are not classified as a modals but sometimes referred as *semi-modals* (see Palmer (1979) and Leech (2003)), a terminology that we adopt here. Semi-modals have different syntactic properties but are semantically equivalent to modals:

(7.2) He is able to speak.

Japanese expression of modality usually requires the use of some specific governing verbs or multi-word expressions, and sometimes specific moods for the verb of the object clause. Among the specific moods, the potential mood expresses capacity and is equivalent to the English modal *can* or semi-modal *be able to*. It does not require any governing auxiliary verb:

(7.3) 飲めた。
nometa
drink
(I) was able to drink.

The *desirative mood* is equivalent to the English verb *want to*. It has a limited inflexion, similar to adjectival inflexion, and can optionally be completed by a copula or stative verb. We can notice that the desirative mood is one of the Japanese modal forms, whereas *want to* is not necessarily considered as a semi-modal in English, but more often seen as a standard verb taking a verbal complement:

(7.4) 飲みたい。
nomitai
drink

(I) want to drink

Among multi-word expressions expressing modality in Japanese, we find 必要がない ("hitsuyou ga nai": there is no need to) where ない ("nai") is the negative form of the auxiliary ある ("aru") , that comes after the Japanese verb.

(7.5)

病院	に	行く	必要	が	ない。
byouin	ni	iku	hitsuyou	ga	nai
hospital	to	go	need	[nominative]	there is no

(You) don't need to go to the hospital

Another example of modal expression is the structure negative [conditional mood + semi-auxiliary なる ("naru") on negative form]. This leads to a typical ending in -なければならない ("-nakereba naranai"). We obtain there a double negation, which literally means *it shall not happen not to*. It is used in cases where *must* or *have to* would be used in English:

(7.6)

私は	働かなければ	ならない
watashi ha	hatarakanakereba	naranai
I	would not work	not happen

I have to work

The results of this overview has shown that English modals and other verbs related to modality expression have various possible translations in Japanese. Sometimes, six or seven possible translations exist. The translation of modality shows cases of both conflational, lexical and sometimes categorial divergences. For example *should* can either be translated into べきだ ("beki da") (ex. 7.7) when a wish from the speaker is expressed, はずだ ("hazu da") (ex. 7.8) or the usual simple form (ex. 7.9) when a possible event to happen is described, or いいですか? ("ii desu ka?") (ex. 7.10) when the source sentence is interrogative and expresses a suggestion.

(7.7) It should be continued when one considers the feelings of the victims. [2]

被害者	の	遺族	の	心情	も	考え合わせれば、
higaisha	no	izoku	no	shinjou	mo	kangaeawasereba,
victims	of	survivors	's	feelings	too	would be taken into consideration,
存続す		べきである。				
sonzoku su		beki dearu				
keep on		(we) should				

[2] Example taken from the English and Japanese editions of the Hiragana Times newspaper in December 1993. A more literal translation of the Japanese sentence would be "(We) should keep on taking the surviving victims' feelings into consideration."

(7.8) He should be here soon.

彼	は	もうすぐ	来る	はずです。
kare	ha	mousugu	kuru	hazu desu
he	[topic/subject]	soon	come	should

(7.9) He should be here any minute.

彼	は	もうすぐ	来る	よ。
kare	ha	mousugu	kuru	yo
he	[topic/subject]	soon	comes	[affirmation]

(7.10) How should I deal with the matter?

この	件	は	どのように	処理すれば	よいのだろう
kono	ken	ha	dono you ni	shori sureba	yoi no darou
this	matter	[topic/subject]	how	would treat	wood be good
か?					
ka?					
[question]?					

We should also notice that the selected modal expressions can be slightly modified for reasons of style or politeness level. In example 7.7, the copula is conjugated in the formal form and べきだ ("beki da") becomes べきである ("beki dearu"), and the verb 存続する ("sonzoku suru") is contracted and becomes 存続す ("sonzoku su"). The verb + modal structure becomes 存続すべきである ("sonzoku su beki dearu"). In example 7.8, the copula is at the polite form, so the modal expression becomes はずです ("hazu desu"). In example 7.9, the affirmative particle よ ("yo") is affixed to the verb. In example 7.10, the adjective いい ("ii") is written in the non contracted form よい ("yoi"), the copula is at the volitional form だろう ("darou"), and the explanatory particle の ("no") is added after the adjective. We obtain the interrogative modal expression よいのだろうか ("yoi no darou ka"). Many other possible stylistic variations exist for this expression: いいのですか ("ii no desu ka"), いいんですか ("iin desu ka"), いいですか ("ii desu ka"), いいんでしょうか ("iin deshou ka"), よいのでしょうか ("yoi no deshou ka"), etc. Given the high number of possible translation for the same English modal expression (a problem evoked in Murata et al. (2005)), the difficulty of the translation task consists in choosing the ones that should be selected by the MT system in the right situation.

7.2.3 Rules and implementation

Rules for the translation of English modals and for the translation of semi-modals and other expressions of modality have been defined, on the basis of

the result of our overview of modality translation. The default translations have been defined, choosing those which were appearing the most frequently in the corpus. However, for the cases where the meaning of the source modal expression was too ambiguous, we have tried to choose the translations that would correspond to the larger range of meanings.

Translation of English modals

In the following examples, structures made of an English modal + a verb are translated in Japanese by structures such as "verb + verbal expression" or just "verb". We will consider English modals as fixed forms, with no morphological variation except positive or negative polarity. Hence, the generated Japanese verbal expressions presented will have a fixed tense too, and will have a morphological variation depending on positive or negative polarity. An agreement depending on politeness level may be possible too, and will be left for further research.

The first rules show the possible translations for *shall, can, would, may*, and *might*.

Shall:	- in a question with pronoun "we" or "I": conjugate the Japanese verb in the volitional form -ましょう (-mashou) and add the question particle か (ka); - otherwise: write the Japanese verb at the usual form.
Can:	write the Japanese verb at the potential mode.
Would:	- in "would like", refer to the specific rule; - in a question: write the Japanese verb at the gerundive form -て (-te) and add the Japanese humble auxiliary いただけます (itadakemasu) with the question particle か (ka): いただけますか (itadakemasu ka); - otherwise, especially for reported speech, no change with the usual form.
May :	- in "May I have... ?": refer to the specific rule; - in a question: write the Japanese verb at the potential form; - otherwise: add the verbal expression かもしれない (ka mo shirenai) after the Japanese verb.
Might:	as for "may", add the verbal expression かもしれない (ka mo shirenai).

The second rules show the possible translations for *must*, depending on the uses. *Must* can either express duty/prohibition (deontic meaning) or express prediction (epistemic meaning). In the studied corpora, *must* was more often used to express obligation than strong prediction.

The next rules show the possible translations for *should*. *Should* is ambiguous as it can both express an information about a fact that will probably happen

Must: - to express obligation: conjugate the Japanese verb in the negative conditional form -なければ (-nakereba), and add the Japanese auxiliary with negative polarity ならない(naranai).
- to express prohibition ("must not"): put the Japanese verb at the potential form, with negative polarity.
- to express a prediction on past events ("must have" + past participle), put the Japanese verb at the past form, and then add the Japanese verbal expression にちがいない (ni chigainai);
- to express a strong prediction, add the verbal expression にちがいない (ni chigainai).

(epistemic meaning), or a wish from the speaker (deontic meaning). However, the second case is more frequent in both travel domain and news corpora.

Should: - to express a wish from the speaker: add the verbal expression べきだ (beki da) after the Japanese verb.
- to tell an information which is not completely sure: no change with the usual present/future expression.

Could is highly ambiguous because it can refer to the past or to hypothetical future, depending on the context. In travel domain and newspaper article corpora, hypothetical future was much more frequent than past. In the travel domain corpus, polite requests were much more frequent than questions related to the past.

Could: - in a polite request (especially with "could you"): write the Japanese verb at the gerundive form -て(-te) and add the Japanese humble auxiliary いただけます(itadakemasu) with the question particleか (ka): いただけますか (itadakemasu ka);
- in a subordinate clause, after "if": write the Japanese verb at the -たら (-tara) conditional form;
- to mention an hypothetical future: write the Japanese verb at the potential form;
- to mention a past fact: write the Japanese verb at the past potential form.

Semi-modals and other expressions of modality

The rules for the translation of English semi-modals are quite similar to the rules for modals. The main difference is that semi-modals have standard morphological inflexion, whereas modals have no inflexion. Consequently, the generated translation of semi-modals must be conjugated in the right tense, depending on the source semi-modal tense. This tense transfer is similar to standard tense tranfer and does not require specific rules. We show here the rules for the translation of semi-modals *have to*, *need to* and *want to*. We have included *want to* in the semi-modal list, even if it is often considered as a standard verb.

Want to: write the Japanese verb at the desirative mode.
Have to: - positive (have to) : as for "must", conjugate the Japanese verb in the negative conditional form - なければ (-nakereba), add the Japanese auxiliary with negative polarity ならない(naranai);
- negative (don't have to) : add the verbal expression 必要がない (hitsuyou ga nai) after the Japanese verb.
Need to/Need: add the verbal expression 必要がある (hitsuyou ga aru) after the Japanese verb.

The overview has also shown the possible translations of other expressions related to modality such as *would like to* and *please*.

Would like to: put the Japanese verb at the desirative mode.
Please: when please associated to a verb, in a polite order or a request, conjugate the verb in the gerundive form - て (-te) and add the polite expression 下さい (kudasai).

We have also included translations for polite expression containing English modals such as *May I have* and *I would like*.

"May I have": add the particleを ("wo") after the noun and the verbal expression 下さい (kudasai).
Would like (before a noun):
- in a question: add the question いかがですか (ikaga desu ka) after the noun;
- otherwise: add the verb にする (ni suru) after the noun.

Implementation and tests

The implementation has been organised in three steps: the improvement of the transfer module, the adaptation of the lexicon and the adaptation of the generation module.

The improvement of the transfer module consisted in the implementation of the transfer rules. It has been done at the tense form transfer level, allowing the rules to be applied when a modal structure is detected in the source sentence. The lexicon has also been adapted consequently, in order to follow the rules and to enable the translation of English semi-modals. Then, the generation module has been adapted, in order to generate the chosen tense forms.

We have only been able to implement over 65 percent of the lexicalised transfer rules presented for modality. Among the remaining 35 percent, some could not be added because of a lack of semantic content in the source content analysis. For instance, the Fips analyser cannot differentiate deontic or epistemic uses of English modals. To do so would require semantic or pragmatic knowledge that is beyond the scope of the Fips parser so far.

Tests have shown mixed results for the translation of modals and semi-modals. The rules have enabled the system to generate grammatically correct outputs, as in example 7.11. However, the lack of sensibility of the system to the semantic and pragmatic context tends to allow the generation of outputs that do not convey well the original source sentence meaning, as in ex. 7.12. So, our approach has shown advantages but it would need to be improved and completed in order to avoid oversimplification of the modality translation problem.

(7.11) Input: They may not arrive.

Output:	人々	は	到着	しない	かもしれない。
	hitobito	ha	touckaku	shinai	kamoshirenai
	they	[topic/subject]	arrive	will not	maybe

(7.12) Input: I think that he must arrive soon.

Output:	*私	は	彼	が	速やかに	到着
	watashi	ha	kare	ga	sumiyaka ni	tochaku
	I	[topic/subject]	he	[nominative]	soon	arrive

しなければならない	と	思う。
shinakereba naranai	to	omou.
must	that	think

literally: I think that he has to arrive soon.

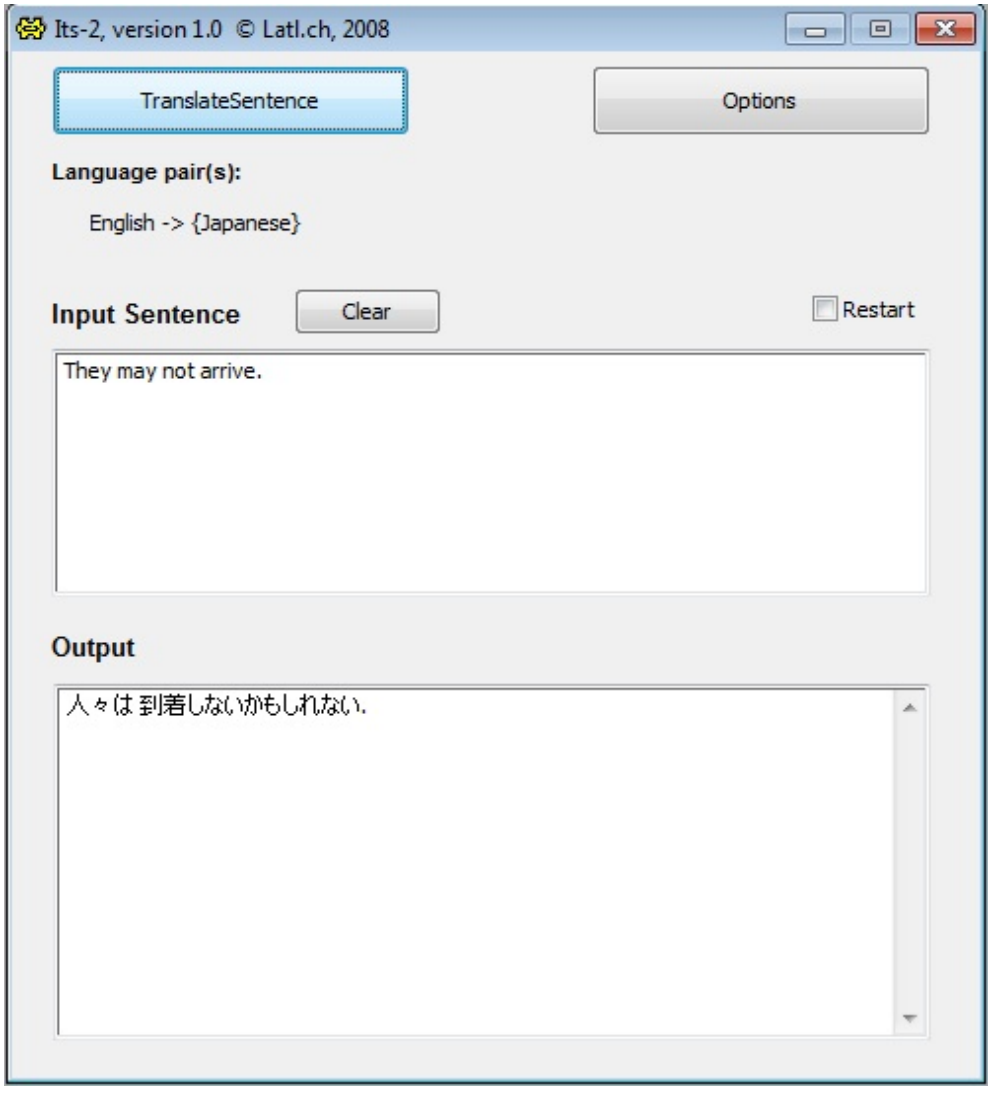


Figure 7.2: *Translation of modality by Its-2*

7.3 Passives and causatives

In English, passive is formed with the auxiliary *be* and the past participle and causative is formed with semi-auxiliaries such as *make*, *let* or *have*. In Japanese, these two moods are expressed with specific verb forms (Farmer (1984); Tsujimura (1996)).

7.3.1 Passives

When a passive clause with a direct transitive verb is translated from English to Japanese, the surface structure subject (which corresponds to the deep structure direct object) is followed by a が ("ga") particle, which is sometimes replaced by a topic particle such as は ("ha"). The agent, if it is mentioned, is followed by a に ("ni") or によって ("niyotte") particle. The Japanese verb is conjugated in the passive form.

(7.13) The book was read by Takahashi.

本	が	高橋	に	読まれた。
hon	ga	takahashi	ni	yomareta
book	[subject]	Takahashi	by	was read.

If a passive clause with an indirect transitive verb is translated from English to Japanese, the same phenomenon happens: the surface structure subject (which corresponds to the deep structure indirect object) is followed by a が ("ga") particle and the agent, if it is mentioned, is followed by a に ("ni") particle. If a verb is ditransitive, as in the following example, the direct object is followed by a を ("wo") particle.

(7.14) Mitsuko has been sold bad sushis.

光子	が	不味い	鮓	を	売られた。
mitsuko	ga	warui	sushi	wo	urareta
Mitsuko	[subject]	bad	sushi	[object]	has been sold

7.3.2 Causatives

When a causative clause is translated from English to Japanese, the agent (*Momoko* in the following example) is followed by a に ("ni") particle. The Japanese verb is conjugated in the causative form.

(7.15) Mitsuko made Momoko drive the car.

光子	が	桃子	に	車	を	運転させた。
mitsuko	ga	momoko	ni	kuruma	wo	untensaseta.
Mitsuko	[subject]	Momoko	[agent]	car	[object]	made drive

If the verb does not have a direct object, the agent can either be followed by a に ("ni", as in example 7.16) or a を ("wo", as in example 7.17) (Farmer (1984); Tsujimura (1996)).

(7.16) Mitsuko made Momoko drive.

光子	が	桃子	に	運転させた。
mitsuko	ga	momoko	ni	untensaseta.
Mitsuko	[subject]	Momoko	[agent]	made drive

(7.17) Mitsuko made Momoko drive.

光子	が	桃子	を	運転させた。
mitsuko	ga	momoko	wo	untensaseta.
Mitsuko	[subject]	Momoko	[object/agent]	made drive

7.3.3 Implementation and tests

As it has been done for modals and semi-modals, the appropriate forms have been assigned to the Japanese verbs when passives or causatives were detected in the source sentence. Correct particles have been assigned the Japanese arguments, always choosing the に ("ni") particle for the agents.

Tests have shown correct results (as in Figure 7.3.3), except in cases of indirect transitive passives where the syntactic parsing was not giving a correct analysis.

7.4 Other complex verbal structures

In this section, we will describe the translation mechanism for complex verbal structures which are not related to modality expression or passive and causative voices. We will especially focus on verbal and sentential complements, and also mention other cases of gerund and infinitive translations.

7.4.1 Comparison between English and Japanese structures

Looking at Japanese-English aligned travel domain and news corpora, we have compared complex verbal structures in both languages. Except for the usual differences in word and component order, many similarities can be found between English and Japanese syntactic structures.

For example, the structure [governing verb + complementizer + completive clause] exists in English, and the same structure in reverse order: [completive clause + complementizer + governing verb] exists in Japanese. The following example illustrates this phenomenon:

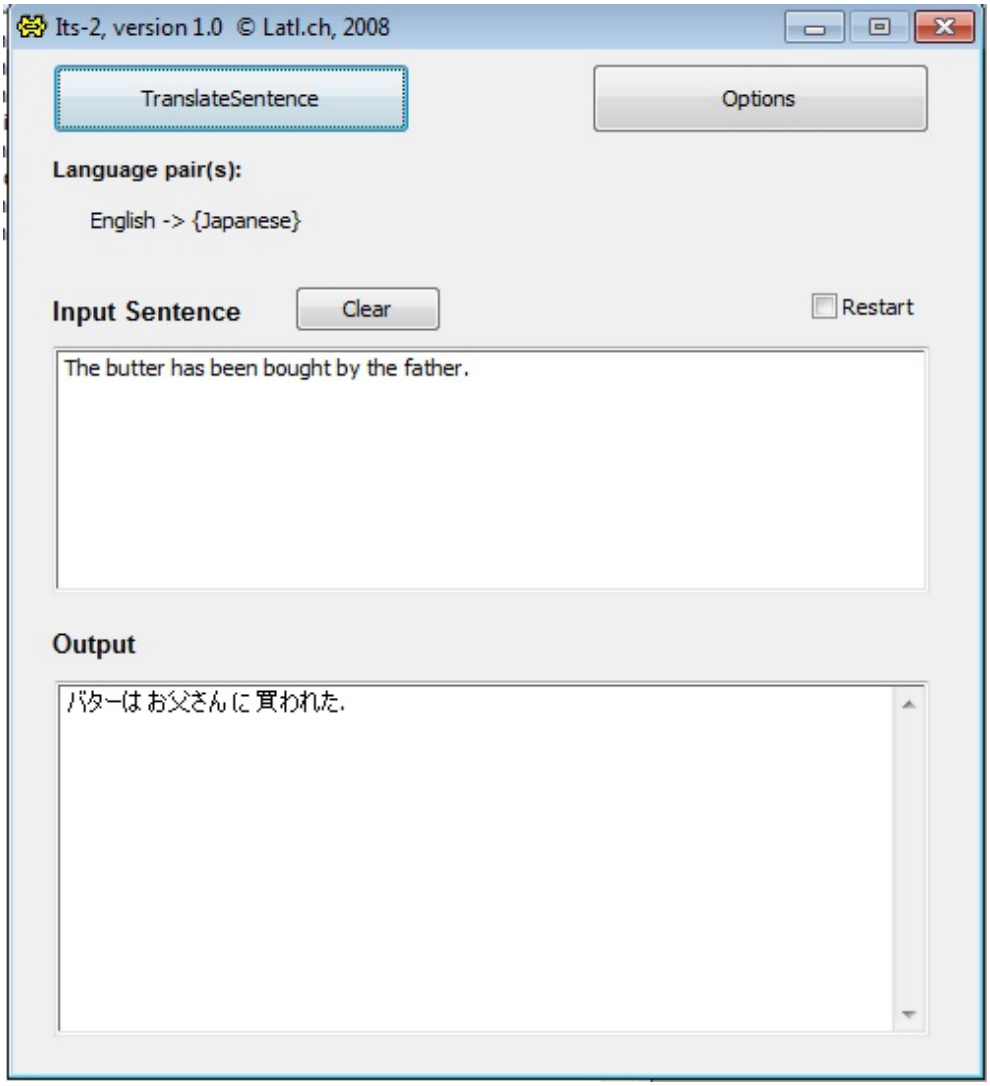
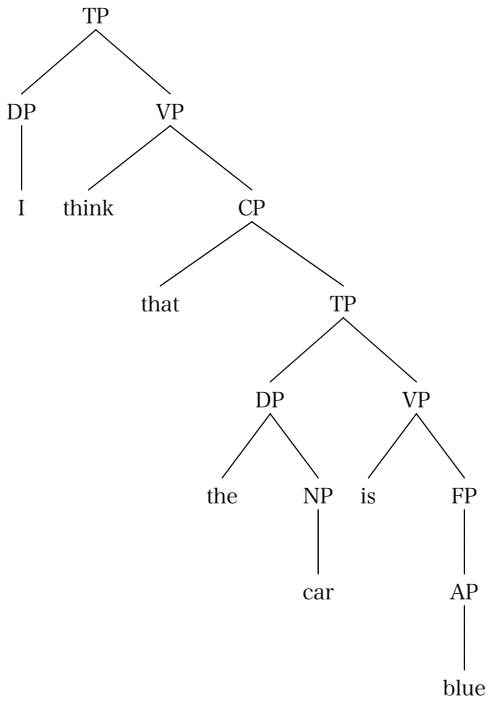


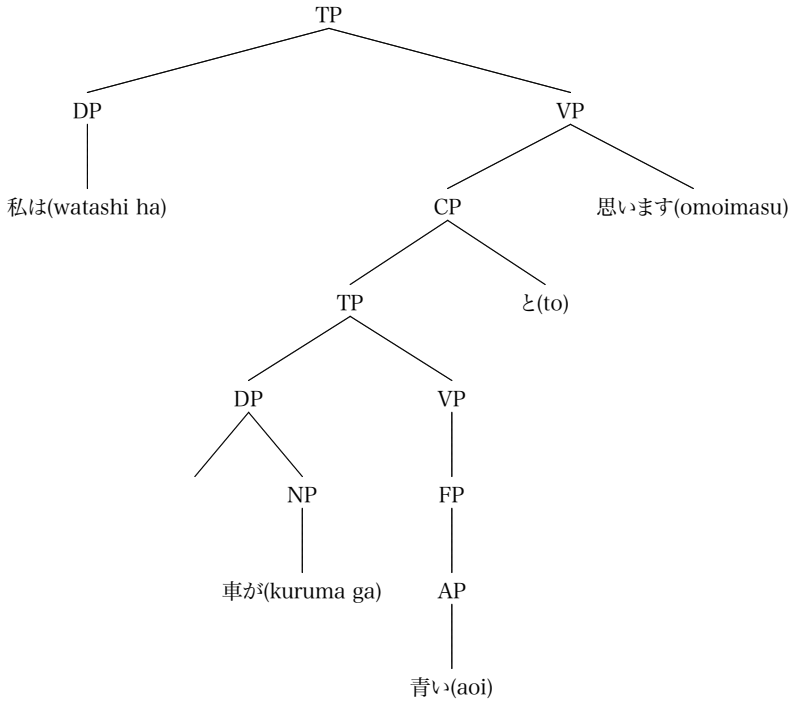
Figure 7.3: *Example of passive clause translation with Its-2*

(7.18) (a) I think that the car is blue



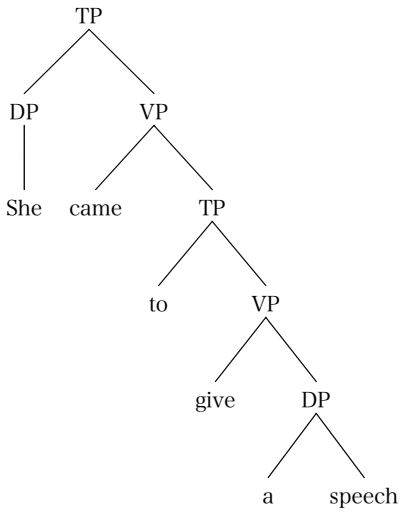
(b)

私	は	車	が	青い	と	思います。
watashi	ha	kuruma	ga	aoi	to	omoimasu
I	[topic]	car	[nominative]	blue	that	think

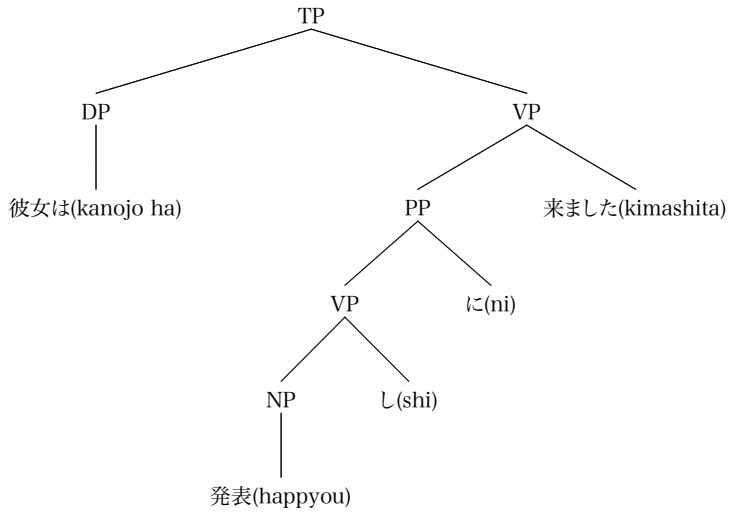


The structure [governing verb + preposition + verb base] also exists in both languages, in reverse orders. It becomes [verb base + postpositional particle + governing verb] in Japanese:

(7.19) (a) She came to give a speech.



(b) 彼女 は 発表 し に 来ました
kanojo ha happyou shi ni kimashita
she [topic] speech do to came



In Table 7.4, an overview of syntactic structures with verbal complements in English and Japanese is given. Globally, a symmetry can be observed between structures in the two languages. Japanese structures end with the governing verb while English ones start with the governing verb. However, minor differences appear sometimes between syntactic structures that are not fully symmetric.

For example, the Japanese nominalisation process consists of a sentence followed by a nominalisation noun or particle. In English, the most similar structure is determiner phrase *the fact* followed by the conjunction *that* and a sentence. The complete Japanese structure is this one : [sentence + nominalisation Noun+ object particle + governing verb], and the complete English one is : [governing verb + "the fact that" + sentence]. The main difference is that there is no conjunction in the Japanese structure and no object case marker in the English structure. This is consistent with the usual syntax, as there are no object case markers in English, and no conjunction after Japanese relative clauses.

7.4.2 Translation of verbal and sentential objects

Even if many syntactic similarities can be observed between existing structures, a study of aligned bilingual corpora shows many cases of structural asymmetry (corresponding to structural divergences in Dorr's classification), where the translation chosen for the English complex verbal structure is not the most syntactically similar one.

English structures	Japanese structures
Governing Verb (+ Conjunction) + Indicative Sentence	Sentence + と (conjunctive particle ≪ to ≫) + GV
GV + Conjunction + Subjunctive Sentence	Conditional Sentence + GV
GV + ″to″ + Infinitive Clause	Base Verb Clause+ に (postpositional particle≪ ni ≫) + GV
GV + Gerundive Clause	Gerundive Clause + GV
GV + ″the fact that″ + Sentence	Sentence + Nominalisation Noun+ object particle + GV

Figure 7.4: Comparison of verbal and sentential object structures

For instance, we can see in sentence 7.20 that the verb *hope* can take a completive clause introduced by the conjunction *that*. In Japanese, the equivalent sentence, with the verb 望む ("nozomu": hope), does not include a completive clause followed by a conjunctive particle (as in example 7.18) but has a nominalised completive clause instead.

In example 7.21, the coordinated verbs of the completive clause of the English sentence are at the gerundive form. In the Japanese sentence, again, the structure is slightly different. Only the first of the three verbs is at the gerundive form. The second one at the verb base form and the last one is at the past form. The whole completive clause is nominalised with the noun こと ("koto": fact), and becomes a direct object of the verb 覚える ("oboeru": remember).

(7.20) [3] I hope that Thai people and Japanese can understand each other better.

私	は	タイ人	と	日本人	が	もっと
watashi	ha	taijin	to	nihonjin	ga	motto
I	[topic/subject]	Thai people	and	Japanese people	[nominative]	more

理解し合う	こと	を	望みます。
rikai shi au	koto	wo	nozomimasu
understand each other	[nominalisation]	[object]	hope

(7.21) I remember my parents coming here and spending months and loving it.

両親が	ここ	を	訪れて	数カ月
ryoushin ga	koko	wo	otozurete	soukagetsu
parents [nominative]	here	[object]	visit (-Te form)	several months

[3] Example taken from the English and Japanese editions of the Hiragana Times newspaper in January 1990.

滞在し、		とても	気に入っていた	こと
touzai shi,		totemo	ki ni itteita	koto
stay (verb base form),		really	loved (continuous past form)	[nominalisation]
を	覚えている。			
wo	oboeteiru			
[object]	remember (continuous present)			

Such examples remind us that verbal or sentential complement tense form selection depends mostly on the lexical properties of the governing verb. The aim of the translation task for verbal or sentential complements is to generate a structure that corresponds to the selectional features of the governing verb, and to choose the most appropriate one when several possible choices exist.

7.4.3 Translation of other gerunds and infinitives

Apart from the object position, infinitives and gerunds can also be found at other positions in the English sentence. For example, they can be found in subject position (see example 7.22 and 7.23), or in adverbial complement position (example 7.24). In both cases, their translation is independent of the syntactic valence of the main verb of the clause.

(7.22) To know is the most important thing.

理解する	の	は	一番	大切な	こと	だ。
rikai suru	no	ha	ichiban	taisetsuna	koto	da
to know	[nominalisation]	[topic/subject]	most	important	thing	it is

(7.23) Fighting will help you to be free.

もめれば	自由人	になる。
momereba	jyuujin	ni naru
fight (conditional)	free person	become

(7.24) I came listening to my mp3 player.

私	は	mp3プレーヤー	を	聞きながら	行きました。
watashi	ha	mp3 player	wo	kikinagara	ikimashita
I	[topic/subject]	mp3 player	[object]	listening (to)	went

Our study of the bilingual corpora has shown that the translation of English gerunds and infinitives in those cases is subject to many possible patterns. Even if the translation structure types depend on the translator's choice and are not clearly predictable, tendencies can still be drawn. When a gerundive or infinitive form is on subject position in English, the translation into Japanese is more likely to contain a nominal form that can either be a nominalised simple form, a

verb base form or a noun. When an English gerund is in adverbial complement position, it is more likely to be translated into one of the Japanese gerundive forms, such as the -て ("-te") form, the -たり ("-tari") form or the -ながら ("-nagara") form.

We have defined transfer rules, which may not always detect the best translation pattern choice, but should enable to generate grammatically correct sentences:

```
if the English gerund is in adverbial complement position
    if it is expressing simultaneity
        write the verb at the -ながら ("-nagara") gerundive form
    else
        write the verb at the -て ("-te") gerundive form
else, if the English infinitive or gerund is in subject position
    write the verb at the simple form, add the nominalisation particle の (no),
    and add the subject particle は (ha) or が (ga).
```

7.4.4 Implementation and tests

As we mentioned in chapter 3.5, information on verb subcategorisation classification has been stored in the lexical database. This information has been especially useful for the translation of verbal and sentential objects, as their conjugation depends mainly on the governing verb properties.

The data was extracted from two sources: mainly from the Case Frames file and also from the dictionary used by the Juman analyser (Kurohashi and Nagao, 2003). The data from the Case Frames file showed that verbs with an argument with conjunctive particle と ("to") usually take a sentential object, and that transitive verbs can take a nominalised sentential object. The file, extracted from the dictionary of the Juman analyser, contained a list of Japanese semi-auxiliaries and light verbs that take verbal objects, mentioning the possible tense form of those verbal objects. It can either be the base verb form or gerundive in -て ("Te" form), as in example 7.25, with semi-auxiliary みる ("miru").

After the compilation of all the information on verbal and sentential complements conjugation, the data, with all the conjugation parameters, was inserted into the lexicon. Then, after this improvement of the lexical database, the MT system has been modified, in order to use the lexical information during the translation process. In the transfer module, procedures that assign the tense forms specified in lexical information of the governing verb to the Japanese verbal objects have been implemented. In the generation module, the verbal

complements are generated with those tense forms, and post-verbal nominalisation particles are added where it is required.

Globally, these improvement have enabled the MT system to generate better translation outputs. For example, in ex. 7.26, 7.25 and Figure 7.5, the English sentences have been translated into Japanese with a correct tense selection in verbal and sentential objects. Still, even if tests have shown good results, some further work would still be required to reach a perfect output. For example, a use of language models would enable to choose between several candidates when the tense of the target sentence verb cannot be clearly determined by syntactical rules.

(7.25) Input: I will try to use it.

Output: 使ってみる。
tsukatte miru
using try

(7.26) Input: He said that the weather is very good.

Output:	彼	は	天気	が	とても	いい	と
	kare	ha	tenki	ga	totemo	ii	to
	he	[topic/subject]	weather	[nominative]	very	good	that

言った。
itta.
said

For the cases where gerundive or infinitive forms are found in subject or adverbial complement position, the rules mentioned above have been applied. These rules have been implemented in the transfer module and the generation of the selected tense forms has been allowed in the generation module. However, a method that would be more sensitive to the sentence context could give a larger range of possible translations for gerunds and infinitives. This should be looked at in future work.

7.5 Evaluation and results

We launched an automatic evaluation of the MT system. We compared BLEU scores[4] obtained by three versions of the MT system: a older version without complex sentence handling or verbal object and modality handling, an intermediate version with complex sentence handling, and a more recent version with

[4]The BLEU score is a method for automatic evaluation of MT system translation quality(Papineni et al., 2002)

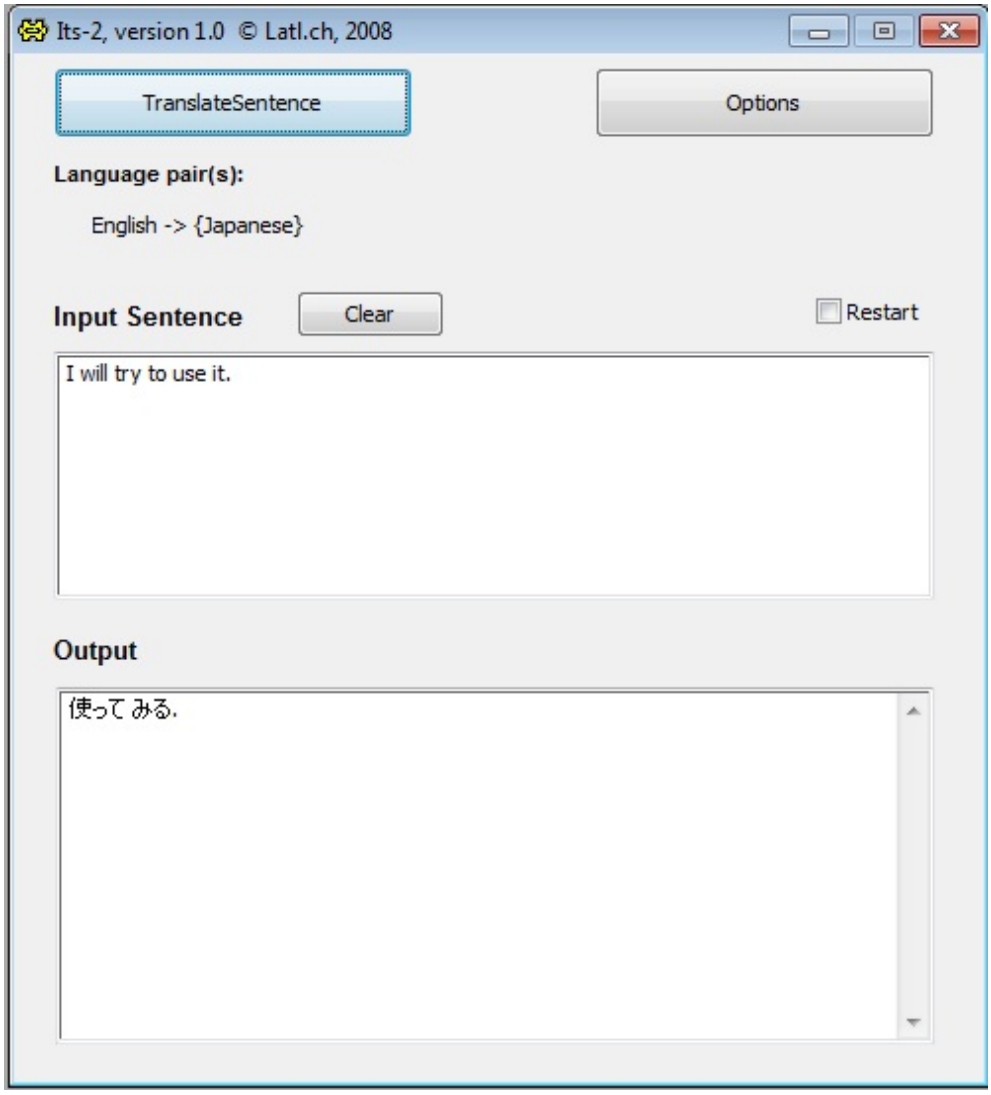


Figure 7.5: *Verbal object translation by Its-2*

Its-2 version	BLEU score
without complex sentence handling or verbal object handling (Dec. 2009)	2.46%
with complex sentence handling (July 2010)	2.49%
with complex sentence handling and verbal object handling (Jan. 2011)	2.52%

Figure 7.6: *Evaluation of Its-2 with BLEU scores obtained on scientific paper abstract translation*

complex sentence handling and verbal object and modality handling. The experiment was launched on a sample of 500 sentences of a scientific paper abstract corpus, available in English and Japanese. The Its-2 output was not segmented into words. As the BLEU score is computed on word comparison, we had to use the Japanese morphological analyser Juman (Kurohashi and Nagao, 2003) to segment the output. We used the evaluation tool available on the NIST website[5] for the BLEU score evaluation. The results showed a progression of +0.03% with the treatment of complex sentences and +0.06% with the treatment of complex sentences, modality and verbal objects, reaching a final BLEU score of 2.52%[6].

Even though the MT system BLEU scores have increased, they remain much lower than the ones obtained by state-of-the-art systems in recent experiments[7].

[5]http://www.itl.nist.gov/iad/mig/tests/mt/2009/

[6]BLEU can vary depending on the test corpus. For example, in an experiment on English-to-Japanese translation of 1000 sentences of the Tatoeba Corpus (that mainly contains everyday language), the June 2012 Its-2 version has obtained a BLEU score of 4.47%. The difference with the scores obtained for scientific paper abstract translation is essentially due to the lower complexity of the sentences and vocabulary in the Tatoeba corpus, compared to those of the paper abstract corpus.

[7]As BLEU scores vary not only depending on the language pair and test corpus, but also on the type of MT system and the number of reference translations used, it is not appropriate to compare BLEU scores between different systems obtained on different test sets. Still, we know that a state-of-the-art English-to-Japanese SBSMT system trained on 3'358'635 reference sentences can reach a BLEU score of 60.46% on Information Technology (IT) domain sentence translation, using two reference translations in the evaluation (Lee et al., 2010). This score would surely have remained high but been lowered down if the evaluation had been carried out with only one reference translation. A baseline English-to-Japanese HMT system, trained on 48'000 sentences, can reach a BLEU sore of 27.88% on everyday language translation, using one reference translation in the evaluation (see Section 8.5).

Even if BLEU scores are usually lower for rule-based or linguistic based MT than for SMT or methods based on machine learning (Callison-Burch et al., 2006), it seems that the major reason for the low level of the obtained scores is the low quality of generated output, which is not as good as state-of-the-art ones (see Lee et al. (2010)).

A linguistic analysis of the generated output showed both a clear improvement on the treatment of the mentioned syntactic structures, and a difficulty to generate fluent output for complex input sentences.

7.6 Conclusion

In this chapter, we have described methods for the translation of modality and complex verbal structures. Our methods are based on syntactic and lexical knowledge, sometimes using data acquired by statistical methods from large corpora or from the web.

On modality, our method has the advantage of generating syntactically correct outputs, when the transfer rules are applied. However, it sometimes causes errors in translation on a pragmatic or semantic point of view. The method should be improved by a more advanced implementation of the rules and a deeper understanding of the sentence context, in order to avoid an oversimplification of the problem.

On verbal and sentential objects, the use of the verb lexical properties has enabled a good selection of tense forms. Some ambiguous cases still remain, and a statistical correction may help to perfect the selection in those cases.

On infinitives and gerundives in subject or adverbial complement position, our method has generated correct translation outputs. The main drawback has been the high homogeneity of the output, that does not reflect the large choice of possible translations existing for a human translator in such cases.

The results of the evaluation have shown a clear improvement that confirmed the validity of the method. However, even if the translation of modality and complex verbal structures has been improved, we have not been able to obtain a good quality of translation for a wide range of complex sentences. This can be explained by several factors, such as errors occurring at the syntactic parsing level, insufficiencies in the English-Japanese bilingual lexicon, syntactic rules that have not been fully implemented in the system, and a lack of bilingual data for the translation of collocations and other multi-word expressions. Those problems should be treated in future work.

Chapter 8

Conclusion and future directions

8.1 Research Questions

As has been said in Section 1.6, the research questions which have motivated this study are the following:

- Can a multilingual linguistics-based MT system such as Its-2, that has been usually used for MT between Western languages, adapt well for translation into a Far-East Altaic language like Japanese?
- Can a procedural linguistics-based approach without interlingual representation overcome syntactic structure differences and structural asymmetries?
- Can statistically obtained data help a linguistics-based MT system to handle structural asymmetries?

We will give conclusions about the first question in Section 8.2, the second question in Section 8.3 and the third question in Section 8.4. In Section 8.5, we will describe possible future directions of research related to the thesis work. Finally, we will give a summary of the contributions of the thesis in Section 8.6.

8.2 Conclusions with respect to the first research question

Throughout the different chapters of this thesis, we have shown that it has been possible to adapt Its-2 for English-to-Japanese translation, without modifying

the English parsing module, developing an English-to Japanese transfer module and a Japanese generation module and storing appropriate lexical data.

However, the quality of the English-to-Japanese translations generated by Its-2, as in the experiment presented in Chapter 7, is lower than the one usually obtained when Its-2 is used for translation between Western languages. Obviously, Its-2 is more affected by errors of syntactic parsing in English-to-Japanese translation, which requires deep reordering, than in translation between Western languages, which only requires limited reordering, because deep reordering can hardly be achieved well on the base of an incorrect parsing.

Therefore, we can conclude that a multilingual LBMT system can adapt for translation into a Far-East Altaic language like Japanese, but that English-to-Japanese translation with Its-2 requires a higher cost in development of the transfer phase than translation between Western languages. Moreover, in the case of complex sentence translation, the results cannot be as high as the ones obtained in translations between Western languages, due to a much higher dependency on syntactic parsing accuracy.

This dependency is not a consequence of the Japanese language structure itself, but of the structural difference between English and Japanese. We can reasonably assume that if the Its-2 parsing component was adapted for a language such as Korean which is syntactically closely related to Japanese, Its-2 could easily adapt for Korean-to-Japanese translation, and give results that would be equivalent to the ones that can be obtained in, for example, English-to French translation.

8.3 Conclusions with respect to the second research question

We have shown in Chapters 4, 6 and 7 that the Its-2 LBMT system can take advantage of the source sentence syntactic parsing to handle syntactic structure differences between English and Japanese, especially for word and constituent reordering.

We have shown that structural asymmetries can be considered and that Its-2 can generate asymmetrical translation in cases where it is required. We focused on asymmetries in verb subcategorisation in Chapters 3 and 7, on grammatical translational ambiguities (which can be considered as a specific type of asymmetry in semantic content) in Chapters 4 and 6, and on asymmetries in word category in Chapter 5 and 7.

A good treatment of these phenomena has been obtained in the translation of example test sentences. Even if the results obtained by Its-2 on evaluations of travel blog page translations (Chapter 6) or scientific paper abstract translations (Chapter 7) remain globally of poor quality, they still highlight the capacity

of the system to achieve reordering and generate asymmetrical translations, showing clear improvements in comparison to earlier versions of the system.

We can therefore conclude that, at an advanced state of development, a procedural linguistics-based approach without interlingual representation can overcome syntactic structure differences and a large part of structural asymmetries in translation, provided it has access to rich lexical data and a robust and accurate syntactic parsing component.

This was indeed known since the ARIANE MT system (Vauquois and Boitet, 1985) and the Mu MT system (Nagao and Tsujii, 1986), and has then been confirmed by the lasting success of English-Japanese non-interlingual LBMT systems such as Honyaku Pikaichi, Yakuse!! Goma, The Honyaku, or Yakushite.net, all developed several years ago and still currently available and, for some of them, still constantly enhanced.

8.4 Conclusions with respect to the third research question

As mentioned in Chapters 3, 6 and 7, we have made use of lexical data obtained and classified by statistical methods (extracted either from text corpora or from the web) for the improvement of the Its-2 lexical database. The data used in Chapter 3 and 7 was dedicated to Japanese verb subcategorisation; the data used in Chapter 6 was dedicated to Japanese conjunctive word classification; the data used in Chapter 7 was dedicated to the expression of modality in English and Japanese.

The insertion of such statistically obtained lexical data has considerably enrich the lexical database, hence helping to improve the quality of the translations generated by Its-2. It has allowed the system to generate correct asymmetrical translations of verb subcategorisation frames. However, only a limited part of the results obtained for modality translation were correct, because our lexical selection system has not been sufficient to solve all the ambiguous cases, where several translations of modal expressions were possible. A statistical or semantic correction would probably help reduce the error rate with those ambiguous cases.

We can then conclude that the use of statistically obtained lexical data can help a linguistics-based MT system to handle structural asymmetries, but that it should be combined to a context-dependent lexical selection when several translations are possible.

8.5 Future directions

In this section, we propose possible directions for future works related to English-Japanese MT with the Its-2 system. We first present possible improvements on the lexical data; after that, we describe a possible selection of the Japanese politeness level. Finally, we evoke possible ways of hybridisation for the system.

8.5.1 Lexical data

Collocations and other multi-word expressions

A domain which has not been deeply investigated in our experiments is the translation or generation of collocations and other multi-word-expressions. This is a crucial question in MT (see Baldwin and Bond (2002)) and would need to be dealt with in future research about Its-2 English-to-Japanese MT.

The architecture of the Its-2 lexical database allows the insertion of collocations or other multi-word expressions. In order to illustrate how valuable correct collocation translation can be, we have carried out a comparison of the translation of 15 test sentences before and after the insertion of the required collocations into the database. All of the translations have been improved, and 14 out of 15 are completely correct after the collocation insertion.

In example 8.1, we can see that the English collocation *take a nap* was incorrectly translated before it was inserted into the lexicon (example 8.1.b). After the insertion, it is correctly translated into the Japanese collocation 昼寝する ("hirune suru", example 8.1.c).

The current translation of *wear* into Japanese depends on the nature of its direct object: for example, if the object is *shirt* as in example 8.2, it is translated into the default translation 着る ("kiru"); if it is *hat* as in example 8.3, the translation is かぶる ("kaburu"). The generation of such Japanese object-verb collocations has been made possible by the collocation insertion.

(8.1) (a) I should take a nap

(b)

*私	は	眠り	を	取る	べきだ.
watashi	ha	nemuri	wo	toru	beki da
I	[topic/subject]	sleep	[object]	take	should

(c)

私	は	昼寝	する	べきだ.
watashi	ha	hirune	suru	beki da
I	[topic/subject]	nap	do	should

(8.2) (a) I am wearing a black shirt.

(b)

私	は	黒い	シャツ	を	着ている.
watashi	ha	kuroi	shirts	wo	kiteiru
I	[topic/subject]	black	shirt	[object]	am wearing

(8.3) (a) I am wearing a hat.

(b)

私	は	帽子	を	かぶっている.
watashi	ha	boshi	wo	kabutteiru
I	[topic/subject]	hat	[object]	am wearing

Large lexical English-Japanese databases such as (Eijiro, 2012) contain collocations, coocurrences or other multi-word expressions. The insertion of a consequent amount of existing data about multi-word expressions into the Its-2 lexical database would help to handle related translation asymmetries, as in example 8.1, and would allow a global improvement of the translation quality.

Related language pairs

Another possible future direction related to lexical data is the development of lexicons for other language pairs. The architecture of Its-2 allows facilitated addition of language pairs related to the ones already implemented, if lexical resources are available.

For example, French-to-Japanese MT has been developed on the basis of Its-2 English-to-Japanese MT (see Chapter 5), but its results have been limited by the deficiencies of the current version of the Its-2 French-Japanese lexicon. A further development of this lexicon would allow a facilitated improvement of the Its-2 French-to-Japanese MT, on the basis of the current Its-2 English-to-Japanese MT.

Similarly, as Korean syntax is closely related to Japanese syntax, the development of Korean and English-Korean lexicons would allow a facilitated development of an Its-2 English-to-Korean MT, on the basis of the current English-to-Japanese MT.

8.5.2 Lexical selection

In this thesis, we have described lexical selection procedures, focusing mainly on the problem of grammatical translational ambiguities. We have noticed several times that the work done here only solves a tiny part of a larger problem, and that more work would need to be achieved about lexical selection, in order to reduce the error rate in case of conceptual translational ambiguities, or just to improve the fluency level of the generated output.

A possible way of enhancing lexical selection would consist in taking the semantic context (see van der Plas (2008); van der Plas (2011); Sumida and Torisawa (2008)) into account. Then, the word with the highest semantic relatedness with the lexical context would be selected. A method such as the one proposed in (Schwab et al., 2011) would allow to compute semantic relatedness considering several neighbour words, at a reduced cost.

Another possible method would consist in more purely statistical selection, using a language model that gives a score to word n-grams[1], allowing to select the words that generate the most probable n-grams, as in (Kurohashi et al., 2005).

The combination of both methods may also be considered.

8.5.3 Politeness level selection

Japanese counts several politeness levels (see Section A.0.4) that influence the words and verb forms used in the sentence (see Sections 3.2.5, 7.2.2, A.0.1). As the impact of politeness expression is much weaker in the English sentence, politeness level remains mainly an underspecified parameter in English-to-Japanese MT.

A possible direction in future development of the Its-2 English-to-Japanese MT would consist in giving to the user a possible choice of language style and politeness level for the output. The user would just need to select a politeness level or style in a list or to keep the default style, before launching the translation process.

An advantage of LBMT, compared to pure data-driven techniques, is that a rule such as the selection of specific tense forms for a given politeness level can be applied consistently on all the generated sentences. In the case of Its-2, the user selection of politeness level style would have three main consequences on the translation process:

- In the transfer phase, lexical selection (especially for pronouns, but also for nouns and other lexemes) would give priority to the words that correspond to the specified politeness level.

- In the generation phase, tense form selection would generate the appropriate forms for verbs or adjectives.

- Also in the generation phase, affixation of the suitable sentence-final particles would be done when it is possible, at the end of sentences.

8.5.4 Hybridisation

As it has been evoked in the introduction of the thesis, LBMT and data-driven MT both present their advantages and drawbacks. Hence, the development of a hybrid version of Its-2 may benefit from the advantages of both linguistics-based and data-driven approaches.

Statistical post-editing (SPE) is a possible way of hybridisation of the system. It does not alter the linguistics-based core system, but adds a final statistical component which is aimed at translating the Japanese output into more fluent Japanese.

[1]A word n-gram is a contiguous sequence of n words

In a first experiment carried out with Asheesh Gulati, SPE with the phrase-based SMT toolkit Moses has reordered of the words of a sentence that Its-2 had incorrectly generated. We can see the incorrect Its-2 output in 8.4.b and the corrected statistically post-edited output in 8.4.c:

(8.4) (a) I can walk to school in ten minutes.

(b)	*私	は	十	に	学校	に	分		歩ける
	watashi	ha	ju	ni	gakkou	ni	pun		arukeru
	I	[topic/subject]	10	to	school	to	minutes		can walk
(c)	私	は	1 0	分		で	学校	に	歩ける。
	watashi	ha	ju	pun		de	gakkou	ni	arukeru
	I	[topic/subject]	10	minutes		in	school	to	can walk

This experiment has been achieved after the English-to-Japanese translation of 4984 sentences of the Tatoeba corpus by Its-2 and the training of the SPE component on the correction of those output sentences with the reference Japanese sentences contained in the corpus. The evaluation task has consisted in translating and correcting 997 other sentences of the Tatoeba corpus. The results have shown a slight improvement of the BLEU score, rising from 4.47% (with Its-2 only) to 4.69% (after SPE). However, manual evaluations do not find a correlation between the BLEU score increase and a global improvement of the translation quality. Instead, the rate of improved translations, such as example 8.4 seems clearly lower than the rate of deteriorated translations, as in (Toue et al., 2011).

In a second experiment described in (Kauffmann and Gulati (2013)), a much larger set of sentences from the Tatoeba corpus has been used: 48'000 sentences for SPE training and 1000 sentences for SPE tuning. Moreover, the generated Japanese language models of the SPE systems have been computed on word 5-grams instead of word 3-grams[2]. Two different systems have been compared for the SPE task: Moses (tuned for unlimited reordering within the sentences), and the HMT system Joshua.

Again, the evaluation task has consisted in the translation and SPE of 1000 sentences from from the Tatoeba corpus. As shown in Table 8.1, SPE has enabled to considerably improve the Its-2 translations, raising the BLEU score from 6.65% with Its-2 alone to 25.23% (when SPE is done with Joshua), and to 25.78% (when SPE is done with Moses). A manual evaluation on a random sample of 45 sentences from the test set has shown that SPE with Joshua gives better output than SPE with Moses (see Kauffmann and Gulati (2013)).

[2]The notion of Japanese word n-gram is, as well as the notion of Japanese word, highly dependent on the the type of word segmentation used. The segmentation used here separates verb and adjective radicals from their endings, so it goes beyond the *word* definition used in this thesis. While many ungrammatical expressions had been selected with the use of 3-grams trained on a smaller corpus, the use 5-grams trained on a much larger corpus has significantly reduced these errors.

System	BLEU
Its-2	6.65%
Its-2 + SPE (with Moses)	25.78%
Its-2 + SPE (with Joshua)	25.23%
Joshua	27.88%

Table 8.1: Comparison of BLEU scores computed on a test set of 1000 sentences

System	score average
Its-2	0.488
Its-2 + SPE (with Moses)	0.522
Its-2 + SPE (with Joshua)	0.593
Joshua	0.580

Table 8.2: Comparison of manual scores between 0 and 1, computed on a random sample of 45 sentences.

This gain in syntactic and semantic accuracy, which had not appeared in the automatic evaluation results, may be a consequence of the hierarchical phrase-based model of Joshua, that seems more powerful than the classical statistical phrase-based model of Moses.

Using the Joshua HMT system alone as a strong baseline for English-Japanese MT, we can see in Table 8.1 and 8.2 that the Its-2+SPE results are roughly equivalent to those of Joshua alone. Joshua, which has been trained in the same conditions as the SPE components, obtains the best score in the automatic evaluation (with a little advance of +2.1% on the highest SPE BLEU score) and Its-2+ Joshua SPE obtains the best score in the manual evaluation (with a slight advance of +1.3% on Joshua). A closer look at the data shows, even if the average scores are almost similar, that the outputs are often different, some sentences being translated better by Its-2+ Joshua SPE, other sentences being translated better by Joshua alone and some (less frequent) sentences being translated better by Its-2 alone.

In future work, research for even more efficient SPE could be carried out.

Two possible methods may be tried: enhancement of the SPE training phase, and assessment of the output quality in order to improve most of the Its-2 output, to keep the best translations unchanged (as in Suzuki (2011)) and to replace those where a significant part of the sentence has been dropped by a HMT translation.

As we have mentioned earlier in this section, the utilisation of language models for lexical selection during the transfer phase is another possible form of statistical correction that may be added to Its-2 in future work.

We can imagine, for a future hybrid version of Its-2, the addition of both statistical lexical selection and statistical post-editing. We can also imagine a more deeply hybrid version of the system, that would combine grammar-based decisions and statistical corrections at several steps of the translation process, following the example of (Ahsan et al., 2010) or (Chen and Eisele, 2010).

8.6 Contributions

The main contributions of this thesis have been:

- A study of the structural asymmetries between English and Japanese and of possible solutions for the generation of asymmetrical translations in non-interlingual procedural linguistics-based machine translation;
- A test of the Its-2 MT system on English-to-Japanese translation;
- An Its-2+SPE English-to-Japanese translation that gives results equivalent in quality to those of a HMT system (Section 8.5);
- An application of the data contained in the Case Frame file for the generation of Japanese verbal phrases in English-to-Japanese MT (Section 3.4 and Section 7.4);
- A method for bilingual verb subcategorisation detection, using monolingual data and basic bilingual verb correspondences (Section 3.4).

Appendix A

Japanese word classification

We describe here the word classification that we have defined and the lexical properties of the different Japanese word categories. We first present usual categories, and then focus on particles and on the copula. Finally, we give some information about the Japanese politeness system.

In the Its-2 Japanese lexicon, the different word categories are:

- Nouns and pronouns,
- Verbs (including copula),
- Adjectives,
- Adverbs,
- Determiners and quantifiers,
- Conjunctions and conjunctive particles,
- Postpositions (preposition equivalent particles),
- Particles,
- Interjections.

The *particles* category has been introduced. However, particles equivalent to English prepositions or conjunctions have been classified as *postpositions* or *conjunctions*, similarly to English and other language classifications. A more detailed description of the classification can be seen in Table 3.2 in Chapter 3.

A.0.1 Usual categories

Nouns

Nouns are invariable. There are in Japanese, like in English, common and proper nouns.

Because Japanese counting system is very specific, nouns have to be divided into several semantic categories for counting. Here is an example:

(A.1)

象	二頭	と	切手	二枚
zou	nitou	to	kitte	nimai
elephant	2[big animal unit]	and	stamp	2[flat object unit]

two elephants and two stamps

To say *two elephants* or *two stamps* in Japanese, the selection of the counter words used for *two* are different, because *elephant* and *stamp* are in different counting categories[1]. An elephant is a big animal, so *two* is said "nitou", where "ni" is the number *2* and "tou" is the counter word for big animals. *Two* is said "nimai" for a stamp, where "mai" is the one for sheets of paper and thin or flat objects.

Among all the existing Japanese counter categories, here are the main ones:

category	**counter words**	**pronunciation**
person	人	nin
thing	つ	tsu
thin and flat things	枚	mai
machines and vehicules	台	dai
small things	個	ko
shoes and socks	足	soku
thin and long things	本	hon
big animals	頭	tou
drinks	敗	hai
small animals	匹	biki

Figure A.1: *Main counter words*

In the electronic lexicon used by Its-2, nouns that correspond to a counting category have their category specified. For example, a stamp 切手 ("kitte") belongs to the "mai" counting category. Countable objects that do not have a

[1]Examples of counter words can also be found in English, such as *bottles* or *pieces* in *2 bottles of beer* or *2 pieces of furniture*.

person	plain, informal	polite	respectful
first	僕 ("boku", male) あたし ("atashi", female) 俺 ("ore", male)	私 ("watashi")	私 ("watakushi")
second	君 ("kimi") お前 ("omae")	あなた ("anata") そちら ("sochira")	あなた様 (anata-sama)
third	かれ ("kare", male) 彼女 ("kanojo", female)		

Figure A.2: *Main singular personal pronouns*

specified category can be counted with the usual counting unit for things つ ("tsu").

Pronouns

There are two types of pronouns: the first are demonstrative and interrogative ones, the second are personal ones.

Demonstrative and interrogative pronouns The demonstrative and interrogative pronouns are:

これ , それ , あれ , どれ
"kore, sore, are, dore"
this one, that one, the one over there, which one?

They do not express singular or plural number, so they can as well mean: *these ones, those ones, the ones over there, which ones.* The location adverbs that have the same structure: こちら ("kochira": here), そちら ("sochira": there), あちら ("achira": over there), どちら ("dochira": where), can also be used instead of the pronouns, to be more polite. In colloquial informal speech, "kochi, sochi, achi, dochi" are also used.

Personal pronouns In Japanese personal pronouns used for subject and object are the same. Many personal pronouns exist or have existed, some of them are often used and others very seldom.

Table A.2 shows the main singular personal pronouns. It shows that, in informal speech, a woman can use あたし ("atashi"), an a man can use 俺

("ore"), instead of using more usual first person pronoun 私 ("watashi"). "atashi" gives a specific feminine and "ore" is masculine and a little rude and strong, whereas "watashi" is completely neutral.

To obtain plural personal pronouns, the suffix -たち ("tachi") or sometimes -等 ("ra") have to be added to the singular ones. Other plural pronouns exist too, like 我々 ("wareware"), which is a formal *we*.

There is also a reflexive personal pronoun: 自分 ("jibun"). It is equivalent to the English *myself, yourself, himself...*

It is important to notice that personal pronouns are not used in Japanese as much as they are in English. With the *pro-drop* phenomenon [2], it is common to find verbs conjugated without any subject pronoun. In these cases the subject is implicit in the context. Speaking to a second person, his or her name is often used instead of the second person pronoun. For example, the sentence of the next example would be used, for a question addressed to Mrs Tanaka.

(A.2) 田中 さん は 来ません か。
tanaka san ha kimasen ka
Isn't Mrs/Mr Tanaka coming?

This sentence means literaly "Isn't Mrs Tanaka coming?". In such a case an English speaking speaker would use the pronoun *you* and say: "Are you coming?".

Verbs

Verbs can be divided into three morphological groups: "godan", "ichidan" and verbs derived from 来る ("kuru": come) and する ("suru": do) (see Kauffmann (2008b)). The copula, which is the only real auxiliary, can be seen as a special kind of verb.

There are transitive and intransitive verbs, and many verbs that can be both transitive or intransitive.

Adjectives

Adjectives are divided into three categories: -いい ("ii") adjectives, -な ("na") adjectives and -の ("no") adjectives. "ii" adjectives are the only ones that are not invariable. Their conjugation depends on tense and polarity. "na" adjectives are followed by the syllable -な ("na") when they occur in prenominal position. Otherwise, when they are used as a predicate, they are followed by the copula that is conjugated in the right tense. "no" adjectives are in fact nouns that take a particular adjectival meaning when they are used as an attribute, followed by the particle の "no" and a noun.

[2] pro-drop: Term used, in generative grammar, to qualify a language in which the use of a subject with a verb is only optional. Spanish and Italian are pro-drop languages too.

Adverbs

Adverbs are invariable. Some adverbs indicate intensity, like とても ("totemo": very). Those intensity adverbs are found before adjectives or before other adverbs. Other adverbs indicate a time, like 最近 ("saikin": lately), or a place, like こちら ("kochira": here), and can be found anywhere before the verb in the sentence. They are sometimes followed by a particle, but sometimes they are not. Time nouns, like 今年 ("kotoshi": this year), often have the same syntaxic behaviour as time adverbs do.

From every "ii" or "na" adjective an adverb can be generated. Many of these generated adverbs are manner adverbs that are more likely to occur in preverbal position, but they can also be in other positions in the sentence.

Determiners and quantifiers

Determiners are non-independent words that come in prenominal position, like in this example:

(A.3)

この	奇麗な	車
kono	kireina	kuruma
this	beautiful	car

The only determiners found in Japanese are demonstrative and interrogative articles, based on the same structure "ko/so/a/do" as demonstrative and interrogative pronouns and position adverbs:

(A.4)

この	その	あの	どの
kono	sono	ano	dono
this/these	that/those	that/those(over there)	which?

Definite or indefinite articles do not exist. In a sentence like:

(A.5)

猫	を	見た。
neko	wo	mita
cat(s)	[object particle]	(I) have seen

(I) have seen cat(s).

猫 ("neko") can mean as well "the cat", "a cat", "cats", "the cats".

As there are no possessive determiners in Japanese, the particle の ("no") is always used to express possession, in a structure like this: (noun or personal pronoun) + の. When a personal pronoun is before の, it behaves like a possessive determiner:

(A.6) 私 の　　オートバイ　です。
watashi no　ootobai　desu
My　motorbike　it is.
It's my bike.

But when there is no noun after の , it is equivalent to a possessive pronoun:

(A.7) 私 の　　です。
watashi no　desu
Mine　it is.
It's mine.

The structure noun +の is often used for the second person, instead of using a personal pronoun:

(A.8) 笹野　君　の　オートバイ　です　か。
Sasano　kun　no　ootobai　desu　ka
Sasano　student　's　motorbike　it is　[question].
Is it your bike? (question asked to Sasano by another student)

Conjunctions

In Japanese conjunctions like もし ("moshi": If), でも ("demo": But/ However), それでも ("soredemo": Nevertheless), are put in the beginning of a sentence, to link it to the previous one. However, those sentences remain two independent sentences. To link two clauses in a complex sentence, or to link two nouns, conjunctive particles are used instead of conjunctions.

A.0.2 Particles

Particles are a very important category in Japanese grammar. There are different kinds of particles, which are very similar in their form. They are all attached to the end of a word, like the Saxon genitive *'s* is in English. That is the reason why they are categorised as 辞 ("ji"): non-independent words, in Japanese lexicon categorisation. You can see main particles in the TableA.3.

However, they can be very different in their syntactic function. Particle classifications change a lot depending on grammars. Four different subcategories have been chosen, to which we add postposition particles and conjunctive particles. Subcategories such as topic particles have no equivalent in English, whereas postpositional particles are very similar to English prepositions, and conjunctive particles are similar to English conjunctions.

Particle Category	**Main Particles**
Topic	は ("ha": *topic*), も ("mo": too) だ゛け ("dake": only), しか ("shika": only)
Case	が ("ga": *nominative*), を ("wo": *accusative*)
Postposition	で ("de": at, by, in), に ("ni": to, at, in, by), へ ("he": to) から ("kara": from, since), まで ("made": to, until), の ("no": of) と ("to": with), より ("yori": more than), ば゛かり ("bakari": about)
Conjunctive:	
Noun coordination	と ("to": and), や ("ya": or) か ("ka": or), とか ("toka": or)
Clause subordination	と ("to": that), から ("kara": because), のに ("noni": since)
Clause coordination	が ("ga": but), けど ("kedo": but)
Postverbal:	
Adverbial	ばかり ("bakari": just), だ゛け ("dake": just), しか ("shika": just)
Nominalisation	の ("no": *nominalisation*)
Sentence final	か ("ka": ?, *question*), よ ("yo": yes,*affirmation*) ね ("ne": isn't-it?), ぞ゛ ("zo": *masculine strong affirmation*) な ("na": *emphasis*), わ ("wa": *feminine affirmation*)

Figure A.3: *Main particles*

Topic particles

Topic particles are usually found just after a noun, a nominalized verb or an adverb. When they are used after a noun which is subject or direct object in the clause, they replace case particles that could be put there instead:

(A.9)	太郎	**は**	本	を	買った。
	Tarou	**ha**	hon	wo	katta
	Taro	**[topic particle]**	book	[object case particle]	bought.

Taro has bought a book.

They can also replace postposition particles, or be found right after them, like in this example:

(A.10)	日本	に	**は**	中国	から	16	世紀	に	渡来した。
	Nihon	ni	**ha**	chugoku	kara	16	seiki	ni	torai shita.
	Japan	to	**[topic]**	China	from	16th	century	in	introduction did.

In Japan, it was introduced from China in the 16th century.

The most frequent topic particles are は ("ha" or "wa")[3] and も ("mo").

Case particles

Case particles are usually found just after a noun or a nominalized verb (see Farmer (1984)).

(A.11)	太郎	は	本	**を**	買った。
	Tarou	ha	hon	**wo**	katta
	Taro	[topic particle]	book	**[object case particle]**	bought.

Taro has bought a book.

We have classified two particles as case particle: を ("wo" or "o")[4] and が ("ga"). Wo is the direct object (accusative) case particle. Ga is usually the verb subject particle (nominative), even if it is sometimes used to express topic too.

Postposition particles

Postpositions particles are quite similar to English prepositions, but instead of coming before a noun, they come after it. Some frequently used postposition particles are で ("de": in/by/at...), に ("ni": to, in...), へ ("he": to), まで ("made": until), の ("no": of)... Some linguists consider that に ("ni") should be classified as a case particle, because it is often used with the indirect

[3] は is transliterated "ha" in the Hepburn romanisation, and "wa" in the Kunrei-shiki romanisation. It is pronounced "wa".

[4] を is transliterated "wo" in the Hepburn romanisation, and "o" in the Kunrei-shiki romanisation. It is pronounced "o".

object (see Tsujimura (1996)). We have chosen to classify it as a postposition particle instead, because it is also often used to express direction, in sentences like:

(A.12) 東京　**に**　行こう!
Tokyo **ni** ikou!
Let's go to Tokyo!

In those sentences, "ni" could be replaced by へ ("he"), another postposition particle that expresses direction, and does not express dative. So it seems better to consider "ni" as a postposition, that has several uses, than to restrict it to a dative case marker.

の ("no") is often analysed as a genitive or possessive case particle, even if, when it is used to link two words, it does not always represent possession. In fact it is very similar to the French preposition *de*. It is often a possession marker, like in this example:

(A.13) アレックス　**の**　本
alex **no** hon
alex 's book

But it can also be used to connect two words without showing possession:

(A.14) パリ　**の**　町
Paris **no** machi
Paris of city
The city of Paris

Conjunctive particles

Conjunctive particles have been classified in three types:

Noun coordination conjunctive particles: These particles can be used to coordonate or enumerate nouns, adjectives or nominalised verbs:

(A.15) パン　と　牛乳
pan to gyûnyû
bread and milk

They can also be used in an enumeration where, instead of having one particle between 2 nouns, there is one after the first and another one after the second:

(A.16) あれ　とか　これ　とか
are toka kore toka
that for example this for example
like this or like that

The main ones are か ("ka": or), と ("to": and), とか ("toka": and for example, and/or). These particles cannot be used to coordinate clauses.

Clause subordination conjunctive particles: These conjunctive particles, such as から ("kara": because), と ("to": that), are used to link two clauses in a complex sentence. We can see how these sentences are structured in Chapter 6, which is dedicated to complex sentence translation.

Clause coordination conjunctive particles: There are no conjunctive particle equivalent to *and* to coordonate two clauses in Japanese. However, there are equivalents to *but* like けど゛ ("kedo") or が゛ ("ga").

Postverbal particles

Adverbial particles: Adverbial particles, like ばかり ("bakari": just) give information about the action of the verb, like adverbs do. Whereas adverbs always occur before the verb, adverbial particles are placed just after the verb:

(A.17) 今来たばかり。

ima	kita	bakari
now	came	just

(I)'ve just arrived.

Nominalisation particle: When の ("no") is added after a verb, or the final verb of a clause, the verb or clause is treated as a noun phrase. It can then, for example, be the subject of another verb, if a the right particle is added after "の":

(A.18) イギリス に 行く の は 楽しい!

igirisu	ni	iku	no	ha	tanoshii!
England	to	go	[nominalisation particle]	[topic]	fun

Going to England is fun!

The same process can happen when nouns such as 事 ("koto": the thing, the fact) or ところ ("tokoro": the place) are added after a verb. Using "koto" is a little more formal than using "no".

Sentence final particles: They are located at the end of a sentence after the final verb. The most salient is the か ("ka") particle. It shows that the sentence is an interrogative one, like the question mark does in English:

(A.19) 今、 光子 は 何 を しています か。

ima,	mitsuko	ha	nani	wo	shiteimasu	ka.
now,	Mitsuko	[topic]	what	[object]	is doing	[question]

What is Mitsuko doing now?

Sentence final particles are used very often in oral speech, being an essential point of Japanese discourse organisation (see Endo (2006)). The particle ね ("ne"), for example, is equivalent to English question tags:

(A.20) 暑い　ね!
atsui　ne!
Hot,　isn't it?

A.0.3 Copula

The copula だ ("da") can be considered as a special kind of stative verb. But unlike other verbs, the copula is a non-independent word, it cannot be found as a separate word. It can have two different uses in the sentence:

Verbal use

In the verbal use, the copula always follows a noun, an adjective or an adverb. Its meaning then is close to the English *it is* or the French *c'est* in null subject clauses, or equivalent to *to be* otherwise.

Auxiliary use

When the copula is used as an auxiliary, it follows a verb or a ii-adjective. Compound tenses have a structure like: verb (with the right termination) + copula (at the right tense). In some cases, the polite present copula form です ("desu") can be optionally added after verbs or adjectives, giving a more polite style to the sentence.

A.0.4 About language style and politeness

In Japanese style and politeness can affect not only the choice of the vocabulary, but also the choice of verb tenses, pronouns (see Table A.2) and grammatical forms. Style can vary depending on factors like the age or the gender of the speaker, or the level of politeness (see Nagao (1989)).

Speaker gender

There are in Japanese, a typically "feminine language" and also, a "masculine language". Even if most Japanese words do not depend on the gender of the speaker and can be used equally by a man or a woman, the use of some special words or grammatical forms show and underline the fact the speaker is a woman or a man. A man may seem stronger and ruder speaking in a masculine language, while a woman may seem more delicate. A good example of these language style effects are personal pronouns of the first person (see Table A.2).

In French, usually, the gender of the speaker does not appear, except in some cases like this one:

(A.21) "Hier je me suis bien amusée." (Yesterday, I had fun.)

Here, the "ée" ending of the past participle, associated with the first person object pronoun *me*, shows that the speaker is a woman.

Politeness levels

There are several politeness levels in Japanese. the choice of a politeness levels implies modifications in the vocabulary and grammatical forms. There are four formal levels, we give here their typical characteristics, even their properties are much more complex:

- 丁寧語 ("teineigo": polite language); it is widely in use and shows a polite distance from a speaker to the other speaker.
- 丁重語 ("techougo": formal polite language); it is quite similar to the polite speech, replacing some words, especially some specific words, by their formal equivalents.
- 尊敬語 ("sonkeigo": respectful or honorific language); it shows respect from the speaker to the subject of a sentence, which can be or not the other speaker.
- 謙譲語 ("kenjougo": humble language); it shows respect from the humble subject of the sentence to the object of the sentence.

We should also consider these two other levels:

- casual or neutral language; It is very often used, either for colloquial speech or in a neutral written document.
- slang.

In the electronic lexicon, words such as honorific verbs or pronouns, that are used at a particular politeness level, have been given a flag showing this politeness level. Words can be defined for one or several politeness levels. A word which has no specific politeness meaning has been given no politeness flag.

Appendix A

Generated translations: examples and comments

A.1 Examples

We show here a set of 52 English sentences or phrases that Its-2 can, for most of them, translate correctly, applying the transfer rules that have been described along the chapters, and using the lexical information for subcategorised verb translation and for collocation translation. The original sentence is given first (*Src*), followed by a possible correct translation (*Ref*), and the Its-2 translation in August 2013 (*Its-2*); we also show the translations produced in August 2013 by Google Translate (*Google*) and Yakushite.net (*Yakushite*), and give some specific comments (*Comments*):

A.1.1 examples for rules 4.6

(A.1) Src: time after lunch

Ref:

昼食	後	の	時間
chuushoku	go	no	jikan
lunch	after	of	time

Its-2: 昼食の後の時間

Google: 昼食後の時間

Yakushite: 昼食後の時間

Comments: All translations correct.

(A.2) Src: there is time after lunch

Ref:	昼食	の	後	に	時間	が	ある
	chuushoku	no	ato	ni	jikan	ga	aru
	lunch	of	after	at	time	(there) is	

Its-2: 時間が昼食の後にある

Google: 昼食後の時間があります

Yakushite: 昼食後の時間があります

Comments: Google Translate and Yakushite.net translate "after lunch" as a noun qualifier instead of an adverbial phrase, which is less suitable here. The Its-2 translation is correct.

A.1.2 examples for rules 4.7

(A.3) Src: she is in front of the house.

Ref:	彼女	は	家	の	前	に	いる
	kanojo	ha	ie	no	mae	ni	iru
	she	[topic/subject]	house	of	front	at	is

Its-2: 彼女は家の前にいる

Google: 彼女は家の前にある

Yakushite: 彼女は家の前にいます

Comments: Google Translate does not translate "is" properly, not taking into account the fact that the subject is a personal pronoun. The other translations are correct.

(A.4) Src: the car is in front of the house

Ref:	車	は	家	の	前	に	ある
	kuruma	ha	ie	no	mae	ni	aru
	(the) car	[topic/subject]	house	of	front	at	is

Its-2: 車は家の前にある

Google: 車が家の前にある

Yakushite: 自動車は家の前にあります

Comments: All translations correct.

(A.5) Src: this is a car

Ref:	これ	は	車	だ
	kore	ha	kuruma	da
	this	[topic/subject]	car	is

Its-2: これは車だ

Google: これは車です

Yakushite: これは自動車です

Comments: All translations correct.

A.1.3 examples for rules 4.8

(A.6) Src: I have a lovely wife

Ref:	私	に	は	美しい	妻	が	います
	watashi	ni	ha	utsukushii	tsuma	ga	imasu
	Me	to	[topic]	beautiful	wife	[nominative]	(there) is

Its-2: 私はりっぱな妻がいる

Google: 私は美しい妻を持っている

Yakushite: 私には美しい妻がいます

Comments: Google Translate incorrectly translates "have", not taking into account the fact that the object is a person. The other translations are correct.

(A.7) Src: I have a lovely bicycle

Ref:	私	は	素敵な	自転車	を	持っている
	watashi	ha	sutekina	jidensha	wo	motteiru
	I	[topic/subject]	beautiful	bicycle	[object]	have

Its-2: 私はりっぱな自転車がある

Google: 私は素敵な自転車を持っている

Yakushite: 私はすばらしい自転車を持っています

Comments: All translations correct.

A.1.4 example for rules 5.2

(A.8) Src: a quiet park and a beautiful tree and a rich tourist

Ref:	静かな	公園	と	美しい	高木	と	金持ちの
	shizukana	kouen	to	utsukushii	kouboku	to	kane mochi no
	quiet	park	and	beautiful	tree	and	rich
	観光客						
	kankoukyaku						
	tourist						

Its-2: 静かな公園と美しい高木と金持ちの観光客

Google: 静かな公園と美しい木と豊かな観光

Yakushite: 静かな公園と美しい木と金持ちの観光客

Comments: Google Translate incorrectly translates "tourist". The other translations are correct.

A.1.5 examples for rules 5.4

(A.9) Src: it looks cold

Ref: 寒そう　です　ね
samusou　desu　ne
looking cold　(it) is　isn't it?

Its-2: 寒そうだ

Google: 寒い見える

Yakushite: 寒いように見えます

Comments: Google Translate generates a grammatically incorrect translation, giving an inappropriate ending to the adjective. The other translations are correct.

(A.10) Src: it seems cold

Ref: 寒そう　です　ね
samusou　desu　ne
looking cold　(it) is　isn't it?

Its-2: 寒そうだ

Google: 寒いようです

Yakushite: 寒いように見えます

Comments: All translations correct.

(A.11) Src: it became cold

Ref: 寒くなった
samukunatta
cold became

Its-2: 寒くなった

Google: 寒い なった

Yakushite: 寒くなりました

Comments: Google Translate generates a grammatically incorrect translation, giving an inappropriate ending to the adjective. The other translations are correct.

A.1.6 examples for rules 5.5

(A.12) Src: it hurts

Ref: 痛い
itai
painful

Its-2: 痛い

Google: それは痛い

Yakushite: それは痛みます

Comments: All translations correct

(A.13) Src: I like flowers

Ref:	私	は	花	が	好き
	watashi	ha	hana	ga	suki
	I	[topic/subject]	flowers	[nominative]	liked

Its-2: 私は花が好きだ

Google: 私は花が好き

Yakushite: 私は花が好きです

Comments: All translations correct

A.1.7 examples for rules described in Section 6.2.3

(A.14) Src: I will buy beers if he comes.

Ref:	彼	が	来れば、	私	は	ビール	を
	kare	ga	koreba	watashi	ha	biru	wo
	He	[nominative]	(would)come	I	[topic/subject]	beer	[object]
	買う。						
	kau						
	buy						

Its-2: 私は彼が来るならビールを買う。

Google: 彼が来れば、私はビールを購入する。

Yakushite: 彼が来れば、私はビールを買うでしょう。

Comments: All translations correct

(A.15) Src: The frogs will come out if it is quiet.

Ref:	静か	ならば、	カエル	は	出て来る。
	shizuka	naraba	kaeru	ha	detekuru
	quiet	if it is	frogs	[topic/subject]	come out

Its-2: かえるは 静かなら出る。

Google: それが静かであればカエルが出てきます。

Yakushite: それが静かであればカエルが出てきます。

Comments: Google Translate and Yakushite.net translate the expletive pronoun "it" into "それが" ("sore ga": this) which seems to be inadequate here, where a null subject seems more appropriate. The translation generated by Its-2 is correct, even if the translation of "come out" into "出る" ("deru") may seem less fluent than 出て来る("detekuru") or 出てきます("detekimasu").

(A.16) Src: He can sing while he's driving.

Ref:	彼	は	運転しながら	歌える.
	kare	ha	unten shinagara	utaeru
	he	[topic/subject]	while driving	can sing

Its-2: 彼は彼が運転しながら歌える。

Google: 彼が運転している間、彼は歌うことができます。

Yakushite: 彼が推進する一方、彼は歌うことができます。

Comments: All the translations are grammatically correct, but, in the three of them, the repetition of the pronoun 彼 ("kare": he) may seem unnatural. A null subject in one of the two verbal clauses would be better.

(A.17) Src: The frogs will come out when he mows the lawn.

Ref:	彼	が	芝生	を	刈る	時、	カエル	は
	kare	ga	shibafu	wo	karu	toki	kaeru	ha
	he	[subject]	lawn	[object]	maws	when	frogs	[topic/subject]
	出て来る	でしょう。						
	detekuru	deshou						
	come out	will						

Its-2: かえるは彼が芝生を刈る時出る。

Google: 彼は芝生をMOWS時カエルが出てきます。

Yakushite: 彼が芝生を刈る時、カエルは出て来るでしょう。

Comments: The verb "mows" has not been translated by Google Translate. The other translations are correct.

(A.18) Src: Provided that you sing, I will go to the concert.

Ref:	ただし	あなた	が	歌えば	私	が	音楽会	に
	tadashi	anata	ga	utaeba	watashi	ga	ongakukai	ni
	if	you	[subject]	would sing	I	[subject]	concert	to
	行く。							
	iku							
	go							

Its-2: ただしあなたが歌えば私が音楽会に行く。

Google: あなたが歌うことを条件に、私はコンサートに行く 予定。

Yakushite: あなたが歌うならば、私はコンサートへ行くでしょう。

Comments: All translations correct.

(A.19) Src: He came, so she left.

Ref:	彼	は	来た。	だから	彼女	が	去った。
	kare	ha	kita	dakara	kanojo	ga	satta.
	He	[subject]	came.	So,	she	[subject]	left

Its-2: 彼は 来た。だから、彼女が 去った。

Google: 彼が来たので、彼女は去った。

Yakushite: 彼は来た、だから彼女が去りました。

Comments: Except the punctuation of the Yakushite.net translation, which should be similar to the one of the Its-2 translation, all the translations are correct.

(A.20) Src: He came but she left.

Ref:	彼	が	来た	けど	彼女	は	去った。
	kare	ga	kita	kedo	kanojo	ha	satta.
	He	[subject]	came	but	she	[topic/subject]	left

Its-2: 彼が来たけど彼女が去った。

Google: 彼が来ましたが、彼女は去った。

Yakushite: 彼は来たが彼女は去りました。

Comments: All translations correct.

(A.21) Src: He came and she left.

Ref:	彼	が	来て、	彼女	は	去った。
	kare	ga	kite	kanojo	ha	satta.
	He	[subject]	coming	she	[topic/subject]	left

Its-2: 彼が 来て 、 彼女が 去った.

Google: 彼が来て、彼女は去った。

Yakushite: 彼は来たそして彼女は去りました。

Comments: All translations correct.

(A.22) Src: He came then she left.

Ref:	彼	は	来て、	そして	彼女	は	去った。
	kare	ha	kite,	soshite	kanojo	ha	satta.
	He	[topic/subject]	coming,	then	she	[topic/subject]	left

Its-2: 彼は来て 、そして彼女は去った 。

Google: 彼は彼女が去った後来た。

Yakushite: 彼は来た、その結果彼女は去りました。

Comments: Google Translates produces an output which does not correspond to the source sentence meaning. The translation produced by Yakushite.net contains an error in verb tense selection. The Its-2 translation is correct.

(A.23) Src: He came, she left.

Ref:	彼	は	来て、	彼女	は	去った。
	kare	ha	kite,	kanojo	ha	satta.
	He	[topic/subject]	coming,	she	[topic/subject]	left

Its-2: 彼は来て、彼女は去った。

Google: 彼が来て、彼女が残した。

Yakushite: 彼は来た、彼女は去りました。

Comments: The translation produced by Yakushite.net contains an error in verb tense selection. The other translations are correct.

(A.24) Src: This is the pen that I bought.

Ref:	これ	は	私	が	買った	ペン	です
	kore	ha	watashi	ga	katta	pen	desu
	This	[topic/subject]	I	[subject]	bought	pen	is

Its-2: これは買ったペンだ。

Google: これは私が買ったペンです。

Yakushite: これは私が買ったペンです

Comments: The translation produced by Yakushite.net contains an error in verb tense selection. The other translations are correct.

A.1.8 examples for rules described in Section 7.2.3: Modals

(A.25) Src: He can swim.

Ref:	彼	は	泳げる
	kare	ha	oyogeru
	he	[topic/subject]	can swim

Its-2: 彼は泳げる。

Google: 彼は泳ぐことができます。

Yakushite: 彼は泳ぐことができます。

Comments: All translations correct.

(A.26) Src: He may come.

Ref:	彼	は	来る	かもしれない。
	kare	ha	kuru	kamoshirenai
	he	[subject/topic]	come	maybe

Its-2: 彼は来るかもしれない。

Google: 彼が来るかもしれない。

Yakushite: 彼は来るかもしれません。

Comments: All translations correct.

(A.27) Src: He might not come.

Ref:	彼	は	来ない	かもしれない。
	kare	ha	konai	kamoshirenai
	he	[topic/subject]	not come	maybe

Its-2: 彼は来るかもしれない。

Google: 彼は来ないかもしれません。

Yakushite: 彼は来ないかもしれません。

Comments: All translations correct.

(A.28) Src: He must come.

Ref:	彼	は		来なければ なりません。
	kare	ha		konakereba narimasen
	he	[topic/subject]		must come

Its-2: 彼は来なければならない。

Google: 彼が来なければなりません。

Yakushite: 彼は来るかもしれません。

Comments: All translations correct.

(A.29) Src: He must not come.

Ref:	彼	は	来て	は	いけません。
	kare	ha	kite	ha	ikemasen
	he	[topic]	coming	[topic/subject]	should not

Its-2: 彼は来られない。

Google: 彼が来ていなければなりません。

Yakushite: 彼は来てはいけません。

Comments: The Google Translate translation is wrong. The other translations are correct.

(A.30) Src: He should come.

Ref:	彼	は	来る	べきだ。
	kare	ha	kuru	beki da
	he	[topic/subject]	come	should

Its-2: 彼は来るべきだ。

Google: 彼は来るべき。

Yakushite: 彼は来るべきです。

Comments: All translations correct.

(A.31) Src: I didn't know if he could come.

Ref:	私	は	彼	が	来られる	か	知らなかった。
	watashi	ha	kare	ga	korareru	ka	shiranakatta
	I	[topic/subject]	he	[subject]	can come	if	didn't know

Its-2: 私は彼が来られるか知らなかった。

Google: 彼が来ることがあれば、私は知りませんでした。

Yakushite: 私は彼は来ることができたかを知りませんでした

Comments: All translations correct. (In many other uses of "could", Its-2 is so far not able to generate a correct translation.)

A.1.9 examples for rules described in Section 7.2.3: Semi-modals

(A.32) Src: I want to eat.

Ref: 私 は 食べたいです。
watashi ha tabetai desu
I [topic/subject] want to eat

Its-2: 食べたい。

Google: 私が食べたい。

Yakushite: 私は食べたいです。

Comments: All translations correct.

(A.33) Src: I wanted to eat.

Ref: 私 は 食べたかったです。
watashi ha tabetakatta desu
I [topic/subject] wanted to eat

Its-2: 食べたかった。

Google: 私が食べたかった。

Yakushite: 私は食べたかったです。

Comments: All translations correct.

(A.34) Src: I had to eat.

Ref: 食べなければ ならなかった。
tabenakereba naranakatta
(I) had to eat

Its-2: 食べなければ ならなかった。

Google: 私が食べていた。

Yakushite: 私は食べなければなりませんでした。

Comments: Google Translate generates an inappropriate translation, which means "I was eating." The other translations are correct.

A.1.10 examples for rules described in Section 7.3.3

(A.35) Src: The book was read by Taro.

Ref: 本 が 太郎 に 読まれた。
hon ga Taro ni yomareta
book [subject] Taro by was read

Its-2: 本が太郎に読まれた。

Google: 本は太郎で読み取った。

Yakushite: 本は太郎によって読まれました。

Comments: Google Translate generates an wrong translation. The other translations are correct.

(A.36) Src: She made him drive the car.

Ref:	彼女	は	彼	に	自動車	を	運転させました。
	kanojo	ha	kare	ni	jidousha	wo	unten sasemashita
	she	[topic/subject]	him	to	car	[object]	made drive

Its-2: 彼女は彼に車を運転させた。

Google: 彼女は彼が車を運転しました。

Yakushite: 彼女は彼に自動車を運転させました。

Comments: Google Translate generates a translation which a bit different from the source sentence, literally meaning "As for her, he drove the car." The other translations are correct.

A.1.11 examples for rules described in Section 7.4.3

(A.37) Src: He lost her phone while driving.

Ref:	彼	は	運転しながら	彼女の	電話	を
	kare	ha	unten shinagara	kanojo no	denwa	wo
	he	[topic/subject]	driving	her	phone	[object]
	失った。					
	ushinatta					
	[lost]					

Its-2: 彼は運転しがら彼女の受話器を失った。

Google: 運転中に彼は彼女の携帯電話を失った。

Yakushite: 彼は運転している間彼女の電話を失いました。

Comments: Its-2 should translate "phone" into 電話 ("denwa") instead of 受話器 ("juwaki"). The two other translations are completely correct.

(A.38) Src: The show ending, he came.

Ref:	ショー	が	終わって、	彼	は	来ました。
	show	ga	owatte	kare	ha	kimashita
	show	[subject]	ending	he	[topic/subject]	came

Its-2: 展覧会 終わって、彼が 来た。

Google: ショーのエンディングは、彼が来た。

Yakushite: ショーが終わって、彼は来ました。

Comments: Google Translates translates "ending" as a noun, which seems unlikely here. A が("ga") particle is missing in the Its-2 translation. The Yakushite.net translation is completely correct.

A.1.12 examples for rules described in Section 7.4.4

(A.39) Src: I remember calling you.

Ref: 私 は あなた を 呼んだ こと を
watashi ha anata wo yonda koto wo
I [topic/subject] you [object] called the fact [object]
覚えています。
oboeteimasu
remember

Its-2: 私はあなたを招くのを思い出す。

Google: 私はあなたを呼び出したのを覚えています。

Yakushite: 私はあなたを呼んだことを覚えています。

Comments: Its-2 translates "calling" into a nominalised present form, but a nominalised simple past form would be better. The other translations are correct.

(A.40) Src: I think that this choice is not good.

Ref: 私 は この 好み が よくない と 思う。
watashi ha kono konomi ga yokunai to omou
I [topic/subject] this choice [subject] not good that think

Its-2: 私はこの好みがよくないと思う。

Google: 私はこの選択が良いではないと思われる。

Yakushite: 私はこの選択が良くないと考えます。

Comments: All translations correct.

(A.41) Src: I will try to eat this dinner.

Ref: この 夕食 を 食べて みる。
kono yûshoku wo tabete miru
this dinner [object] eating try to

Its-2: この夕食を食べてみる。

Google: 私はこの夕食を食べてしようとします。

Yakushite: 私はこのディナーを食べようとするでしょう

Comments: The translation generated by Google Translate is ungrammatical. The other translations are correct.

A.1.13 examples of collocation translation

(A.42) Src: I had a dream.

Ref: 私 は 夢 を 見た。
watashi ha yume wo mita
I [topic/subject] dream [object] saw

Its-2: 私は夢を見た。

Google: 私は夢を見た。

Yakushite: 私は夢を見ました。

Comments: All translations correct.

(A.43) Src: This is the dream I had.

Ref:	これ	は	私	が	見た	夢	だ。
	kore	ha	watashi	ga	mita	yume	da
	this	[topic/subject]	I	[subject]	saw	dream	is

Its-2: これは私が見た夢だ。

Google: これは私が持っていた夢です。

Yakushite: これは私が持っていた夢です。

Comments: The collocation "have a dream", that becomes here "dream I had", is not translated well by Google Translate and Yakushite.net. The Its-2 translation is correct.

(A.44) Src: I had a dream yesterday and I'll have it tomorrow again.

Ref:	私	が	夢	を	昨日	見て、	あす
	watashi	ga	yume	wo	kinou	mite,	asu
	I	[subject]	dream	[object]	yesterday	saw	tomorrow
	もう一度	見る。					
	mou ichido	miru					
	one more time	see					

Its-2: 私が夢昨日 見て 、 私があすもう一度見る。

Google: 私は昨日、夢を持っていたと私は明日再びそれを持っているよ。

Yakushite: 私は昨日夢を見たそして私は再び明日それを持っているでしょう。

Comments: The three MT systems repeat the subject translation 私 ("watashi") twice in the sentence, which may seem unnatural in Japanese. The collocation "have a dream" is not translated well here by Google Translate. When "dream" is referred to by the pronoun "it" in "have it", both Google Translate and Yakushite.net translate "have" incorrectly, and only Its-2 translates it (according to the collocation translation) into 見る("miru": see).

(A.45) Src: My father wears a hat.

Ref:	私の	お父さん	は	帽子	を	かぶる。
	watashi no	otousan	ha	boshi	wo	kaburu
	my	father	[topic/subject]	hat	[object]	wear

Its-2: 私のお父さんは帽子をかぶる。

Google: 私の父は帽子を身に着けている。

Yakushite: 私の父は帽子を被っています。

Comments: Its-2 and Yakushite.net uses the specific verb かぶる("kaburu") for the translation of "wear" in the collocation "wear a hat". Google Translate uses a generic expression. All translations seem to be correct.

(A.46) Src: My father wears a T-shirt.

Ref:	私の	父	は、	Tシャツ	を	身に着けている。
	watashi no	chichi	ha	T-shirt	wo	mi ni tusketeiru
	my	father	[topic/subject]	hat	[object]	is wearing

Its-2: 私のお父さんはTシャツを着る。

Google: 私の父は、Tシャツを身に着けている。

Yakushite: 私の父はティーシャツを身に付けています。

Comments: All translations correct.

(A.47) Src: She ate an apple but went back home.

Ref:	彼女	が	りんご	を	食べた	けど	帰った。
	kanojo	ga	ringo	wo	tabeta	kedo	kaetta
	she	[subject]	apple	[object]	ate	but	went back home

Its-2: 彼女がりんごを食べたけど帰った。

Google: 彼女はリンゴを食べたが、戻って家に帰った。

Yakushite: 彼女はりんごを食べたが家に戻りました。

Comments: All translations correct (for the English collocation "go back home", and for the whole sentence).

A.1.14 examples of subcategorised verb translation

(A.48) Src: I understand the book.

Ref:	私	は	本	が	わかる。
	watashi	ha	hon	ga	wakaru
	I	[subject]	book	[nominative]	understand

Its-2: 私は本がわかる。

Google: 私は本を理解しています。

Yakushite: 私は本を理解します。

Comments: All translations correct

(A.49) Src: He wrote her a letter.

Ref:	彼	は	彼女	に	手紙	を	書いた。
	kare	ha	kanojo	ni	tegami	wo	kaita
	he	[subject]	her	to	letter	[object]	wrote

Its-2: 彼は彼女に手紙を書いた。

Google: 彼は彼女に手紙を書いた。

Yakushite: 彼は彼女に手紙を書きました。

Comments: All translations correct

(A.50) Src: She will inform him of the situation.

Ref:	彼女	は	彼	に	状況	を	知らせる	でしょう。
	kanojo	ha	kare	ni	joukyou	wo	shiraseru	deshou
	she	[subject]	him	to	information	letter	[object]	wrote

Its-2: 彼女は彼に立場を告げる 。

Google: 彼女は状況に彼をお知らせいたします。

Yakushite: 彼女は彼に状況を知らせるでしょう。

Comments: He, the argument structure in the translation generated by Google Translate is not correct. The translation generated by Its-2 has a correct argument structure, but the translation of "situation" into 立場 ("tachiba") is less suitable than 状況 (joukyou). The translation generated by Yakushite.net is correct.

(A.51) Src: He looks at the map.

Ref:	彼	は	地図	を	見ます。
	kare	ha	chizu	wo	mimasu
	he	[subject]	map	[object]	sees

Its-2: 彼は地図を見る。

Google: 彼は、地図を見ます。

Yakushite: .彼は地図を見ます。

Comments: All translations correct

(A.52) Src: He looks for the map.

Ref:	彼	は	地図	を	捜します。
	kare	ha	chizu	wo	sagashimasu
	he	[subject]	map	[object]	looks for

Its-2: 彼は地図を探す。

Google: 彼は、マップを探します。

Yakushite: .彼は地図を捜します。

Comments: All translations correct

A.2 Global comments

We have seen here in these 52 examples that a large part of the transfer rules described in the thesis have been implemented with success. A comparison with Google Translate and Yakushite.net has shown that Google Translate, which does not use transfer rules, is not programmed for a systematic respect of grammatical constraints, and seems especially weak for the selection of the verb depending on the subject (or object) type (see A.1.2, A.1.3).

Still, Its-2 for English-Japanese remains so far a prototype, and has weaknesses in rules coverage, lexical coverage, and lexical selection (as in examples A.37 and A.50). Its level of implementation and the potential quality of its translations on a random selection of sentences is much lower than those of Google Translate, and maybe even much lower than those of Yakushite.net.

However, with the help of the deep structure parser Fips for English, Its-2 seems to be the one of the three systems that has the best architecture for the selection of the preposition translation depending on the grammatical function of the preposition phrase (see A.1.1) and for the detection and translation of verb-object collocations when the verb is in a relative clause following the noun (as in example A.43) or when the noun is referred to by a pronoun in an anaphora (as in example A.44, see Wehrli (2013)).

Bibliography

Abney, S. (1987). *The English Noun Phrase in its Sentential Aspect*. MIT Press.

Ahsan, A., Kolachina, P., Kolachina, S., Sharma, D. M., and Sangal, R. (2010). Coupling Statistical Machine Translation with Rule-based Transfer and Generation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas*.

ALC (2012). Space ALC. www.alc.co.jp.

Alegria, I., de Ilarraza, A., Labaka, G., Lersundi, M., Mayor, A., and Sarasola, K. (2007). Transfer-based MT from Spanish into Basque: reusability, standardization and open source. *Lecture Notes in Computer Science*, 4394/2007:374 -- 384.

Amano, S., Hirakawa, H., and Tsutsumi, Y. (1989). AS-TRANSAC: The Toshiba Machine Translation System. In *Proceedings of MT Summit II*, Munich, Germany.

Ashizaki, T. (1989). Outline of the JICST machine translation system. In *Proceedings of MT Summit II*, pages 44--49, Munich, Germany.

Asia Online (2012). Language Studio. www.languagestudio.com.

Baker, M. (2003). Verbal Adjectives as Adjectives without Phi-Features. In *Proceedings of the Fourth Tokyo Conference on Psycholinguistics*, pages 1--22.

Baldwin, T. and Bond, F. (2002). Multiword Expressions: Some Problems for Japanese NLP. In *Proceedings of the eighth annual meeting of the association of Natural Language Processing (Japan) (NLP 2002)*, pages 379--382, Keihanna, Japan.

Barnett, J., Mani, I., Rich, E., Aone, C., Knight, K., and Martinez, J. C. (1991). Capturing Language-Specific Semantic Distinctions in Interlingua-Based MT. In *Proceedings of MT Summit III*, Washington, DC, USA.

Barreiro, A., Scott, B., Kasper, W., and Kiefer, B. (2011). Openlogos machine translation: philosophy, model, resources and customization. *Machine Translation*, 25(2):107--126.

Blanchon, H., Boitet, C., and Choumane, A. (2006). Traduction automatisée fondée sur le dialogue et documents auto-explicatifs : bilan du projet LIDIA. *TAL, volume 47*, pages 175--204.

Boitet, C. (2005). Gradable quality translations through mutualization of human translation and revision, and UNL-based MT and coedition. In *Universal Networking Language, advances in theory and applications (presented at the 2nd Workshop on UNL and Other Interlinguas)*, pages 393--410, Mexico.

Boitet, C., Bey, Y., Tomokiyo, M., Cao, W., and Blanchon, Hervé. (2006). IWSLT-06: experiments with commercial MT systems and lessons from subjective evaluations. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2006)*, Kyoto, Japan.

Boitet, C., Blanchon, H., Seligman, M., and Bellynck, V. (2009). Evolution of MT with the Web. In *Proceedings of International Conference "Machine Translation 25 Years On", Cranfield, England.*

Boitet, C., Boguslavskij, I., and Cardenosa, J. (2007). An Evaluation of UNL Usability for High Quality Multilingualization and Projections for a Future UNLplusplus Language. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 4394 of *Lecture Notes in Computer Science*, pages 361--373. Springer Berlin Heidelberg.

Bond, F. (2005). *Translating the Untranslatable, A Solution to the Problem of Generating English Determiners.* CSLI Publications.

Breen, J. (2009). JMdict/EDICT. http://csse.monash.edu.au/~jwb/wwwjdicinf.html.

Bresnan, J., editor (1982). *The Mental Representation of Grammatical Relations.* The MIT Press, Cambridge, MA.

Bresnan, J. (2001). *Lexical-Functional Syntax.* Blackwell Publishers, Oxford.

Brockett, C., Aikawa, T., Aue, A., Menezes, A., Quirk, C., and Suzuki, H. (2002). English-Japanese Example-Based Machine Translation Using Abstract Linguistic Representations. In *Proceedings of COLING 2002*, Taipeh, Taiwan.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings EACL 2006*, pages 249--256.

Carrera, J., Beregovaya, O., and Yanishevsky, A. (2009). Machine Translation for Cross-Language Social Media. www.promt.com/company/technology/overview.

Chen, Y. and Eisele, A. (2010). Integrating a Rule-based with a Hierarchical Translation System. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010)*, Malta. European Language Resources Association.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics, 33(2)*, page 201–208.

Chiang, D. (2010). Learning to translate with source and target syntax. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1443--1452, Uppsala, Sweden.

Chiang, D., Lopez, A., Madnani, N., Monz, C., Resnik, P., and Subotin, M. (2005). The Hiero machine translation system: extensions, evaluation, and analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT 2005)*, pages 779--786, Vancouver, British Columbia, Canada.

Chomsky, N. (1972). *Topics in the Theory of Generative Grammar*. Mouton.

Chomsky, N. (1995). *The Minimalist Program*. MIT Press.

Collins, M., Koehn, P., and Kučerová, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 2005)*, pages 531--540, Ann Arbor, Michigan, USA.

Costa-Jussa, M. R., Farrus, M., Marino, J. B., and Fonollosa, J. A. (2010). Automatic and human evaluation study of a rule-based and a statistical Catalan-Spanish machine translation systems. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC2010)*, pages 1707--1711, Valletta.

Cowan, B., Kučerová, I., and Collins, M. (2006). A discriminative model for tree-to-tree translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 232--241, Sydney, Australia.

Cross language (2012). Cross language software. www.crosslanguage.co.jp/index2.html.

Culicover, P. and Jackenfoff, R. (2005). *Simpler Syntax*. Oxford University Press.

Dorr, B. J. (1994). Machine Translation Divergences: A Formal Description and Proposed Solutions. *Computational Linguistics*, 20:597--633.

Dugast, L., Senellart, J., and Koehn, P. (2007). Statistical post-editing on SYSTRAN's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation (StatMT 2007)*, pages 220--223, Prague, Czech Republic.

Dyer, C., Weese, J., Setiawan, H., Lopez, A., Ture, F., Eidelman, V., Ganitkevitch, J., Blunsom, P., and Resnik, P. (2010). cdec: a decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*, pages 7--12, Uppsala, Sweden.

Ehara, T. (2007). Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation. In *Proceedings of the MT Summit XI, Workshop on Patent Translation*, Copenhagen, Denmark.

Eijiro (2012). Electronic Dictionary Project. www.eijiro.jp.

Eisele, A., Federmann, C., Uszkoreit, H., Saint-Amand, H., Kay1, M., Jellinghaus, M., Hunsicker, S., Herrmann, T., and Chen, Y. (2008). Hybrid Machine Translation Architectures within and beyond the EuroMatrix project. In *Proceedings of the 12th EAMT conference*, Hamburg, Germany.

Endo, Y. (2006). *A Study of the Cartography of Japanese Syntactic Structures*. PhD thesis, Université de Genève.

Farmer, A. (1984). *Modularity in Syntax*. MIT Press.

Font Llitjós, A. and Vogel, S. (2007). A walk on the other side: adding statistical components to a transfer-based translation system. In *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 72--79, Rochester, New York.

Forcada, M. L., Tyers, F. M., and Ramirez-Sanchez, G. (2009). The Apertium machine translation platform: Five years on. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 3--10, Alicante, Spain.

Frederking, R. E. and Brown, R. D. (1996). The Pangloss-Lite Machine Translation System. In *Proceedings of Second Conference of the Association for Machine Translation in the Americas*, Montreal, Canada.

Fujitsu (2012). Atlas V14. http://software.fujitsu.com/jp/atlas.

Gesmundo, A. and Anderson, J. (2011). Heuristic Search for Non-Bottom-Up Tree Structure Prediction. In *Proceedings of EMNLP 2011*.

Goh, C.-L. and Sumita, E. (2011). Splitting Long Input Sentences for Phrase-based Statistical Machine Translation. In *Proceedings of the 17th annual meeting of the association of Natural Language Processing (Japan) (NLP 2011)*, pages 802--805, Toyohashi, Japan.

Google (2012a). Google Translate. http://translate.google.com.

Google (2012b). Inside Google Translate. http://translate.google.com/about/.

Gulati, A. (2011). Hybrid machine translation: an overview. Master's thesis, Université de Genève.

Halpern, J. (2008a). Japanese part of speech codes. http://cjk.org/cjk/samples/jappos.htm.

Halpern, J. (2008b). Orthographic Variation in Japanese. http://cjk.org/cjk/reference/japvar.htm.

Halpern, J. (2008c). Principal Japanese Lexical Ressources. http://cjk.org/cjk/samples/japsam.htm.

Hashimoto, S. (1934). 国語法要説 *(An outline of the Japanese grammar)*. Meiji Shoin, Tokyo.

Hildebrand, A. S. and Vogel, S. (2010). CMU system combination via hypothesis selection for WMT'10. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics*, pages 307--310, Uppsala, Sweden.

Hinds, J. (1986). *Japanese*. Croom Helm.

Hino, S., Murakami, J., Tokuhisa, M., and Murata, M. (2011). 統計翻訳における英辞郎を利用したパラレルコーパスの効果 (Effects of a parallel corpus using English expressions for statistical machine translation). In *Proceedings of the 17th annual meeting of the association of Natural Language Processing (Japan) (NLP 2011)*, pages 400--403, Toyohashi, Japan.

Hutchins, J. (1986). *Machine Translation: Past, Present, Future*. Ellis Horwood limited.

Hutchins, J. and Somers, H. (1992). *An introduction to machine translation*. Academic Press.

IBM (2012). Honyaku no Ôsama. www.ibm.com/software/jp/internet/king.

Ikehara, S., Shirai, S., Yokoo, A., and Nakaiwa, H. (1991). Toward an MT System without Pre-Editing --- Effects of New Methods in ALT-J/E ---. In *Proceedings of MT Summit III*, Washington DC.

Ikeya, A. (1991). A contextuel approach to Japanese Adjectives. In *Proceedings of the sixth Japanese-Korean joint conference on formal linguistics*, pages 64--90.

Imamura, K., Sumita, E., and Matsumoto, Y. (2003). Feedback cleaning of machine translation rules using automatic evaluation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)- Volume 1*, pages 447--454, Sapporo, Japan.

Jin, Y. (2010). A hybrid-strategy method combining semantic analysis with rule-based MT for patent machine translation. In *The 6th IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'10)*, Beijing.

Kaji, H. (1987). HICATS/JE : A Japanese-to-English Machine Translation System Based on Semantics. In *Proceedings of MT Summit*, Hakone, Japan.

Kauffmann, A. (2008a). Description de la structure de la phrase japonaise pour l'analyse syntaxique. In *Proceedings of TALN-RECITAL 2008*, Avignon, France.

Kauffmann, A. (2008b). *Japanese morphology, lexicon and syntactic parsing, in the context of Japanese-French automatic translation*. DEA thesis, University of Geneva.

Kauffmann, A. and Gulati, A. (2013). Comparative Evaluation of Statistical Post-Editing in English-Japanese MT. In *Proceedings of the 19th annual meeting of the association of Natural Language Processing (Japan) (NLP 2013)*, Nagoya, Japan.

Kauffmann, A., Kawahara, D., and Kurohashi, S. (2011). Treatment of Complex Sentences, Modality and Verbal Structures in Linguistics-Based MT. In *Proceedings of the 17th annual meeting of the association of Natural Language Processing (Japan) (NLP 2011)*, Toyohashi, Japan.

Kawahara, D. and Kurohashi, S. (2006). Case Frame Compilation from the Web using High-Performance Computing. In *Proceedings of The 5th International Conference on Language Resources and Evaluation (LREC-06)*.

Kawahara, D. and Kurohashi, S. (2010). Acquiring Reliable Predicate-argument Structures from Raw Corpora for Case Frame Compilation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1389--1393, Valletta, Malta.

Kinoshita, S., Phillips, J., and Tsujii, J.-I. (1992). Interaction between Structural Changes in Machine Translation. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 679--685, Nantes, France.

Kodensha (2012). Kodensha MT systems. www.kodensha.jp/company/history.html.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague.

Komachi, M., Nagata, M., and Matsumoto, Y. (2006). Phrase Reordering for Statistical Machine Translation Based on Predicate-Argument Structure. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2006)*, Kyoto, Japan.

Kuno, S. (1973). *The structure of the Japanese language*. MIT Press.

Kuroda, K. (2007). Evénement causatif et ses structures prédicatives. In *Proceedings of Forces of Grammatical Structures (FIGS 07)*, Paris, France.

Kuroda, S.-Y. (1979a). *Aux quatre coins de la linguistique*. Editions du Seuil.

Kuroda, S.-Y. (1979b). *Generative Grammatical Studies in the Japanese language*. Garland Publishing.

Kurohashi, S. and Nagao, M. (2003). Building a japanese parsed corpus. In Abeillé, A., editor, *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 249--260. Springer Netherlands.

Kurohashi, S., Nakazawa, T., Kauffmann, A., and Kawahara, D. (2005). Example-Based Machine Translation Pursuing Fully Structural NLP. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2005)*, Pittsburg, USA.

Kuwae, K. (1984). *Manuel de japonais*. l'Asiathèque.

Language Weaver (2012). SDL-LW. www.sdl.com/en/language-technology/products/automated-translation.

Lee, Y.-S., Zhao, B., and Luo, X. (2010). Constituent Reordering and Syntax Models for English-to-Japanese Statistical Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 626--634, Beijing.

Leech, G. (2003). Modality on the move: the English modal auxiliaries 1961-92. In *Modality in contemporary English*, volume 44 of *Topics in English linguistics*, pages 223--240. Mouton de Gruyter, Berlin.

Lepage, Y. and Denoual, E. (2005a). ALEPH: an EBMT system based on the preservation of proportional analogies between sentences across languages. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2005)*, Pittsburg, USA.

Lepage, Y. and Denoual, E. (2005b). Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, 19(3-4): 251--282.

Lepage, Y. and Lardilleux, A. (2007). The GREYC Machine Translation System for the IWSLT 2007 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2007)*, pages 49--54, Trento, Italy.

Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Khudanpur, S., Schwartz, L., Thornton, W. N. G., Weese, J., and Zaidan, O. F. (2009). Joshua: an open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (StatMT 2009)*, pages 135--139, Athens, Greece.

Lin, S.-C., Wang, J.-C., and Wang, J.-F. (2005). Translation Divergence Analysis and Processing for Mandarin-English Parallel Text Exploitation. In *Proceedings of 17th Conference on Computational Linguistics and speech Processing ((ROCLING 2005))*, Tainan, Taiwan.

Linguatec (2013). Personal Translator Demo. www.linguatec.net/onlineservices/pt.

LionBridge (2012). iTranslator MT system. http://itranslator.lionbridge.com.

Liu, Y., Lü, Y., and Liu, Q. (2009). Improving tree-to-tree translation with packed forests. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL (ACL 2009) and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, pages 558--566, Suntec, Singapore.

Lucy LT (2012). Kwik lucy. www.lucysoftware.com/machine-translation/lucy-lt-kwik-translator-/lucy-lt-quick-translator.html.

Mahesh, R., K.Sinha, and Thakur, A. (2005). Translation divergence in English-Hindi MT. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT 2005)*, pages 245--254, Budapest, Hungary.

Makinouchi, A. (1970). *Algorithmes de traduction automatique du japonais en français*. PhD thesis, Université Grenoble 1.

Mangeot, M. and Kuroda, K. (2003). Divergences interlinguistiques dans le dictionnaire multilingue Papillon. In *Proceedings of MTT 2003*, Paris, France.

Masuoka, T. and Takubo, Y. (1992). 基礎日本語文法 *(Fundamental Grammar of Japanese)*. Kurishio Shuppan, Tokyo.

Microsoft (2012). Bing Translator. www.microsofttranslator.com.

Miyagawa, S. (2005). Unifying agreement and agreementless languages. In *Proceedings of the third Workshop on Altaic in Formal Linguistics (WAFL3)*.

Mochizuki, M., Nakazawa, T., and Kurohashi, S. (2011). 構造を持った定型表現の自動獲得と機械翻訳での利用 (Automatic acquisition of structured fixed expression and use in machine translation). In *Proceedings of the 17th annual meeting of the association of Natural Language Processing (Japan) (NLP 2011)*, pages 975--978, Toyohashi, Japan.

MT Labs (2012). Yakuse!! GOMA. http://www.mtlabs.co.jp/english/product.htm.

Murata, M., Uchimoto, K., Ma, Q., Kanamaru, T., and Isahara, H. (2005). Error Analysis of Translation of Tense, Aspect, and Modality in Machine Translation System. In *Proceedings of FIT 2005*, pages 77--80.

Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. In *Proc. of the international NATO symposium on Artificial and human intelligence*, pages 173--180, Lyon, France.

Nagao, M. (1989). *Machine Translation, How far can it go?* Oxford University Press.

Nagao, M. and Tsujii, J. (1986). The Transfer Phase of the Mu Machine Translation System. In *Proceedings of the 11th International Conference on Computational Linguistics (COLING 86)*, pages 97--103, Bonn, Germany.

Nagara, S. (1990). *Japanese for everyone, a functional approach to daily communication*. Gakken.

Nakaiwa, H., Yokoo, A., and Ikehara, S. (1994). A system of verbal semantic attributes focused on the syntactic correspondence between Japanese and English. In *Proceedings of the 15th conference on Computational Linguistics (COLING 94)*, pages 672--678, Kyoto, Japan.

Nakamura-Delloye, Y. (2003). Analyse syntaxique du japonais. Master's thesis, Institut National des Langues et Civilisations Orientales.

Nakamura-Delloye, Y. (2005). Système AlAleR, Alignemement au niveau phrastique des textes parallèles français-japonais. In *Proceedings of RECITAL 2005*, Dourdan, France.

Nakamura-Delloye, Y. (2007). *Alignement Automatique de Textes Parallèles Français-Japonais*. PhD thesis, Université Paris Diderot (Paris 7).

Nakazawa, T. and Kurohashi, S. (2010). Fully Syntactic EBMT System of Kyoto Team in NTCIR-8. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-8)*, pages 403--410, Tokyo.

Nakazawa, T., Yu, K., Kawahara, D., and Kurohashi, S. (2006). Example-based Machine Translation based on Deeper NLP. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2006)*, Kyoto, Japan.

Namai, K. (2002). The word status of Japanese adjectives. *Linguistic Inquiry*, 33:340--349.

Nec (2012). Crossroad. http://www.nec.co.jp/middle/meshplus/enterprise/top.html .

NICT (2009). The EDR electronic dictionary. http://nict.go.jp/r/r312/EDR/index.html.

NIJL (2009). Bunrui Goihyo. http://kokken.go.jp/en/publications/bunrui_goihyo .

Nishida, F., Takamatsu, S., and Kuroki, H. (1980). English-Japanese translation through case-structure conversion. In *Proceedings of the 8th conference on Computational Linguistics (COLING 80)*, pages 447--454, Tokyo, Japan.

Nishiyama, K. (2005). Morphological Boundaries of Japanese Adjectives: Reply to Namai. *Linguistic Inquiry*, 36:134--143.

Oberon (2001). Component Pascal language report. http://oberon.ch/pdf/CP-Lang.pdf.

Oki (2012). Yakushite.net. www.yakushite.net.

Palmer, F. R. (1979). *Modality and the English Modals*. Longman.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311--318, Philadelphia.

Pause, P. E. (1997). Interlingual strategies in translation. *Machine Translation and Translation Theory*, pages 175--190.

Quirk, C. and Menezes, A. (2006). Dependency Treelet Translation: The convergence of statistical and example-based machine translation? *Machine Translation*, 20:43--65.

Robitaille, X., Sasaki, Y., Tonoike, M., Sato, S., and Utsuro, T. (2006). Compiling French-Japanese Terminologies from the Web. In *Proceedings of EACL-2006*.

Russo, L., Loaiciga, S., and Gulati, A. (2012). Improving machine translation of null subjects in Italian and Spanish. In *Proceedings of European ACL conference (EACL 2012)*, Avignon, France.

Russo, L. and Wehrli, E. (2011). La traduction automatique des séquences clitiques dans un traducteur à base de règles. In *Proceedings of TALN 2011*, Montpellier, France.

Ryu, K., Mizuno, A., Matsubara, S., and Inagaki, Y. (2004). Incremental Japanese Spoken Language Generation in Simultaneous Machine Interpretation. In *Proceedings of Asian Symposium on Natural Language Processing to Overcome language Barriers*, pages 91--95, Hainan Island, China.

Saint-Jacques, B. (1966). *Analyse structurale de la syntaxe du japonais moderne*. C.Klincksieck.

Sakai, T., Koyama, M., Suzuki, M., and Manabe, T. (2003). Toshiba KIDS at NTCIR-3: Japanese and English-Japanese IR. In *Proceedings of the third NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering (NTCIR-3)*, Tokyo.

Sanchez-Martinez, F., Forcada, M. L., and Way, A. (2009). Hybrid Rule-Based - Example-Based MT: Feeding Apertium with Sub-sentential Translation Units. In *Proceedings of the 3rd Workshop on Example-Based Machine Translation*, pages 11--18, Dublin, Ireland.

Sasaki, M. and Murata, T. (2005). A Pattern-Based Machine Translation System — Yakushite Net MT Engine. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2005)*, Pittsburg, USA.

Sasano, R. and Kurohashi, S. (2011). A Discriminative Approach to Japanese Zero Anaphora Resolution with Large-scale Lexicalized Case Frames. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, Chiang Mai, Thailand.

Sata, I. (1993). Construction of a Bilingual Dictionary for DUET-E/J, Toward High Performance MT. In *Proceedings of Premières journées franco-japonaises sur la traduction assistée par ordinateur*, pages 117--122.

Schwab, D., Goulian, J., and Guillaume, N. (2011). Désambiguïsation lexicale par propagation de mesures sémantiques locales par algorithmes à colonies de fourmis. In *Proceedings of TALN 2011*, Montpellier, France.

Scott, B. E. (1977). Linguistic and computational motivations for the Logos machine translation system. www.mt-archive.info/Scott-1977.pdf.

Seretan, V. and Wehrli, E. (2007). Collocation translation based on sentence alignment and parsing. In *Proceedings of TALN 2007*, Toulouse, France.

Shen, W., Delaney, B., and Anderson, T. (2006). The MIT-LL/AFRL IWSLT-2006 MT System. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2006)*, Kyoto, Japan.

Shirai, S., Bond, F., and Takahashi, Y. (1997). A Hybrid Rule and Example-based Method for Machine Translation. In *Proceedings of NLPRS-97*, Phuket, Thailand.

Shirai, S., Ooyama, Y., Ikehara, S., Miyazaki, M., and Yokoo, A. (1998). Nihongo GoiTaikei ni tsuite. *ci.nii.ac.jp/naid/110002946985*, 98(106):47--52.

Simard, M., Ueffing, N., Isabelle, P., and Kuhn, R. (2007). Rule-based Translation With Statistical Phrase-based Post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 203--206, Prague.

Softissimo (2012). Reverso. www.reverso.net.

Somers, H. (1999). Review Article: Example-based Machine Translation. *Machine Translation*, 14:113--157.

Sumida, A. and Torisawa, K. (2008). Hacking Wikipedia for Hyponymy Relation Acquisition. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*, Hyderabad, India.

Suzuki, H. (2011). 統計的後編集手法を適用したルールベース翻訳と文レベルの自動品質評価との融合 (Integration of sentence-level automated quality assessment with rule-based translation using statistical post-editing techniques). In *Proceedings of the 17th annual meeting of the association of Natural Language Processing (Japan) (NLP 2011)*, pages 1119--1122, Toyohashi, Japan.

Systran (2012). Systranet. www.systranet.com.

TaroWatanabe and Sumita, E. (2011). Machine Translation System Combination by Confusion Forest. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, page 1249–1257, Portland, Oregon.

Thurmair, G. (2009). Comparing different architectures of hybrid Machine-Translation systems. In *Proceedings of MT Summit XII*, Ottawa, Canada.

Toshiba (2012). The Honyaku. http://hon-yaku.toshiba-sol.co.jp.

Toue, K., Izuha, T., and Murakami, J. (2011). 日英方向におけるハイブリッド翻訳とルールベース翻訳の人手評価 (Manual evaluation of Japanese-to-English hybrid translation and rule-based translation). In *Proceedings of the 17th annual meeting of the association of Natural Language Processing (Japan) (NLP 2011)*, pages 1127--1130, Toyohashi, Japan.

Tsujii, J. and Fujita, K. (1991). Lexical Transfer based on bilingual signs: Towards interaction during transfer. In *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics (EACL 91)*, pages 275--280, Berlin, Germany.

Tsujimura, N. (1996). *An introduction to Japanese Linguistics*. Blackwell Publishers.

Tyers, F. M. and Nordfalk, J. (2009). Shallow transfer rule-based machine translation for Swedish to Danish. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, page 28–33, Alicante, Spain.

Uchida, H. (1989). ATLAS II: A Machine Translation System Using Conceptual Structure as an Interlingua. In *Proceedings of MT Summit II*, Munich, Germany.

Uchida, H. and Zhu, M. (2005). UNL2005 for Providing Knowledge Infrastructure. In *Proceedings of the Semantic Computing Workshop (SeC2005)*, Chiba, Japan.

Ueffing, N., Stephan, J., Matusov, E., Dugast, L., Foster, G., Kuhn, R., Senellart, J., and Yang, J. (2008). Tighter Integration of Rule-Based and Statistical MT in Serial System Combination. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, United Kingdom.

UNDL Foundation (2013). UNL explorer. www.undl.org/unlexp.

van der Plas, L. (2008). *Automatic lexico-semantic acquisition for question answering*. PhD thesis, University of Groningen.

van der Plas, L. (2011). Combining Syntactic Co-occurrences and Nearest Neighbours in Distributional Methods to Remedy Data Sparseness. In *Proceedings of the NAACL workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, Boulder, US.

Vauquois, B. and Boitet, C. (1985). Automated Translation at Grenoble University. *Computational Linguistics*, 11(1):28--36.

Wang, C., Collins, M., and Koehn, P. (2007a). Chinese Syntactic Reordering for Statistical Machine Translation. In *Proceedings of EMNLP-CoNLL 2007*.

Wang, W., Knight, K., and Marcu, D. (2007b). Binarizing Syntax Trees to Improve Syntax-Based Machine Translation Accuracy. In *Proceedings of the 2007 Joint Conference on Empirical Method in Natural Language Processing and Computational Natural Language Learning*, pages 746--754, Prague.

Wehrli, E. (2007). Fips, a "Deep" Linguistic Multilingual Parser. In *Proceedings of ACL 2007*, Prague, Czech Republic.

Wehrli, E. (2013). Collocations and anaphora resolution in machine translation. In *Proceedings of the International Congress of Linguists (ICL 19)*, Geneva, Switzerland.

Wehrli, E. and Nerima, L. (2008). Traduction multilingue: le projet MulTra. In *Proceedings of TALN 2008*, Avignon, France.

Wehrli, E. and Nerima, L. (2009). L'analyseur syntaxique Fips. In *Proceedings of 11th International Conference on Parsing Technologies (IWPT 2009)*, Paris, France.

Wehrli, E., Nerima, L., and Scherrer, Y. (2009a). Deep Linguistic Multilingual Translation and Bilingual Dictionaries. In *Proceedings of the fourth Workshop on Statistical Machine Translation (EACL 2009)*, pages 90--94, Athens, Greece.

Wehrli, E., Nerima, L., Seretan, V., and Scherrer, Y. (2009b). On-line and off-line translation aids for non-native readers. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, page 299–303, Mrągowo, Poland.

Wiechetek, L. (2008). Rule-based MT approaches such as Apertium and GramTrans. http://uit.no/Content/84555/mt.pdf.

Wilks, Y. (2009). *Machine Translation, its scope and limits*. Springer.

Witkam, T. (1988). DLT: an industrial R & D project for multilingual MT. In *Proceedings of the 12th conference on Computational linguistics (COLING 88) - Volume 2*, pages 756--759, Budapest, Hungry.

Wu, X. and Tsujii, J. (2011). A Term Translation System Using Hierarchical Phrases and Morphemes. In *Proceedings of the 17th annual meeting of the association of Natural Language Processing (Japan) (NLP 2011)*, pages 806--809, Toyohashi, Japan.

Xiong, D., Zhang, M., and Li, H. (2011). Enhancing Language Models in Statistical Machine Translation with Backward N-grams and Mutual Information Triggers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL2011)*, page 1249–1257, Portland, Oregon.

Yahoo! Japan (2012). Yahoo! Japan Honyaku. honyaku.yahoo.co.jp.

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL 2001)*, pages 523--530, Toulouse, France.

Yasuhara, H. (1993). An Example-Based Multilingual MT System in a Conceptual Language. In *Proceedings of MT Summit IV*, Kobe, Japan.

Zaharin, Y. (1990). Generation of synthesis programs in ROBRA (ARIANE) from string-tree correspondence grammars: or a strategy for synthesis in machine translation. In *Proceedings of the 13th conference on Computational Linguistics (COLING 90) - Volume 2*, pages 425--430, Helsinki, Finland.

Zaidan, O. F. and Callison-Burch, C. (2010). Predicting human-targeted translation edit rate via untrained human annotators. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT 2010)*, pages 369--372, Los Angeles, California.

Zhang, H., Zhang, M., Li, H., Aw, A., and Tan, C. L. (2009). Forest-based Tree Sequence to String Translation Model. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, page 172–180, Singapore.

Zhang, R., Yamamoto, H., Paul, M., Okuma, H., Yasuda, K., Lepage, Y., Denoual, E., Mochihashi, D., Finch, A., and Sumita, E. (2006). The NiCT-ATR Statistical Machine Translation System for IWSLT 2006. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2006)*, Kyoto, Japan.

Zollmann, A., Venugopal, A., Och, F., and Ponte, J. (2008). A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008) - Volume 1*, pages 1145--1152, Manchester, United Kingdom.

UNIVERSITÉ
DE GENÈVE
FACULTÉ DES LETTRES

La Faculté des lettres, sur le préavis d'une commission composée de Messieurs les professeurs Jacques MOESCHLER, président du jury; Eric WEHRLI, directeur de thèse; Sadao KUROHASHI (Université de Kyoto); Christian BOITET (Université Joseph Fourier, Grenoble) ; Timothy BALDWIN (Université de Melbourne), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 21 Janvier 2014

Le Doyen : Nicolas ZUFFEREY

Thèse N° 797