A LEXICAL SEMANTIC STUDY OF FOUR-CHARACTER SINO-JAPANESE COMPOUNDS AND ITS APPLICATION TO MACHINE TRANSLATION

by

Mayo Kudo
BA, University of Victoria 2004

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

In the
Department of Linguistics

SIMON FRASER UNIVERSITY

2007

APPROVAL

Name: **Mayo Kudo**

Degree: **Master of Arts**

Title of Thesis: **A Lexical Semantic Study of Four-Character Sino-Japanese Compounds and its Application to Machine Translation**

Examining Committee:

Chair: **Dr. John D. Alderete**
Assistant Professor, Department of Linguistics

Dr. María Teresa Taboada
Senior Supervisor
Assistant Professor, Department of Linguistics

Dr. Nancy Hedberg
Supervisor
Associate Professor, Department of Linguistics

Dr. Chung-hye Han
Supervisor
Assistant Professor, Department of Linguistics

Dr. Kaori Kabata
External Examiner
Assistant Professor, Department of East Asian Studies
University of Alberta

Date Defended/Approved: September 12, 2006

ABSTRACT

Four-Character Sino-Japanese compounds are a productive word formation process in Japanese. There are many morpho-syntactic analyses on these compounds. However, little has been done on their lexical semantic structure. In this thesis I will provide a syntactically motivated classification system for these compounds, and a lexical semantic analysis of their structure. The lexical semantic analysis is extended to a potential application in Japanese-English Machine Translation.

A lexical semantic analysis reveals that for compounds with a deverbal head, there is an argument relation between the constituents if the head's lexical semantic requirement is fulfilled by the non-head constituent, while the relation is adjunct if it is not fulfilled. The constituents of compounds with a regular noun head are in an attributive relation, and the relation cannot seem to be determined by Lexical Semantics. Compounds with a de-adjectival head require more examples to draw firm conclusions because these compounds are rare.

Keywords: lexical semantics; Japanese compounds; machine translation

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to Dr. María Teresa Taboada, my senior supervisor, for her helpful academic and technical advice, thorough proofreading of my work, and continuous encouragement throughout my MA. I cannot imagine how the state of my mind would have been without her support. I cannot thank her enough for being an attentive listener, reading my drafts from the beginning of my thesis work, providing me with helpful suggestions, and helping me shape my thoughts. I would also like to thank my committee members, Dr. Chung-hye Han and Dr. Nancy Hedberg, and my external examiner, Dr. Kaori Kabata, for the time and energy they have devoted to reading my thesis and their insightful comments.

Special thanks to Loreley Hadic Zabala, Yasuko Sakurai, Tim Choi, Niki Efstathopoulou, and Yudong Liu for their fellowship and friendship. Especially, I would like to thank Loreley and Yasuko for proofreading my thesis in the middle of sunny summer. Lorely and Yasuko, I owe you big time. Thank you, Loreley, for being such a nice office buddy. Thank you, Yasuko, for listening to my incomprehensible or incohesive utterances in Japanese. I got epiphany one day while I was talking to you on the phone. Thank you, Tim and Niki, for pushing me to work. And finally, thank you very much, Yudong, for your help with Unix. Without you, I would not have finished my thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

3MS	3rd person masculine subject
3N	3rd person neutral
ACC	Accusative
ADV	Adverb
ANP	De-adjectival noun phrase
ASP	Aspect or mood (general)
CAUS	Causative
CONJ	Conjunctive
COP	Copula
CopP	Copula phrase
CS	Cislocative (Iroquoian) (see Baker 1988)
GEN	Genitive
N	Noun
NOM	Nominative
O	Object
PP	Prepositional phrase
S-J	Sino-Japanese
SL	Source language
SUF	Suffix
TL	Target language
TOP	Topic
V	Verb
Vi	Intransitive verb
VNP	Deverbal noun phrase
VP	Verb phrase

CHAPTER 1: INTRODUCTION

1.1 Introduction

Noun compounds have long been the subject of study in Natural Language Processing. They pose multiple problems in the automatic processing of language. A noun compound can be defined as a word that consists of more than one noun, expressing a concept that is related to the nouns it consists of, but different in the sense that it is often a subcategory of the head of the compound. Head is, as will be defined in Chapter 2, the element that determines the properties of a word or compound. An example of noun compound in English is *dog food*. *Dog food* is composed of two nouns *dog* and *food*, and is a subcategory of food. Likewise, an example of Japanese noun compound is *juuyou-mondai* (important-issue) 'important issue', which consists of two nouns, *juuyou* and *mondai*, and is a subcategory of the noun *mondai* 'issue'. This thesis explores the relation between the constituents of noun compounds that can then be applied to Natural Language Processing.

Natural Language Processing is a field at the intersection of Computer Science and Linguistics that studies how natural languages as opposed to computer languages can be converted into formal representations and how computers can generate natural languages based on formal representations. Natural Language Processing covers areas such as automated text summarization, machine translation, speech recognition, question and answering, and information retrieval. Noun compounds are one of the issues in Natural Language Processing because first, compounding is a productive word formation process. New compounds can be coined as a need arises. Therefore, it is not practical to encode all possible compounds in the dictionary that the system uses. It is more efficient to treat noun compounds without relying entirely on dictionaries. Secondly, there are numerous associations between the constituents of the compounds, and there is no single compounding rule that accounts for all types of compounds. There have been many algorithms that attempted to identify possible constituent relations of noun compounds.

However, not many algorithms succeeded in treating compounds in the language that was dealt with. Machine Translation, which is translation from one natural language to another done by a machine, is no exception to such obstacles. In Machine Translation, a text in one human language is morphologically and syntactically analyzed by some algorithms and gets converted into another human language. During the morphological and syntactic analysis of the text, noun compounds need to be first recognized as noun compounds because it is possible for a machine to analyze them as two or more separate noun phrases without any association between them. Also, noun compounds in one language are not necessarily expressed by similar noun compounds in the target language into which the text needs to be converted. In order to correctly analyze and translate noun compounds, noun compounds need to be treated as a separate problem from regular words, and processed with specific algorithms.

This thesis attempts to tackle the problem of noun compounds in Japanese-English machine translation. There are numerous types of noun compounding in Japanese. However, four-character Sino-Japanese compounds are chosen for this study because of their high productivity. These compounds are highly frequent in newspaper articles, academic and official documents, but the chance of each compound appearing multiple times is low (Baldwin, 2004). This suggests that Sino-Japanese compounds need to be treated separately from the rest of the text by an algorithm. There are some four-character Sino-Japanese compounds that are highly idiomatic. These are excluded from this study because these compounds are not productive. Further, since the meaning of the compound is not derived from the constituents of the compound, it is impossible to convey the idiomatic meaning by simply translating the constituents of the compound.

A number of researchers have investigated four-character Sino-Japanese compounds. However, there is no unified classification of these compounds suggested in the literature. Consequently, the types of four-character Sino-Japanese compounds found in the language have not been extensively studied. A unified Sino-Japanese compound classification may provide a further understanding of the Sino-Japanese compounds. As mentioned earlier, there are numerous possible associations between the compound constituents. The association is constrained by pragmatics or lexical semantic structure of the compound constituents. In fact, many lexical semanticists have been trying to identify

the lexical semantic constraints that play a role in determining possible associations between the compound constituents. For instance, it has been suggested in English and Japanese that content words and morphemes can be represented by a set of lexical semantic features and some argument structure, and the plausibility of morphologically complex words can be determined by the interaction of these features and the argument structure of the constituents the word contains. Lexical Semantics is an area of Linguistics which explores and tries to understand how words are represented in our mind. Many lexical semantic frameworks treat words as being composed of smaller features, although the types of features used in each framework differ from one another. The details of each framework are discussed in Chapter 3. It has been attested that lexical semantic features can identify the compound constituent relations in synthetic compounds, which are ones whose head is derived from a verb (Lieber, 2004: 58; Takeuchi et al., 2003b). Therefore, Lexical Semantics may be one of the keys to understanding the properties of at least some four-character Sino-Japanese compounds. The lexical semantic properties of Sino-Japanese compounds or noun compounds in general have not been extensively studied in Japanese Linguistics. In this study, the lexical semantic structures of four-character Sino-Japanese compounds are identified and examined to see if they can be used in Japanese-English machine translation.

In order to explore the lexical semantic structures of four-character Sino-Japanese compounds, a parallel bilingual Japanese-English corpus, the Utiyama Corpus (2003), is used. In previous research, it was common to use a monolingual Japanese corpus when collecting Sino-Japanese compounds. A monolingual Japanese corpus is useful when the focus of investigation is solely Japanese. However, since one of the foci of this thesis is to apply the study of Sino-Japanese Lexical Semantics to Japanese-English machine translation, a parallel bilingual corpus is superior; it allows us to observe how four-character Sino-Japanese compounds are translated.

Some scholars have attempted to develop a translation algorithm for four-character Sino-Japanese compounds. However, many ended up analyzing the morpho-syntactic structure of these compounds, and have not come up with a translation algorithm. In this study, a translation algorithm for one type of four-character Sino-

Japanese compounds is suggested based on the analysis of their lexical semantic structures.

1.2 Goals of This Study

As mentioned earlier, there is no unified classification system for four-character Sino-Japanese compounds. A linguistically-motivated classification system for these compounds is the first step to broaden the understanding of the properties of these compounds. The morpho-syntactic properties of Sino-Japanese compounds have been extensively studied, but their lexical semantic structure has not been well investigated. This thesis attempts to provide a lexical semantic analysis of these compounds that can be applied to Natural Language Processing, especially Machine Translation. Based on the proposed classification system and the lexical semantic analysis of Sino-Japanese compounds, an automatic classification algorithm and a translation algorithm for Sino-Japanese compounds that can be applied to Machine Translation are suggested. Regarding the translation algorithm, this thesis suggests only an algorithm for compounds whose head is a deverbal noun since these are the most commonly used four-character Sino-Japanese compounds, according to Takeuchi et al. (2003b).

In summary, the goals of this thesis are to:

- Develop a classification system of four-character Sino-Japanese compounds, based on a corpus study,
- Identify the lexical semantic structure of four-character Sino-Japanese compounds,
- Suggest an automatic classification algorithm for Sino-Japanese compounds, and
- Propose a translation algorithm for Sino-Japanese compounds whose head is deverbal.

1.3 Organization

The organization of this thesis is as follows. Chapter 2 introduces Japanese noun types as well as the types of compounding found in Japanese. Japanese nouns come from three sources, native, Sino-Japanese, and foreign. This chapter introduces the characteristics of each noun type with a focus on Sino-Japanese nouns. It also explains how compounds are generally formed in Japanese. The major approaches in Machine Translation and the previous approaches to Sino-Japanese compound translation are presented in Chapter 3. The major approaches in Machine Translation are the transfer approach and the interlingua approach. This chapter briefly explains the mechanisms of each approach. Regarding the classification and translation of Sino-Japanese compounds, there are also two major approaches, statistical approach and lexical approach. Each approach is explained with some examples. Chapter 4 provides a syntactically motivated classification system for Sino-Japanese compounds. Its focuses are on the justification of the proposed classification and the identification of different syntactic behaviours in each group of compounds. Chapter 5 presents a lexical semantic analysis of Sino-Japanese compounds. It introduces the framework used in this study, and demonstrates how this framework is applied to four-character Sino-Japanese compounds. One of the goals of this study is to find a tool to distinguish one type of Sino-Japanese compounds from others. This chapter discusses how Lexical Semantics can be used to distinguish some types of compounds. Chapter 6 proposes a Sino-Japanese compound classification algorithm and a deverbal compound translation algorithm. The classification algorithm is based on the classification system proposed in Chapter 4. The translation algorithm uses the translation patterns that are found in the Utiyama Corpus, which is a parallel bilingual Japanese-English corpus. The thesis is concluded in Chapter 7, summarizing the findings of each chapter, and the contributions of the study.

CHAPTER 2: JAPANESE NOUNS

2.1 Nouns

All languages have nouns or noun-like elements (Greenberg, 1963). This is one of the language universal properties that are true to all languages. Hopper and Thompson (1984) justify that noun and verb are universal lexical categories from the point of view of discourse. In human communication, nouns are necessary to introduce participants, whether they are animate or inanimate, and verbs are used to describe events or actions. As in all world's languages, Japanese has a major part of speech category called nouns. Nouns are, as Tsujimura (1996) mentions, words that can take a demonstrative such as *kono* 'this', and *sono* 'that', and *ano* 'that' as shown in Example 1. English has a two-way distinction of demonstratives, proximal and distal. However, Japanese has a three-way distinction of demonstratives. The proximal demonstrative is *kono* 'this', which identifies the object that is close to the speaker. There are two distal demonstratives, *sono* 'that' and *ano* 'that'. *Sono* is used when the object the speaker is referring to is closer to the hearer while *ano* is used when the object is far from both the speaker and the hearer. In Japanese, as in English, nouns can also take modifiers that precede them and these modifiers take the genitive case particle *-no*. Another characteristic of nouns mentioned by Tsujimura is that nouns can take the conjunctive *to* 'and', which cannot be used with other parts of speech such as verbs and adjectives as illustrated in Example 2.

Example 1: Use of Demonstratives

a. kono saihu
   this wallet
   'this wallet'

b. ano hon
   that book
   'that book'

c. sono kasa
that (distal) umbrella
'that umbrella'

Example 2: Use of the Conjunctive *–to*

a. ari to kirigirisu
ant CONJ grasshopper
'ant and grasshopper'

b. watashi to anata
I CONJ you
'you and I'

c. *atarasii to kirei
new CONJ pretty

d. *asobu to neru
play CONJ sleep

Within the category of nouns, there are three subcategories: regular nouns, de-adjectival nouns, and deverbal nouns. Deverbal and de-adjectival nouns are not commonly found in Indo-European languages such as English. Sections 2.1.1 and 2.1.2 describe the characteristics of these nouns. In the present study, I adopt the term 'deverbal noun[1]' as opposed to the term 'verbal noun' following Takeuchi et al. (2003a, 2003b, 2001). Some researchers such as Kageyama (1982), Shibatani (1990), Takano (2003) and Yokoyama and Sakuma (1996) use the term 'verbal nouns' for deverbal nouns. However, I consider the term 'deverbal nouns' more appropriate than the term 'verbal nouns' because as will be argued in Chapter 4, these nouns are assumed to be verbs that are de-verbalized. Some argue that they are nouns that are verbalized. However, I will reject this analysis in Chapter 4. Characteristics of deverbal nouns are explained in 2.1.2. Likewise, I adopt the

[1] There is a controversy over whether deverbal nouns are nouns or verbs as will be discussed in 4.5.2. For the time being, deverbal nouns are considered verbs that underwent nominalization. Deverbal nouns can also be viewed as elements that have underspecified category (Manning, 1993). In any case, if they are considered as nouns, the term 'verbal noun' is more appropriate.

term 'de-adjectival nouns' as opposed to the conventional 'adjectival nouns' because they are nouns that are derived from adjectivals[2]. Nonetheless, regular nouns, deverbal nouns, and de-adjectival nouns are all categorically nouns. Regular nouns denote objects, deverbal nouns are typically activity nouns, and adjectival nouns are usually stative nouns.

2.1.1 De-adjectival Nouns

De-adjectival nouns are nouns that possess both characteristics of nouns and adjectives. They behave like nouns in that they can take demonstrative determiners and the copula, as illustrated in Example 3. Example (3a) contains the de-adjectival noun *kirei*, taking a copula. The copula can normally accompany a noun as in (3b) while it cannot take a regular adjective as illustrated in (3d). As Example (3c) shows, adjectives, when they are predicate adjectives, do not need a copula to form a sentence. However, de-adjectival nouns behave differently from regular nouns in that they cannot take any grammatical case markers such as the nominative marker *-ga*, the accusative marker *–o* and so on to function as a subject or an object (Shibatani, 1990), but only function as modifiers or predicate adjectivals. In addition, the adjective-making derivational suffix *rasii* can attach to regular nouns, but cannot be attached to de-adjectival nouns as shown in Example 4. The adjectival characteristics of de-adjectival nouns include the fact that they can take the noun-making derivational suffix *-sa*, which cannot be attached to regular nouns. De-adjectival nouns can also precede regular nouns as modifiers (Tsujimura, 1996). Further, as regular adjectives, they can be modified by intensifying adverbs such as *totemo* 'very' and *chou* 'super'.

Example 3: Noun-like Properties of De-adjectival Nouns (Shibatani 1990: 215 slightly modified)

a. ano hito wa kirei da
that person TOP pretty COP
'that person is pretty'

[2] As will be explained in 2.1.1, adjectivals are different from adjectives in Japanese.

b. ano hito wa gakusei da
that person TOP student COP
'that person is a student'

c. ano hito wa utukusii
that person TOP beautiful
'that person is beautiful'

d. *ano hito wa utukusii da
that person TOP beautiful COP

Example 4: Use of *rasii*

a. gakusei rasii
student Adj. making suffix
'student-like'

b. kicchin rasii
kitchen
'kitchen-like'

c. *kirei rasii
pretty

d. *suteki rasii
splendid

2.1.2 Deverbal Nouns

Many deverbal nouns are of Sino-Japanese origin although there are deverbal nouns that come from native or foreign vocabulary (Tsujimura, 1996). Deverbal nouns are ones that have dual functions as nouns and verbs. The noun-like characteristics of deverbal nouns are that they can take demonstratives, they can function as subject or object by taking the nominative case marker *-ga* or the accusative case marker *-o*, and they can undergo syntactic operations that normally apply to regular nouns (Shibatani, 1990). Deverbal nouns can also be relativized (Kageyama, 1982). The verb-like characteristic of deverbal

nouns is that they can function as verbs when accompanied by the dummy verb *-suru* 'do' (Shibatani, 1990). This dummy verb *-suru* can only attach to nouns that denote activities (Kageyama, 1982). Another prominent characteristic of deverbal nouns is that they can take suffixes that carry temporal meanings such as *tyuu* 'while', *go* 'after', *gatera* 'at the same time' and *izen* 'before' (Iida, 1987; Kageyama, 1982; Shibatani, 1990; Tsujimura, 1996).

2.2 Origin of Japanese Nouns

Japanese nouns can be roughly divided into three groups according to their origin: native, Sino-Japanese, and foreign nouns. Native nouns have their origin in Japanese. Sino-Japanese nouns mostly come from Chinese. However, there are some exceptions, which will be explained in 2.2.1. Foreign nouns have their origin in foreign languages excluding Chinese. Foreign nouns are further explained in 2.2.2.

2.2.1 Sino-Japanese Nouns

Sino-Japanese nouns are ones that are written solely in Chinese characters and are originally of Chinese origin. According to Shibatani (1990), Chinese words were first introduced to the Japanese language as early as before the first century A.D., but they were systematically introduced into Japanese around A.D. 400. The main functions of these words were to record official documents and store academic writings until the 19th century. Chinese words were eventually integrated into colloquial Japanese by the 19th century because of the Meiji Restoration in 1867, which is the period in which Japan went through the march of modernization. During this period, the Japanese government coined many terms, which were mostly translations of English, by utilizing already existing semantically appropriate Chinese characters. According to the 1886 revision of the dictionary *wa-ei gorin shusei* 'Japanese-English Glossary' by J. C. Hapburn, more than 10,000 coined terms had been added since the first edition published in 1867 (Shibatani, 1990). Because of the fact that many so-called Sino-Japanese nouns are coined based on English terms, Sino-Japanese nouns today are not all of Chinese origin. In this study, Sino-Japanese nouns include coined words as well as original Sino-Japanese nouns because, first, the structure of the coined words resembles that of the

original Sino-Japanese noun compounds, and second, both coined words and original Sino-Japanese are equally productive in terms of word formation.

Ueno (1980) reveals that the proportion and characteristics of Sino-Japanese nouns resembles that of Latinate words in English. In English, Latinate words account for 55 percent of the English vocabulary although words of Germanic origin have much higher frequency of occurrence than Latinate words. The prominent characteristic of Latinate words is that they are typically words of abstract concepts and academic vocabulary. The status and proportion of Sino-Japanese words are analogous to those of Latinate words in English because the proportion of Sino-Japanese vocabulary in the Japanese language is similar to that of Latinate words in English. Also, Sino-Japanese words tend to be used in academic or literary writings and express abstract concepts as are English Latinate words. Consequently, many technical words are often Sino-Japanese nouns (Shibatani, 1990).

2.2.2 Foreign Nouns

Foreign nouns have their origin in languages other than Japanese or Chinese. These nouns come from various languages such as Dutch, Spanish, Portuguese, Korean, Thai, and Indonesian. Foreign words have been adapted to the Japanese language in many ways, but the processes of adaptation generally fall into two categories. The first route foreign nouns underwent was the substitution of Chinese characters as briefly described in 2.2.1. When English terms were introduced, semantically appropriate Chinese characters were substituted for the original English terms. Therefore, the pronunciation of the original words was not at all preserved while their semantics was retained.

Other foreign words took another route. Japanese has three distinctive writing systems, *hiragana*, *katakana*, and *kanji*. *Hiragana* is generally used for native grammatical words, *katakana* is used for foreign words, and *kanji* is used for Sino-Japanese and native content words. *Hiragana* and *katakana* are syllabary writing systems. That is, one character represents a syllable. *Kanji* is the Chinese writing system, which is logographic. One character represents some semantic entity. When foreign words came into contact with the Japanese language, the *katakana* writing system was

used to phonetically represent the original pronunciation although the pronunciation was modified to fit Japanese phonotactics. The current trend is to use the *katakana* writing system to maintain the original phonology of the borrowed words (Shibatani, 1990).

Foreign words have two routes they can take when adapting to the Japanese language. Some words only took one route, but others took both, resulting in the creation of synonyms. For instance, the word for gasoline in Japanese took two routes, one which uses Chinese characters, *gyuunyuu* 牛乳, and the other that is written in *katakana*, *miruku* ミルク.

2.3 Compounding

2.3.1 Compounding in Japanese

Compounding is one of the most productive word formation processes in Japanese. As mentioned earlier, Japanese nouns are classified into three groups according to their origin: native, Sino-Japanese, and foreign words. Within each group, there are three types of nouns that are grouped according to their distinct characteristics: regular nouns, de-adjectival nouns, and deverbal nouns. These nouns can be combined with a noun of any type to form compound nouns regardless of their origin. Also, nouns can combine with words that belong to other parts of speech to form compound nouns. Shibatani (1990) shows all possible noun compounds as reproduced in Example 5. Compound nouns can also be combined with another word or even compound to form complex compounds as illustrated in Example 6.

Example 5: Japanese Compounds (Shibatani, 1990)

Native Compounds:

Types	Compounds	English Translation
a. N-N	aki-zora autumn-sky 秋-空	'autumn sky'
	kona-yuki powder-snow 粉-雪	'powdery snow'
b. Adj-N	maru-gao round-face 丸-顔	'round face'
	tika-miti close-street 近-道	'short cut'
c. Vi-N	watari-dori migrate-bird 渡り-鳥	'migratory birds'

Sino-Japanese Compounds[3]:

Types	Compounds	English Translation
a. N-N	ki-soku rule-model 規-則	'rule'
	hu-bo[4] father-mother 父-母	'parents'
b. Adj-N	syoo-u little-rain 少-雨	'slight rain'
	koo-ri high-interest 高-利	'high interest'
c. V-N	si-ketu stop-blood 止-血	'stopping of bleeding'

[3] The two-character "compounds" presented here as compounds in Shibatani (1990), but they are considered non-compound context words in this thesis. Chapter 4 discusses the reasons in detail.
[4] *hu-bo* 'parents' is literally father-mother. Therefore, in this paper, it is considered a dvandva compound explained in 2.3.2. However, Shibatani (1990) classifies it as a Sino-Japanese compound.

	satu-jin kill-person 殺-人	'manslaughter'

Hybrid Compounds:

Types	**Compounds**	**English Translation**
a. S-J-native	dai-dokoro work table-place 台-所	'kitchen'
b. native-S-J	to-kei time-measure 時-計	'clock'
c. S-J-foreign	sekiyu-sutoobu oil-stove 石油-ストーブ	'oil stove'
d. foreign-S-J	taoru-ji towel-fabric タオル-地	'towel cloth'
e. native-foreign	ita-tyoko bar-chocolate 板-チョコ	'chocolate bar'
f. foreign-native	garasu-mado glass-window ガラス-窓	'glass window'
g. foreign-foreign	teeburu-manaa table-manner テーブル-マナー	'table manner'

Example 6: Complex Compounds

a. roudou - kumiai		'labour union'
labour union		
労-働 組-合		
b. paato - sha - in		'part-time employee'
part time company member		
パート 社 員		
c. dauzu - isohurabon - sesshu - ryou		'intake of soy
soy bean isoflavone intake amount		isoflavone'
大豆 イソフラボン 摂取 量		

2.3.2 Native Compounds and Sino-Japanese Compounds

Native compounds and Sino-Japanese compounds are often written in Chinese characters. Therefore, at a glance, it may be hard to distinguish between the two. However, there is a crucial difference between native compounds and Sino-Japanese compounds. That is, the constituent order of the compounds reflects the word order of the origin of their language. One of the major differences between Chinese and Japanese is that Chinese is an S-V-O language whereas Japanese is an S-O-V language (Shibatani, 1990). The structure of morphologically complex native nouns and Sino-Japanese nouns reflects this word-order difference between Japanese and Chinese. Morphologically complex native compounds usually exhibit O-V order while Sino-Japanese compounds have V-O order. As mentioned earlier, many of the currently used Sino-Japanese nouns are not of Chinese origin. However, these coined words also conform to the Chinese word order.

Another characteristic of Sino-Japanese compounds is their high productivity. Namiki (2001: 278-279) states that "[Sino-Japanese] compounds have the unique property that they are recursively formed; in other words, there is no principled limit to the number of constituents that compounds may have, due to the fact that a compound noun freely becomes the base of another compound noun". Some examples of recursively formed Sino-Japanese compounds are provided in Example 7. These examples are taken from Namiki (2001), but slightly modified.

Example 7: Examples of Recursively Formed Compound

a. 国会　対策　委員長
Kokkai taisaku iinchou
Diet measure chairperson
'chairperson for negotiation in the Diet'

b. 中高年　労働　移動　支援　特別　助成金
Chuukounen roudou idou sien tokubetu joseikin
'a special grant-in-aid for supporting aged and middle-aged people who have changed their jobs' (The Asahi, morning edition, 6/10/99 (Thurs.), p4)

In Example (7a), the first noun *kokkai* 'Diet' consists of two bound morphemes, *kok* 'country' and *kai* 'meeting'. This word can become a base when forming a compound. When the second noun *taisaku* 'measure' is attached to the base *kokkai*, the whole compound can then become the base of the compound *kokkai – taisaku – iinchou* 'chairperson for negotiation in the Diet'. Likewise, Example (7b) shows that words are recursively combined to form a complex compound.

2.3.3 Dvandva Compounds

Besides the compound nouns listed in Example 5, Japanese has another type of Sino-Japanese noun compounds called dvandva compounds[5]. Dvandva compounds are not commonly found in modern Indo-European languages. Kageyama (1982) defines dvandva compounds as ones where each member of a dvandva compound has an independent head and all the heads are equal in weight. He states that "each member carries its own accent and a slight pause is put between every two members. This type of coordination may be equated with mere parataxis" (Kageyama, 1982: 235-236). Generally, the two entities of the compounds can be connected by the conjunction *and*. For instance, *oya-ko* 'parent and child', *ten-ti* 'sky and earth', *ni'ti-bei* 'Japan and the United States' are all instances of dvandva compounds because each member contains a

[5] These compounds are sometimes called copulative or coordinative compounds because the constituents of the compound can be connected by 'and'. The term *dvandva* originates in Sanskrit, where *dva* means 'two'.

head and these heads have equal weight. Dvandva compounds can be attached to another noun to form complex compound nouns. For example, *ni'ti-bei houmon* 日米訪問 'Japan and the United States visiting each other' is an instance of a Sino-Japanese compound whose first member is a dvandva compound and the second member a deverbal noun. English also has some instances of dvandva compounds although they are rare, and are not productive. For example, *singer-songwriter* and *teacher-apprentice* are instances of dvandva compounds because the two constituents are joined as if they were connected by a connective 'and'.

2.4 Head Structure

2.4.1 Williams' (1981b) Righthand Head Rule

The notion 'head' is generally defined as an entity that determines the category of the entity of which the head is part. Williams (1981b) provides the definition of 'head' as follows shown in (1).

(1) The definition of 'Head' (Williams, 1981b)

If both X and the head of X are eligible members of category C, then

$X \in C \equiv$ head of $X \in C$

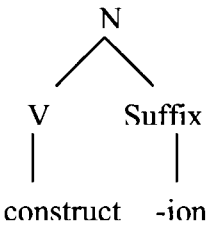
In other words, the head has the lexical-category determining property. Williams investigates English morphology, and observes that there is a general tendency for a suffix to determine the category of a word of which it is a part, but prefixes do not normally possess such property although there are some exceptions as illustrated in Example 8, which is taken from Williams (1981b). Based on the general behaviour of English suffixes and prefixes, Williams proposes the Righthand Head Rule (RHR), which states that "the head of a morphologically complex word to be the righthand member of that word" (Williams, 1981b).

Williams shows some exceptional prefixes that determine the lexical category of the whole word as shown in Example 9. He claims that the prefix *en-* verbalizes its following constituent, and the base does not have the lexical determining property.

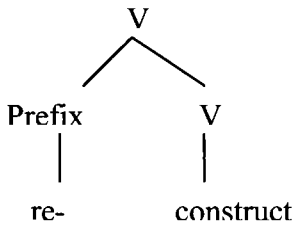
Prefixes such as *en-* are exceptional to the Righthand Head Rule, but he reports that the majority obey the Righthand Head Rule in English, as shown in Example 10. Examples 8-10 are taken from Williams (1981b).

Example 8: Category-Determining Property of Suffixes

a.



b.

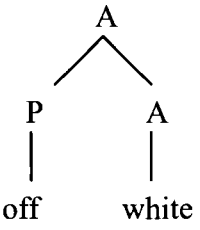


Example 9: Exceptions to Right Head Rule

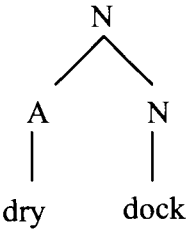
a. en-cage
b. en-dear
c. en-noble
d. en-case

Example 10: Examples of Right Head Rule

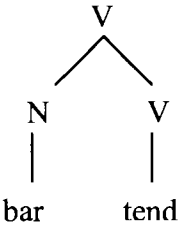
a.



b.



c.



2.4.2 Japanese Head Structure

Tsujimura (1996) and Kageyama (1982) claim that Williams' Righthand Head Rule generally holds for morphologically complex Japanese nouns because, as shown in

Example 11, many Japanese suffixes possess the category-determining property. The data in Example 11 are taken from Tsujimura (1996).

Example 11: Category-Determining Property of Japanese Suffixes (Tsujimura, 1996)

a. hanas (V) + i speak	+ kata (N) way	→ hanas-i-kata (N) 'way of speaking'
b. ama-zuppai (Adj) sweet & sour	+ sa (N) degree-marking	→ ama-zuppa-sa (N) 'the degree of sweet & sour'
c. otoko (N) + rasii (Adj) man -ly	+ sa (N) degree-marking	→ otoko-rasi-sa (N) 'the degree of manliness'

In English, the rightmost constituent of the compound assigns the lexical category of the whole compound as shown in Example 8. It is also the case for Japanese compounds as shown in Example 12. These examples are taken from Kageyama (1982). In Example 12, they all have nouns as their second member, and these determine the lexical category of the compound, regardless of the lexical category of the other member of the compound.

Example 12: Right Head Rule in Japanese Compounds (Kageyama, 1982)

a. hai (N) ash	+ sara (N) dish	→ hai-zara (N) ashtray
b. nagai (Adj) long	+ hanasi (N) talk	→ naga-banasi (N) long talk
c. tabe (V) eat	+ mono (N) thing	→ tabe-mono (N) food

Although the RHR generally holds in Japanese compounds, there are some exceptions as was the case in English. Many morphologically complex two-character Sino-Japanese deverbal compounds[6] violate the Righthand Head Rule because they are left-headed.

[6] As will be discussed in Chapter 4, these compounds are termed as 'nouns' that are normally composed of two Chinese characters. These are considered non-compounds in this study because each Chinese character normally does not have an independent entry as a Sino-Japanese noun in Japanese dictionaries.

Also, as Kageyama (1982) reports, there are a large number of prefixes with de-adjectival contexts and determiner/quantifier-like prefixes that determine the lexical category of the whole compound. Although some Sino-Japanese compounds are left-headed, when left-headed Sino-Japanese compounds are recursively compounded with other Sino-Japanese compounds, the rightmost constituent always becomes the head regardless of the headedness of the base compound.

2.5 Summary

Japanese nouns have a nine-dimensional classification. They are classified according to their origin and properties. Within each type, nouns can be grouped into three types, regular nouns, de-adjectival nouns, and deverbal nouns, based on their characteristics. Types of Japanese nouns and their examples are illustrated in Table 1.

Table 1: Types of Japanese Nouns

	Native	Sino-Japanese[7]	Foreign
Regular noun	kasa 'umbrella'	kaigai 'overseas'	spuun 'spoon'
De-adjectival noun	n/a	humei 'unknown'	kuuru 'cool'
Deverbal noun	kaimono 'shopping'	kenkyuu 'research'	dessan 'draw'

Regarding noun compounds, native, Sino-Japanese, foreign, and dvandva compounds, have been introduced in this chapter. Native compounds and Sino-Japanese compounds are written in Chinese characters, and are distinguished by their constituent order, which reflects the word order of their original language. Native compounds follow Japanese word order, which is O – V, while Sino-Japanese compounds have V - O order, which is the Chinese word order. Foreign compounds are distinguished from other compounds by their orthography. Foreign words and compounds usually use *katakana*, which phonetically transcribes their original pronunciation. Dvandva compounds are ones whose constituents are semantically related to each other, and carry equal weight.

[7] When Sino-Japanese nouns appear in a text, they are always morphologically complex because each morpheme cannot stand alone. The examples in Table 1 are all morphologically complex.

Japanese nouns and compounds, although there are some exceptions, follow the Right Head Rule (Williams, 1981b), which predicts that the rightmost entity within a word determines the lexical category of the whole word. Some Sino-Japanese nouns (compounds in Shibatani (1990)) do not seem to obey the Right Head Rule. However, when they are combined with another Sino-Japanese noun, the rightmost constituent is always the head. At this point, all the examples of Sino-Japanese compounds consist of two characters. However, as will be discussed later in this thesis, these two-character compounds will be considered as non-compound in this study. A further explanation of the treatment of these two-character compounds in this study and the definition of four-character compounds will be provided in Chapter 4.

CHAPTER 3:
JAPANESE COMPOUNDS
AND MACHINE TRANSLATION

This chapter introduces the two major approaches in Machine Translation, which are the transfer approach and the interlingua approach. The transfer approach is explained in 3.1 and the interlingua approach is discussed in 3.2. This thesis attempts to unveil the structure of four-character Sino-Japanese compounds that can be useful in Machine Translation. This chapter discusses how these Sino-Japanese compounds have been treated in previous research. Previously, two major approaches have been used to solve the issue of Sino-Japanese compounds. One approach uses statistical techniques and the other uses Lexical Semantics. Each approach is described and explained with some examples.

3.1 Machine Translation

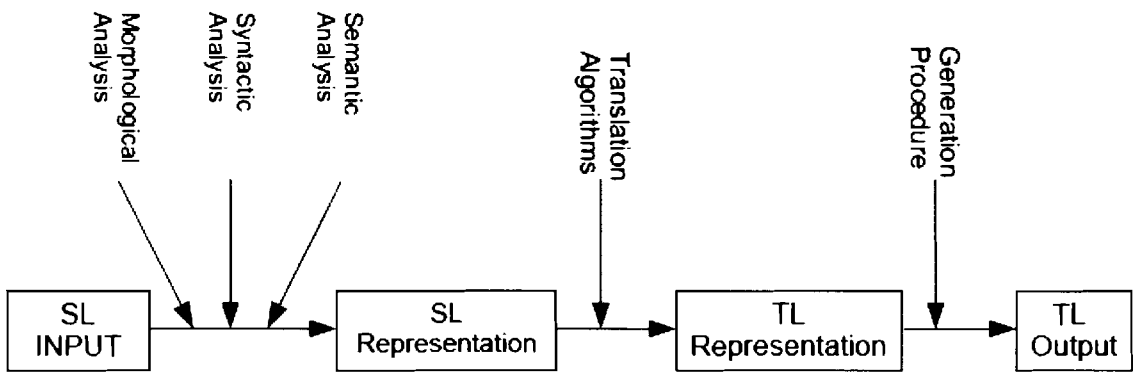
Machine Translation is, as the word indicates, translation by a machine. A machine translates a natural language to another, using algorithms. There are two general approaches to Machine Translation, the transfer approach and the interlingua approach, which are roughly explained in 3.1.1 and 3.2.2. The current trend is to go somewhere between the two approaches, taking the good aspects of both. Regarding algorithms, there are also two types, ones that primarily use human-constructed rules and others that use statistical techniques.

3.1.1 Transfer Approach

The transfer approach is an approach in which the morphology, syntax and semantics of the source language input is analyzed and converted into a source-language dependent representation, which is then transferred into a representation of the language in which the input needs to be translated. A machine generates the target language output from the

target-language dependent representation, using some generation procedure (Trujillo, 1999). Figure 1 roughly describes the procedures of the transfer approach.

Figure 1: Transfer Approach

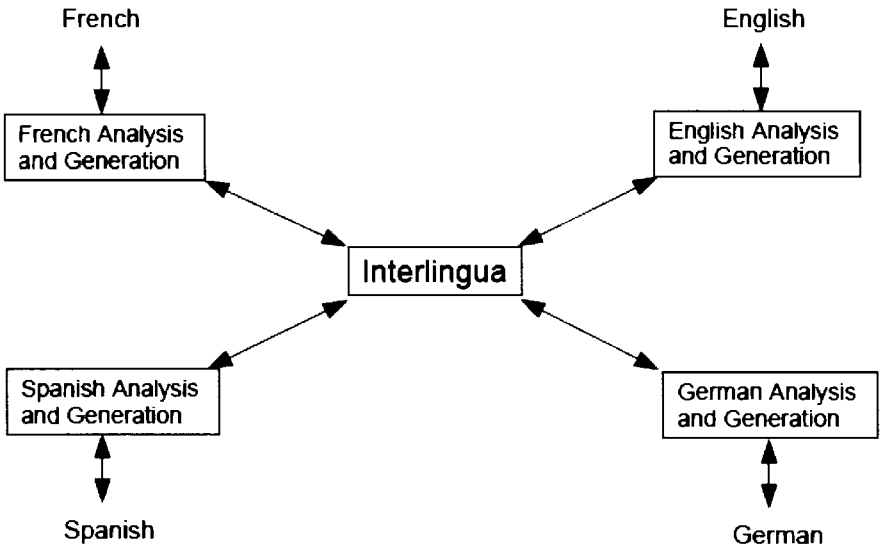


3.1.2 Interlingua Approach

In the interlingua approach, the natural language input is analyzed and converted into a representation that is common to all languages. This language-neutral representation is called interlingua. As Figure 2 shows, the target language translation is generated through the language-neutral representation. Theoretically, the languages the system can handle can be freely translated from one to another. The merit of the interlingua approach is that it can handle multiple languages while keeping the size of the machine translation program small. The transfer approach can handle multiple languages. However, the size becomes big because source language and target language representations are needed for each language pair. For instance, if a machine translation program is a Japanese-Spanish-English translator, the machine needs a translation module for the Japanese-Spanish pair, Spanish-Japanese pair, Japanese-English pair, English-Japanese pair, Spanish-English pair, and English-Spanish pair. The interlingua approach only needs one representation

for all the languages the translation program deals with, in addition to rules for mapping to and from the interlingua[8], keeping the size of the program small.

Figure 2: Interlingua Approach (Trujillo 1999: 167)



3.2 Previous Work on Japanese Compounds in Machine Translation

Noun compounds are one of the most difficult elements to translate in Machine Translation because the thematic association between the constituents of compounds is not clear. For instance, the compound *isi-kettei* 'decision making' is made up of two words, *isi* 'intention, mind' and *kettei* 'decision, settlement'. The first constituent *isi* is a regular noun, and the second member *kettei* is a deverbal noun. In this compound, the first member *isi* acts as the theme of the deverbal noun *kettei* because if one were to paraphrase the compound, it would become *isi-o kettei suru* 'to make up someone's mind'. In other cases, the first member of the compound acts as the instrument of the second member. The compound *kikai-honyaku* 'machine translation' (*kikai* 'machine' and *honyaku* 'translate') has the first member acting as the instrument of the second member because the compound can be paraphrased as 'translation done by a machine'.

[8] It still needs an analysis module and a generation module for each language.

Because the constituents of compound nouns are associated with no apparent lexical semantic rules, it is hard to translate noun compounds in Japanese-English Machine Translation. Many algorithms have been developed to translate Japanese compound nouns into English. However, not many algorithms have been successfully implemented in Machine Translation systems.

As mentioned in Chapter 2, Japanese nouns have three origins, native, Sino-Japanese and foreign. These nouns can combine with another noun regardless of their origin to form a new compound. Although Japanese has many types of compounds, it is far more common to develop an algorithm to translate four-character-Sino-Japanese compound nouns rather than other types of compounds because of their high productivity (Baldwin & Tanaka, 2003b; Takeuchi et al., 2003b; Tanaka, 1993). New terms are coined with existing Chinese characters to form new noun compounds. Also, they are the most frequently used compounds in newspaper articles, academic papers, and official documents (Tanaka, 1993).

Baldwin and Tanaka (2004) investigated the frequency of noun compounds in English and Japanese. For English, they used the British National Corpus (Burnard, 2000) and the Reuters Corpus (Rose et al., 2002), and they used the Mainichi Newspaper corpus ("Mainichi Newspaper Co." 2001) for Japanese. Token coverage in Table 2 refers to "the percentage of words which are contained in NN compounds" in each corpus (Baldwin & Tanaka, 2004). Table 2 shows that, although the size of the corpus is different, Japanese has a significantly larger stock of compounds than English. The Mainichi Newspaper Corpus contains 340 million words, and there are 889,000 types of Sino-Japanese compounds. The average frequency of these Sino-Japanese compounds, which is named as average token frequency, is 11.1 as shown in Table 2. There are many types of Sino-Japanese compounds. Sino-Japanese compounds can be recursively formed as shown in 2.3.2, and the number of possible Sino-Japanese compounds is virtually unlimited. Table 2 also shows that the chance of one Sino-Japanese compound appearing multiple times is low. Therefore, it is impractical to encode Sino-Japanese compounds in a dictionary. Some Japanese dictionaries include Sino-Japanese compounds. However, the coverage is low. Therefore, when translating Japanese to English by Machine

Translation, Sino-Japanese compounds need to be translated without the use of dictionaries.

Table 2: Corpus Occurrence of NN compounds (Baldwin and Tanaka 2004: 24)

	BNC	Reuters	Maichini
Token Coverage	2.6%	3.9%	2.9%
Total Number of Compound Types	265 K	166 K	889 K
Average Token Frequency	4.2	12.7	11.1
Singletons	60.3%	44.9%	45.9%
Size of Corpus	84M	108M	340M

The noun compound translation algorithms that have previously been proposed fall into one of two commonly used methods, statistical or lexical semantic (Takeuchi et al., 2003b). Each approach is described in 3.3 and 3.4 respectively.

3.3 Statistical Approaches

3.3.1 Statistics in Machine Translation

Since large corpora became available, statistics has been a tool to solve problems that could not be handled by linguistically-oriented rules in Natural Language Processing. One of the most frequently used statistics in Machine Translation is Conditional Probability, namely, n-grams. N-grams compute the probability of the occurrence of an entity given the immediately preceding entities in a string of entities. For example, n-grams can be used to guess the likelihood of a word given the previous word in a sentence. The probability of the word *the* given a space and a period is higher than that of the word *elephant* given a space and a period because in English many sentences start with the determiner *the*. In Machine Translation, and Natural Language Processing in general, bi-grams and trigrams are the most commonly used n-grams. Bi-grams compute the probability of the occurrence of an entity given one immediately preceding entity. Likewise, trigrams calculate the probability of the occurrence of an entity given two immediately preceding entities.

3.3.2 Previously Proposed Statistical Approaches

3.3.2.1 Baldwin and Tanaka (2003a)

Baldwin and Tanaka (2003b) propose a template method for translating noun compounds from Japanese to English and vice versa. Their focus is translating noun compounds that can be translated compositionally. For example, *kazan-katudou* 火山-活動 'volcanic activity' consists of two words, *kazan* and *katudou*, which mean 'volcano' and 'activity' respectively. This compound is translated compositionally because each constituent is translated at the word level and concatenated with the modification of the first constituent's part of speech. However, some compounds are not translated compositionally into English. Each constituent of the compound *zangyaku-koui* 残虐-行為 'atrocity' is translated as 'cruel' and 'behaviour' respectively. If the compound were to be translated compositionally, the translation would be 'cruel behaviour'. However, the best translation is 'atrocity'. This type of compounds is not their focus although Baldwin and Tanaka tested the intelligibility of non-compositional compounds when translated compositionally.

Baldwin and Tanaka's method of translating compositional compounds is two-fold. The first step is the generation of translation candidates and the second is the selection of the optimal translation from the translation candidates. Baldwin and Tanaka do not specify which type of Japanese compound nouns are their focus. However, the given examples show that their focus is most likely nouns that are written in Chinese characters, that is, native nouns and Sino-Japanese nouns.

Before the generation phase, the Reuter Corpus and Mainichi Corpus are run through a part-of-speech tagger, and only Noun-Noun compounds are isolated examining Noun-Noun sequences preceded and followed by another part of speech other than a noun. Baldwin and Tanaka also measured the entropy of the preceding and following elements of the Noun-Noun compounds to filter out noun compounds that are part of a larger compound. Entropy is another kind of conditional probability that is used for statistical estimation and pattern recognition (Berger et al., 1996). In the generation phase, the constituents of the noun compound are translated at the word level, and run

through a set of translation templates to generate translation candidates. If any of the constituents of the noun compound cannot be translated at the word level, the generation of translation candidates fails. For instance, *nikka-kankei* 日加-関係 'the relationship between Canada and Japan' consists of three parts, *ni* 'Japan', *ka* 'Canada' and *kaikei* 'relationship, and both *ni* and *ka* are abbreviations. Although the abbreviation for Japan, which is *ni*, may be encoded in a Japanese dictionary, that may not be the case for the abbreviation for Canada. In fact, the EDICT bilingual dictionary (Breen, 1995), which is a freely available machine-readable dictionary, does not contain the abbreviation for Canada.

The translation templates are built based on the noun compounds encoded in the ALTDIC bilingual dictionary (Ikehara et al., 1991) and the EDICT bilingual dictionary (Breen, 1995). Based on the noun compounds found in the dictionary and their translation, Baldwin and Tanaka made translation templates, which are set as gold-standard translation templates. They found a total of 28 templates for Japanese-English translation, and 4 templates for English-Japanese translation. Some examples of the templates are shown in Table 3.

Table 3: Examples of Translation Templates (Baldwin and Tanaka 2003: 81)
N = noun and Adj = Adjective

Templates	Examples
$[N_1\ N_2]_J \rightarrow [N_1\ N_2]_E$	市場・経済 (market · economy) 'market economy'
$[N_1\ N_2]_J \rightarrow [N_2\ N_1]_E$	賛成・多数 (agreement · majority) 'majority agreement'
$[N_1\ N_2]_J \rightarrow [N_2$ of (the) $N_1]_E$	政権・交替 (government · change) 'change of government'

For the selection phase, a machine-learning algorithm, namely Support Vector Machines (Vapnik, 2000), is used to choose the optimal output from the translation candidates generated in the first phase. Support Vector Machines classify data by constructing an N-dimensional hyperplane that optimally separates the cluster of data into two categories. Among a cluster of data, one needs to choose some pieces of data that define a category to construct vectors. Support Vector Machines find the optimal hyperplane that separates the vectors in such a way that a cluster of data of the same category is placed on one side

of the hyperplane, and the other cluster on the other side. In Baldwin and Tanaka's translation candidate selection phase, each translation candidate, including the gold-standard translation, constructs a vector. The vectors that are constructed by the gold-standard translation are treated as positive exemplars, and others are treated as negative exemplars. Support Vector Machines only classify data into two categories by returning a value for each exemplar that is either closer to +1 or -1. Baldwin and Tanaka consider this value to be the translation quality rating. They ran Support Vector Machines over some training exemplars, and ranked the translation candidates according to their value, whether it is closer to the positive class (+1) or the negative class (-1).

Baldwin and Tanaka's method has advantages and disadvantages. Their method is able to cover low-frequency compounds because during the selection of the gold-standard templates, compounds were divided into three groups according to their frequency. The disadvantage is that their method is designed for compounds that can be translated compositionally, and it is not suitable for ones that are translated non-compositionally. For instance, Baldwin and Tanaka's method may translate the aforementioned non-compositional compound, *zangyaku-koui*, as *cruel behaviour*, not *atrocity*. For this particular compound, the compositionally translated one still preserves the meaning of the compound. However, if compounds are somewhat idiomatic or language-specific expressions will not convey the accurate meaning if translated compositionally.

Their method was evaluated using the standard measures of precision, recall, and F-score against the model translations. Precision is the percentage of inputs for which they generated a correct translation. Recall refers to the percentage of inputs for which they could generate a translation. F-score is the mean of precision and recall. Their results show that their method performs at an F-score of 0.68 on compounds that are translated compositionally, and 0.66 on a random sample of 500 noun compounds.

3.3.2.2 Kobayashi et al. (1994)

Kobayashi et al. (1994) propose a statistically-based segmentation method for Sino-Japanese compounds that are composed of four or more characters using the co-occurrence frequency of the Sino-Japanese compound constituents and a thesaurus. They first extracted four-character Sino-Japanese compounds from the Tanaka Corpus (Tanaka,

1993), which is a collection of newspaper articles and contains 130,552 four-character Sino-Japanese compounds. These compounds are segmented in the middle, meaning two two-character nouns[9], because it has been attested that more than 90% of four-character Sino-Japanese compounds are made up of two two-character Sino-Japanese nouns (Hisamitsu & Nitta, 1996; Nomura, 1973; Tanaka, 1993; Yokoyama & Sakuma, 1996). Each constituent of the compounds is then checked if it is encoded in the thesaurus *Bunrui Goi Hyo* 'the Classification Chart of Words' ("The National Institute for the Japanese Language", 1991), which arranges words by hierarchical thesaurus categories such as parts of speech and properties of words. Kobayashi et al. discarded compounds for which one or both constituents were not found in the thesaurus and the remaining got their constituents assigned the thesaurus category to which they belong. The number of each thesaurus category co-occurrence patterns that appear in the Tanaka Corpus is calculated, and used as a knowledge base for the segmentation algorithm of Sino-Japanese compounds that contain more than four characters.

In summary, the segmentation procedure for multiple-character Sino-Japanese compounds is as follows. First, all possible segmentations for the input are generated, and each constituent is assigned a thesaurus category. Using the knowledge of the thesaurus category co-occurrence acquired from the four-character Sino-Japanese compounds, the preference of the co-occurrence of the thesaurus categories is calculated to select the optimal segmentation for the input.

Kobayashi et al.'s segmentation method was tested on 954 four-character compounds, 710 five-character compounds, and 786 six-character compounds. These compounds do not overlap with ones that appear in the Tanaka Corpus. Kobayashi et al. report that their method identifies the correct segmentation for 96% of the four-character compounds. However, the performance deteriorates as the compound becomes longer. The decrease in performance may be because the words are missing from the dictionary or the heuristics they adopted are not yet suitable for compounds consisting of five or six characters.

[9] One may argue that two-character Sino-Japanese nouns are compounds. However, these are treated as single nouns, denoting one concept, as discussed in Chapter 4.

Kobayashi et al.'s study focuses on developing a segmentation method. Although they do not mention how these compounds can be translated in Machine Translation, their study is important because the translation of compounds occurs after the segmentation phase in Machine Translation. The translation phase depends on the performance of the segmentation algorithm. Therefore, it is essential to look at segmentation algorithms as well as translation algorithms. In this study, the proposed classification algorithm and the translation algorithm, which will be presented in Chapter 6, depend on the segmentation by a Japanese morphological analyzer/part of speech tagger, called ChaSen. The details of ChaSen can be found in 4.3.2.

3.4 Lexical Semantic Approaches

3.4.1 Lexical Semantics

Lexical Semantics is an area of Linguistics which explores and tries to understand how words are represented in our mind. It is generally agreed that a word can be decomposed into morphemes, but there is no correspondence between the sound and the meaning of the morpheme. However, many agree that the morphemes have an underlying semantic structure that is compositional, that is, the meaning of the morpheme can be broken down into smaller pieces. Lexical Semanticists attempt to identify the correspondence between the meaning of the morpheme and the structural unit. Jackendoff (1992: 10) states that "there is a form of mental representation called *conceptual structure* that is common to all natural languages and that serves as the syntax of thought". In general, word meaning or morpheme meaning is decomposed into primitives or features, and the argument structure of words or the properties of the argument structure are represented using those features. There has been a variety of sets of features and semantic representation schemas. For example, Jackendoff (1990), Pustejovsky (1995), Wierzbicka (1996), Szymanek (1988) and Lieber (2004) all use a different set of features and representation schemas.

Jackendoff (1972, 1983, 1987b, 1990, 1991, 1996a, 1996b) is largely concerned with the structural description of the meaning of verbs. Using primitives such as BE, GO, STAY, ORIENT, CAUSE, TO, FROM, THING, and PATH, and in his later work, [bounded] and

[internal structure], he argues that the arguments and the meaning of a word is hierarchically arranged as shown in Example 13.

Example 13: An Example of Jackendoff' s Lexical Conceptual Structure (Jackendoff 1990: 45)

Sentence: John ran into the room.

Syntactic Structure: $[_S\ [_{NP}$ John$]\ [_{VP}$ ran $[_{PP}$ into $[_{NP}$ the room$]]]]$

Conceptual Structure:

$[_{Event}$ **GO** $([_{Thing}$ **JOHN**$],\ [_{Path}$ **TO** $([_{Place}$ **IN** $([_{Thing}$ **ROOM**$])])])])]$

Lieber (2004: 6) states that "[Jackendoff's primitives are] insufficiently cross-categorial to allow a full description of the semantics of nouns and adjectives". Expanding Jackendoff's framework, Lieber develops a framework that accounts for nouns, adjectives and prepositions in addition to verbs.

Wierzbicka (1972, 1980, 1985, 1988, 1996) also believes that word meaning is decompositional. However, unlike Jackendoff or other lexical semanticists, she argues that primitives should be words that cannot be further spelled out in simpler words. In other words, she states that these primitives themselves must be words because she argues that the primitives do not assign meaning to the sentence but rather the meaning is only assigned by syntax or sentences in natural languages. The number of primitives used in her framework is 56 in Wierzbicka (1996), and the primitives include I, YOU, HERE, NOW, DO, HAPPEN, to name a few.

Szymanek (1988: 32) proposes another representation of English and Czech word formation. He assumes that any derived word must be somehow related to the lexeme from which it is derived. His representation schema is not feature- or primitive-based. The structure of morphologically-complex words are represented with labelled brackets and tree diagrams, and word formation rules are represented in the form of $X \rightarrow Y / A \_ B$ which states that X becomes Y in the environment in which A and/or B occur. His

focus is on the morpho-syntactic structure of morphologically-complex words, and he largely ignores the lexical semantic aspects of them.

Pustejovsky, the founder of the Generative Lexicon, argues that lexical semantic representation is composed of four levels, argument structure, event structure, qualia structure, and lexical inheritance structure. For each structure, Pustejovsky (1995: 58) defines them as follows:

> These include the notion of argument structure, which specifies the number and type of arguments that a lexical item carries; an event structure of sufficient richness to characterize not only the basic event type of a lexical item, but also internal, subeventual structure; a qualia structure, representing the different modes of predication possible with a lexical item; and a lexical inheritance structure, which identifies how a lexical structure is related to other structures in the dictionary, however it is constructed.

These four levels are linked by three devices, type coercion, selective binding, and co-composition, operating at different levels of the lexical semantic representation to generate lexicons. The definition of the three devices is provided in Table 4. One of the main goals of his lexical semantic model is to capture the fact that a word can potentially have an infinite number of senses depending on how it is used in a sentence while people only store limited number of senses in the lexicon. Therefore, syntax plays a crucial role in his framework. In fact he states "without an appreciation of the syntactic structure of a language, the study of lexical semantics is bound to fail. There is no way in which meaning can be completely divorced from the structure that carries it" (Pustejovsky, 1995: 5).

Table 4: Definition of Three Devices (Pustejovsky, 1995: 61-62)

Type Coercion	Where a lexical item or phrase is coerced to a semantic interpretation by a governing item in the phrase, without change of its syntactic type.
Selective Binding	Where a lexical item or phrase operates specifically on the substructure of a phrase, without changing the overall type in the composition.
Co-composition	Where multiple elements within a phrase behave as functors, generating new non-lexicalized senses for the words in composition. This also includes cases of underspecified semantic forms becoming contextually enriched, such as *manner co-composition, feature transcription,* and *light verb specification.*

Lieber (2004), expanding Jackendoff's Lexical Conceptual Structure, develops a framework of lexical semantic representation in order to reveal the structure of English non-inflectional word formation, derivation, compounding, and conversion. Lieber's lexical semantic representation consists of two parts, event structure template and semantic representation. Like Jackendoff's Lexical Conceptual Structure, her lexical semantic representation is hierarchically arranged. Details of her framework will be further described in 5.1.1.

3.4.2 Previously Proposed Lexical Semantic Approaches

3.4.2.1 Yokoyama and Sakuma (1996)

Yokoyama and Sakuma (1996) use lexical semantic features to classify the constituents of four-character Sino-Japanese compounds and investigate which features can co-occur and which cannot. They paid attention to the fact that the last constituent of Sino-Japanese compounds determines the category and the property of the whole compound. For instance, the compound *keisan-ki* 計算-機 'calculator' can be segmented into two parts, *keisan* 'calculation' and *ki* 'machine', and the last constituent *ki* 'machine' determines the lexical category of the whole word. It also determines the property of the whole compound: that it is a machine because a calculator is a machine that calculates, not a calculation that makes a machine. In other words, the last constituent is the head, and the rest of the constituents are the modifier to the head as explained in Chapter 2.

Yokoyama and Sakuma's classification of nouns and lexical semantic features used for their study are shown in Table 5. Nouns are grouped into three types: concrete nouns, phenomenon nouns, and abstract nouns. Under concrete nouns, there are nouns that are perceived as organizations and ones that are products. Product nouns are further classified into three types according to the properties of the nouns. Phenomenon nouns express events that occurred naturally. They include disaster nouns and phenomenon nouns. Abstract nouns are likewise classified into 28 types according to their properties as shown in Table 5.

Yokoyama and Sakuma assigned to each lexical semantic feature approximately five Sino-Japanese nouns, a total of 155 nouns that can be the head of Sino-Japanese compounds. These nouns are then combined with the same 155 nouns one by one to generate 24,025 compounds. Yokoyama and Sakuma report that only 22.7% of these generated compounds are legitimate compounds. Legitimate compounds are ones that carry a semantic meaning. Yokoyama and Sakuma then analyzed the lexical semantic features of those legitimate compounds to determine which features can co-occur in Sino-Japanese compounds. Based on their analysis, they posited rules[10] on the permissible co-occurrence of lexical semantic features, and tested their rules on the compounds generated by the 155 nouns. Their rules isolated possible Japanese compounds from impossible compounds with 70.4% accuracy. Among ones that their rules identified as impossible compounds, 21% of them were judged as possible Japanese compounds by humans.

Yokoyama and Sakuma admit that they may need a finer set of lexical semantic features and more precise rules to isolate legitimate compounds from illegitimate compounds with better accuracy. Nonetheless, their study contributes to a better understanding of the structure and the properties of Sino-Japanese compounds, and in turn, a better translation algorithm for Sino-Japanese compounds.

[10] Yokoyama and Sakuma do not provide the exact rules they posited.

Table 5: Lexical Semantic Features by Yokoyama and Sakuma (1996: 307)

Types of Nouns	Lexical Semantic Feature	Lexical Semantic Feature
Concrete Nouns	Organization	Organization (ORG)
	Product	Tool (IMP)
		Material (MAT)
		Artificial Product (ART)
Phenomenon Nouns	Nature/Biological	Disaster (DIS)
		Phenomenon (PHE)
Abstract Nouns	Action/Movement	Action (ACT)
		Movement (MOV)
	Philosophy	Idea (IDE)
	Language	Character (CHA)
		Figure (FIG)
		Rule (NOR)
		Information (INF)
		Technical Terms (SCI)
	Property	Property (PRO)
		Ability (ABI)
	Space/Place	Space (SPA)
		Building (BUI)
	Time	Term (TER)
		Results (RES)
		Process (TRA)
	Quantity	COE
		Measure (MEA)
	Others	Method (MET)
		Occupations (OCC)
		Condition (EXT)
		Sphere (SPH)
		Scientific Sphere (SCH)
		Standard (STA)
		Degree (DEG)

		Circumstance?[11] (CIR)
		Shape (SHA)
		Finance (MON)
		Relationship (SOC)

3.4.2.2 Miyazaki et al. (1993)

Miyazaki et al. (1993) identify the issue in Machine Translation that each compound constituent is translatable with the use of a dictionary, but when the constituents are combined to form a compound, it is no longer translatable because the compound is not encoded in the dictionary. Miyazaki et al. make use of the fact that noun compounds are segmentable in order to develop an algorithm for noun compound translation. They analyzed the internal structure of noun compounds and the particle that is attached to the noun compound in the text. The particles attaching to the noun compounds are one of the keys to translation because these particles often indicate the compound's part of speech in English when the compounds are translated. The noun compounds in question include all types of compounds listed in Example 5 in Chapter 2, including Sino-Japanese compounds.

Based on their analysis of noun compound structures, they identified 15 structure types, which include proper nouns, ones that contain numeric expressions, date, or occupation. Proper nouns, and nouns that contain numeric expressions, date, or occupations are beyond the focus of this thesis, and left for future research. Excluding those types, Miyazaki et al. identify six structure types shown in Table 6.

[11] Because Yokoyama and Sakuma do not provide any explanation on their abbreviations, it is not certain what COE and CIR stand for. The rest are guessed from the abbreviation.

Table 6: Constituency of Japanese Compounds (Miyazaki et al., 1993)

1	Prefix + N	[極・[超・[短波] [extremely [super[short wave]]] 'super short wave
2	N + Suffix	[[[大型]・機]・用] [[[jumbo] machine] for] 'for large machins'
3	Deverba N + N	[[処理]・[手順]] [[processing] [procedure]] 'processing procedure'
4	N + Deverbal N	[[データ]・[処理]] [[data] [processing]] 'data processing'
5	De-adjectival + N	[[特別]・[料金]] [[special] [fee]] 'special fee'
6	N + De-adjectival	[[人気]・[絶頂]] [[populariy] [peek]] 'the peek of popularity'

The square brackets in Table 6 show constituency. Miyazaki et al. report that it is always the case that the left constituent modifies the right constituent. For instance, the word for *super shortwave* has three constituents. The leftmost constituent modifies the adjacent constituent *super short wave*, which consists of two constituents, the left of which modifies the right. For ones that do not have any affix, Miyazaki et al. point out that the first member of the compound is either an adjunct or an internal argument of the last member of the compound.

Some rules on identifying the structure of noun compounds are constructed based on their analysis. Miyazaki et al. tested their rules on 201 noun compounds, and their rules yielded the correct constituent relation for 94.6% of the noun compounds. The aim of their study was to compact an existing machine-readable Japanese dictionary by processing Sino-Japanese compounds with an algorithm because it was inefficient to encode all possible Sino-Japanese compounds in the dictionary. They implemented their rules in the ALT-J/E dictionary (Ikehara et al., 1991) and tested if their method could keep the size of the dictionary relatively compact. They state that if they included frequently-used noun compounds in the dictionary, the dictionary would need to encode 130,000 nouns and compounds. However, their rules of noun compounds keep the dictionary compact because the dictionary only needed to encode 70,000 words to handle frequently-used noun compounds.

3.4.2.3 Takeuchi et al. (2003b)

Another lexical semantic approach is proposed by Takeuchi et al. (2003b). They only analyzed four-character Sino-Japanese compounds whose head is a deverbal noun

because the majority of Sino-Japanese compounds have a deverbal noun as their head. The non-head element can be an internal argument or an adjunct of the deverbal head. The association between the head and the non-head element is analyzed in the study of Takeuchi et al. using a formal lexical semantic framework, namely, Lexical Conceptual Structure. They used a set of lexical conceptual structure types previously provided by Jackendoff (1990), Kageyama (1993), and Pustejovsky (1995) as shown in Table 7.

In Table 7, 'x' denotes an external argument and 'y' and 'z' denote internal arguments as stated by Levin and Hovav (1995). The terms 'external argument' and 'internal argument' originate in Williams (1981a). It has been proposed that the direct object of a transitive verb is the true argument of the verb whereas the agent is not. Kratzer (1996) proposes that the agentive subject is projected under Voice Phrase, which is projected above VP. In any case, the term 'internal argument' refers to the direct object of a transitive verb and 'external argument' refers to the agentive subject of the verb.

According to Takeuchi et al., Types 1 – 4, 8, 9 represent different types of transitive verbs, Types 11 and 12 represent intransitive verbs, Type 5 represents ergative verbs, and Types 6, 7, and 10 represent unaccusative verbs.

Table 7: Types of Lexical Conceptual Structure (Takeuchi et al., 2003b)

	Structure Types	English Examples
1	[x ACT ON y]	calculate, operate
2	[x CONTROL [BECOME [y BE AT z]]]	process, translate
3	[x CONTROL [BECOME [y NOT BE AT z]]]	shield, deter
4	[x CONTROL [y MOVE TO z]]	transmit, propagate
5	[x=y CONTROL [BECOME [y BE AT z]]]	recover, close
6	[BECOME [y BE AT z]]	become saturated, be distributed
7	[y MOVE TO z]	move, transmit
8	[x CONTROL [y BE AT z]]]	maintain, protect
9	[x CONTROL [BECOME [x BE WITH y]]]	recognize, predict
10	[y BE AT z]	exist, locate
11	[x ACT]	hold a meeting, queue
12	[x CONTROL [BECOME [[FILLED]y BE at z]]]	sign-name

Takeuchi et al. first divided the four-character Sino-Japanese compounds in the middle as did Kobayashi et al. (1994) and many other researchers. That is, the four-character compounds divide into two two-character nouns, because up to 96% of four-character compounds have the last two characters as their head and the first two as their modifier or the internal argument to the head (Hisamitsu & Nitta, 1996; Kobayashi et al., 1994; Nomura, 1973; Tanaka, 1993; Yokoyama & Sakuma, 1996). Takeuchi et al. examine the lexical semantic structure of the head, which is the second member of the compound, using the lexical structure types illustrated in Table 7. After determining the structure of the head, the non-head constituents are examined. The non-head constituents can be an internal argument of or an adjunct to the head depending on the compounds. For example, *kikai-hon'yaku* 機械翻訳 'machine translation' and *kikai-sousa* 機械操作 'machine operation' both contain *kikai* 機械 'machine' as the first member of the compound. The *kikai* in the first compound *kikai-hon'yaku* is an adjunct because the compound can be paraphrased as 'translation done by a machine' while the second compound *kikai-sousa*, has an argument relation because it can be paraphrased as 'a machine operates'. In order to disambiguate which non-head constituent is the internal argument or adjunct, Takeuchi et al. first classified the non-head constituents into

+accusative and –accusative. In other words, they examined the non-head constituents as to whether they can take the accusative case (ACC). If they can take the accusative case, the non-head constituents are +ACC, if not, they are –ACC. There are some non-head constituents that cannot take ACC. These are regarded as adjuncts because deverbal heads are often derived from a transitive verb, which takes an accusative noun as their direct object, which is the internal argument. The non-head constituents that can take the accusative case are further classified into four groups that are illustrated in Table 8. If the label of the non-head constituent is compatible with the lexical structure of the head, the non-head constituent is considered an internal argument of the head. If the label is not compatible, the non-head constituent is considered an adjunct to the deverbal head. For instance, a +ON modifier can be an internal argument of the deverbal heads that have the structure [x ACT ON y]. The noun *kikai* 'machine' is +ON because it can be the internal argument of the verb *operate* which has the structure of [x ACT ON y] shown in Table 7. Likewise, +EC modifier can be an internal argument of the deverbal heads that have, for instance, the structure [x CONTROL [BECOME [y BE AT z]]]. The modifier of the compound *kikai-honyaku* 'machine translation' is not +EC because it cannot be an internal argument of the verb which has the structure [x CONTROL [BECOME [y BE AT z]]]. Therefore, the non-head constituent is identified as an adjunct to the head. The algorithm for classifying the deverbal heads and their non-head constituent proposed by Takeuchi et al. is shown in Table 9. Their classification method can determine the relations of 858 compounds with 99.3% accuracy.

Takeuchi et al.'s study focuses on the classification and the semantic representation of Sino-Japanese compounds that have a deverbal noun as their head, and does not develop an algorithm that directly leads to the translation of those compounds. However, their study potentially leads to a translation algorithm of Sino-Japanese compounds into English because if the semantic structure of the compounds is available to the machine translation system, it is possible to translate those compounds from the semantic representation by adding translation rules.

Table 8: Lexical Semantic Categories of Modifiers (Takeuchi et al., 2003b)

Lexical Structure Category	Definitions
ON	ON stands for the predicate in 'ACT ON'. *koshou* (fault) and *seinou* (performance) are +ON while *heikou* (parallel) and *rensa* (chain) are –ON.
EC	EC stands for an External argument that Controls an internal argument. *imi* (semantic) and *kairo* (circuit) are +EC while *kikai* (machine) and *densou* (transmission) are –EC.
AL	AL stands for alternation verbs. *fuka* (load) and *jisoku* (flux) are +AL while *kakusan* (diffusion) and *senkei* (linearly) are –AL.
UA	UA stands for UnAccusative verbs. *jiki* (magnetic) and *joutai* (state) are +UA while *junjo* (order) and *heikou* (parallel) are –UA.

Table 9: Takeuchi et al.'s Algorithm for Noun Compound Classification (Takeuchi et al., 2003b)

Step 1	If the modifier has the category –ACC, then declare the relation as adjunct and terminate. If not, go to the next.
Step 2	If the lexical semantic structure of the deverbal head is 10, 11, or 12 in Table 2, then declare the relation as adjunct and terminate. If not, go to the next.
Step 3	The analyzer determines the relation from the interaction of lexical meanings between a deverbal head and a modifier noun. In the case of '-ON', '-EC', '-AL', or '-UA', declare the relation as adjunct and terminate. If not, go to the next.
Step 4	Declare the relation as internal argument and terminate.

3.5 Summary

This chapter introduced two major approaches in Sino-Japanese compound translation. One is statistically oriented while the other focuses on the lexical semantic structure of Sino-Japanese compounds. Some methods presented in this chapter are not specifically for translating Sino-Japanese compounds. For instance, Miyazaki et al.'s study was conducted to handle the astronomical number of compounds in Machine Translation while keeping the size of the dictionary relatively compact. The studies that utilize Lexical Semantics introduced in 3.3.2 focus on figuring out the structure and properties of Sino-Japanese compounds, and have not presented an algorithm to translate those compounds from Japanese to English. However, understanding the structure and the properties of the compounds is an important step towards developing a translation algorithm.

CHAPTER 4:
CLASSIFICATION OF SINO-JAPANESE COMPOUNDS

4.1 Classification of Nouns

This chapter presents a classification system of four-character Sino-Japanese compounds. In the literature, four-character Sino-Japanese compounds are classified in certain ways without any concrete justification. Also, researchers tend to focus on one type of compounds, and do not provide any classification system that accounts for all types of four-character Sino-Japanese compounds. This chapter provides a classification system that is linguistically justified, and accounts for all types of four-character Sino-Japanese compounds. In Chapter 2, nouns are classified by two criteria, their type and origin. Japanese has three types of nouns, regular nouns, de-adjectival nouns, and deverbal nouns, and nouns come from Japanese, Chinese, and other foreign languages. Further, nouns can be combined with other nouns to form compounds.

Regarding Sino-Japanese compound nouns, Shibatani (1990), Kageyama (1993), and Tsujimura (1996) assume that each Chinese character of the Sino-Japanese "compound nouns" constitutes a morpheme containing a semantic meaning, and when a morpheme is combined with another character or morpheme to form a word, this word is considered a Sino-Japanese compound. Therefore, in this view, each character or morpheme is treated as a content word although it cannot normally stand alone. However, what they call "compounds", especially two-character Sino-Japanese "compounds", are normally encoded in the dictionary as non-compound content words because each Chinese character of these "compounds" does not normally appear alone in a sentence nor are they usually encoded in dictionaries. For example, the word *kikai* 機械 'machine' contains two morphemes, *ki* and *kai*. These two Chinese characters are usually used with another Chinese character to represent a concept. In fact, Hayashi (1992) mentions that the majority of Sino-Japanese "compounds" consist of two Chinese characters containing two semantic meanings, but expressing a single concept. There are some Chinese

characters that sometimes appear alone in a sentence. For instance, the word *kaigai* 海外 'overseas' is a Sino-Japanese "compound" because it consists of two morphemes *kai* and *gai*. Each Chinese character, 海 and 外, can be used alone in a sentence, meaning 'sea' and 'outside' respectively. However, when used alone, these Chinese characters are no longer a Sino-Japanese noun, but a native noun; native Japanese nouns are also written in semantically appropriate Chinese characters. Also, the pronunciation of these two characters changes when they appear alone. 海 'sea' is pronounced *umi* and 外 'outside' is pronounced *soto* when used as a native noun, and *kai* and *gai* when use as a Sino-Japanese noun.

Taking into account the structure and the properties of Sino-Japanese compounds and the fact that the majority of two-character Sino-Japanese compounds are encoded in the dictionary as a content word expressing a single concept, it is convenient to treat two-character Sino-Japanese "compounds" as non-compound nouns in Machine Translation. In this study, so-called two-character Sino-Japanese compounds in the traditional Japanese Linguistics are considered single nouns, and are called Sino-Japanese nouns. Many proper nouns, numeric expressions, and dates are written with four Chinese characters. However, those expressions are not considered in this study.

Hisamitsu and Nitta (1996) state that Sino-Japanese noun compounds may contain abbreviations. For instance, *kaisei-daitempou-sekou* 改正大店法施行 'application of the revised large retail shop law' contains three words, *kaisei* 'revision' *daitempou* 'large retail shop law' and *sekou* 'application', one of which, *daitempou*, is an abbreviation. It stands for *dai-kibo kouri tempo hou* 大規模小売店舗法 (*dai-kibo* 'large scale', *kouri* 'retail', *tempo* 'shops', *hou* 'law'). Although it has been attested that some Sino-Japanese compounds may contain an abbreviation, these compounds will not be discussed in the present study either, because they are beyond its scope.

4.2 Four-Character Sino-Japanese Compounds

Many four-character Sino-Japanese noun compounds consist of two two-character Sino-Japanese nouns. The constituents of four-character Sino-Japanese compounds can be any type; regular noun, de-adjectival noun, deverbal noun, prefix, or suffix, as long as they

are Sino-Japanese. For frequently used four-character Sino-Japanese noun compounds that are composed of two two-character Sino-Japanese nouns, the combinations of the constituents are shown in Table 10, which is provided by Yokoyama and Sakuma (1996). Yokoyama and Sakuma collected four-character Sino-Japanese compounds from the Tanaka Corpus (Tanaka, 1993), and identified approximately 11,000 compounds that appear more than ten times[12]. They classified these compounds by the part of speech of the head, and the non-head constituent, and calculated the distribution of the compounds. The results are shown in Table 10. As can be seen, the most common combination is Regular Noun + Regular Noun, which constitutes 53% of the compounds. There is no De-adjectival + De-adjectival combination in commonly used four-character Sino-Japanese compounds. Although rare, this combination is possible. For example, *shousai-humei* 詳細-不明 (*shousai* 'detailed', *humei* 'unknown') 'detail unknown' contains two de-adjectival nouns, and is a Sino-Japanese compound.

Table 10: Classification of Sino-Japanese Compounds (Yokoyama & Sakuma, 1996)

The number '1' in the top cell stands for the first constituent, and the number '2' stands for the second constituent.

1 \ 2	Regular Noun	De-adjectival Noun	Deverbal Noun
Regular Noun	53%	3%	13%
De-adjectival Noun	6%	0%	6%
Deverbal Noun	13%	2%	4%

Although 90% of Sino-Japanese compounds consist of two two-character Sino-Japanese nouns, there are other structures. Tanaka (1993) collected 130,552 Sino-Japanese compounds that appeared in newspaper articles, and compiled all possible structures of four-character Sino-Japanese compounds shown in Table 11.

[12] Yokoyama and Sakuma state that they identified compounds that have the frequency of 10, but they did not define what they meant by it. For the purpose of this paper, I take it as compounds that occurred at least ten times in the corpus.

Table 11 confirms that the vast majority consist of two two-character Sino-Japanese nouns. Types 18 and 22 were not found in the Tanaka corpus. Therefore, these types are not considered in this study.

Tanaka (1993) states that there are a number of compounds that consist of three or four morphemes modifying each other (Types 2-12, 15-20), or compounds which cannot be further segmented (Type 23). Regarding the structure of compounds that consist of more than three morphemes, Tanaka's analysis is accurate. However, ChaSen, which is a Japanese morphological analyzer/part of speech tagger, appears to analyze most of these compounds such as Types 6, 8-11, 15-17, and 19 as ones that consist of two two-character content words. The details of ChaSen are provided in 4.3.2. The compounds in Types 6, 8-11, 15-17, and 19 often contain two-character dvandva compounds. Dvandva compounds are normally considered as having two equally weighted heads. Thus, they were analyzed as having two independent morphemes modifying each other in Tanaka (1993). However, the allowable combinations of the morphemes are limited, and the combination of these morphemes creates a concept that is somewhat separate from each of these morphemes. ChaSen appears to segment two-character dvandva compounds as content words expressing a single concept. Since ChaSen segments them as content words expressing a single concept, two-character dvandva compounds are considered content words in this study.

Others are not analyzed as two two-character words because these compounds contain a content word and one or two affixes attaching to a content word. For instance, *zen-daitouryou* 前-大統領 'former president' consists of a prefix, *zen* 'former', and a content word *daitouryou* 'president'. This type of compound is frequently found in the Utiyama Corpus. However, these are not dealt with in this study because the relation of the affix to the content word is quite clear, and in turn, the translation of these compounds into English is easier than ones whose constituent relation is not clear such as ones that consist of two content words.

For this study, only compounds whose structure is identified as two two-character content words by ChaSen are analyzed. The detail of the compounds selected from the corpus is shown in 4.4.

Table 11: Structure of Four-Character Sino-Japanese Noun Compounds (Tanaka, 1993)

The numbers in the Structure column represent the number of characters. The square brackets show the constituency.

	Structure	Number of Types	Number of Tokens	Examples
1	[[2] [2]]	71,322	368,660	内需. 拡大 domestic demand . enlargement 'increase of domestic demand'
2	[[1 [2]] 1]	1,318	3,798	大. 都市. 圏 big. city . area 'metropolitan area'
3	[1 [[2] 1]]	1,416	3,201	旧. 主流. 派 former . mainstream . clique 'a group of people who formerly followed the mainstream'
4	[[[2] 1] 1]	1,958	2,785	選挙. 区. 制 election. section . system 'electoral district law'
5	[1 [1 [2]]]	188	447	前. 副. 知事 former. vice. Premier 'former vice-premier'
6	[1 [[[1] [1]] 1]]	16	26	他. 府県. 人 other. prefecture. people 'people from other prefectures'
7	[1 [[1] [1] [1]]]	2	3	他. 市. 町. 村 other. city. town. village 'other city, town, or village'
8	[[[[1] [1]] 1] 1]	119	865	中. 長. 期. 的 medium. Long. Period. ADV 'medium to long period'
9	[[1 [[1] [1]]] 1]	12	29	歓. 送. 迎. 会 pleasure. Send. Welcome. Party 'farewell and welcome party'
10	[[2] [[1] [1]]]	116	177	国会. 内. 外 Congress. inside. outside. 'inside and outside of the Congress'
11	[[[1] [1]] [2]]	148	1,262	中. 小. 企業 medium. small. enterprise 'small to medium sized enterprise'
12	[[[1] [2]] 1]	6	42	産. 婦人. 科 birth. woman. section obstetrics and gynecology
13	[1 [3]]	740	6,427	前. 大統領 former. president 'former president'
14	[[3] 1]	941	7,874	自営業. 者 self-employment. person 'self-employed person'
15	[1 [[1] [1] [1]]]	5	92	各. 市. 町. 村 each. city. town. village each city, town, and village
16	[[[1] [1] [1]] 1]	28	135	市. 町. 村. 長 city. town. village. chief 'premier'
17	[2] [1] [1]	39	41	国内. 売. 買 domestic. sell. buy 'domestic trade'
18	[1] [2] [1]	0	0	
19	[1] [1] [2]	285	1,199	男. 女. 平等 man. woman. equality 'gender equality'

20	[1] [1] [1] [1]	42	700	市. 区. 町. 村 city. ward. town. village 'city, ward, town, and village'
21	[1] [3]	7	7	敵. 防衛網 enemy. defense network 'defense network'
22	[3] [1]	0	0	
23	[4]	76	418	一生懸命 'very hard'
24	Proper nouns (excluding place names)	24,818	99,051	中曽根派 'followers of Nakasone'
25	Place names	2,619	9,789	神奈川県 'Kanagawa Prefecture'
26	Numeric expressions	545	1,450	数百万円 'a few million yen'
27	Other	23,786	49,213	連書記長 'secretary of (a) group'
	Total	130,552	557,721	

4.3 Data

4.3.1 The Utiyama Corpus

The Utiyama Corpus (Utiyama & Isahara, 2003), which is a parallel bilingual Japanese-English newspaper articles corpus containing 260,000 translation pairs, was used to collect Sino-Japanese compounds for this study. This corpus is chosen because Sino-Japanese compounds frequently appear in newspaper articles and technical documents as mentioned in Chapter 2. The Japanese component of the corpus was taken from *The Yomiuri Shimbun* (Yomiuri Newspaper), and its English translation was taken from *The Daily Yomiuri*. The Utiyama Corpus provides Japanese sentences and their translations translated by human translators. This corpus is superior to the Tanaka corpus, which was used by Yokoyama and Sakuma (1996), because it allows us to observe the patterns of Sino-Japanese compound translation.

The Utiyama corpus was put through a Japanese morphological analyzer, ChaSen (Matsumoto et al., 1999), to tag the corpus with parts of speech. The description of ChaSen is provided in 4.3.2. In the present study, Sino-Japanese compounds are initially classified into nine types according to the part of speech of the head and the non-head constituent as shown in Table 12. For each type, approximately four hundred compounds are randomly selected, except ones that have a de-adjectival because compounds with a de-adjectival are very rare.

4.3.2 ChaSen

ChaSen (Matsumoto et al., 1999) is a Japanese morphological analyzer/parser that segments Japanese text into free morphemes and bound morphemes and tags these morphemes with their part of speech and pronunciation. The pronunciation is indicated in the Japanese syllabary writing, which can be *hiragana* or *katakana* (See 2.2.2). When the morpheme is conjugated, ChaSen also provides the base form and the conjugation type and form. Regarding nouns, it differentiates three types of nouns, deverbal nouns, regular nouns, and de-adjectival nouns, identified in Chapter 2. The procedure for selecting the optimal part of speech for the morphemes in ChaSen is two-fold. First, ChaSen assigns all possible parts of speech to morphemes and content words with their costs, which are determined by the frequency of the morpheme or the word appearing as a particular part of speech in a part-of-speech tagged corpus[13]. Because they use logarithms for cost calculation, the smaller the number, the more frequent the morpheme is in the corpus. After determining the frequency of each morpheme, ChaSen runs bi-grams on the morphemes and calculate the frequency of two morphemes co-occurring in the corpus. The frequency of co-occurrence is set as their co-occurrence cost. The sum of the morpheme cost and the co-occurrence cost is calculated, and the smallest cost is selected as the optimal part of speech.

4.4 Selection and Classification of Sino-Japanese Compounds

For this study, a total of 1860 compounds were selected from the corpus. These compounds were manually extracted. However, these can be extracted automatically as did Baldwin and Tanaka (2003b, 2004) and others. The extraction method by Baldwin and Tanaka (2003a, 2004) is introduced in 3.3.2.1. The manually extracted compounds are classified into nine types according to the noun type of the head and the non-head constituent as did Yokoyama and Sakuma (1996) shown in Table 10. The number of compounds for each type in the present study is shown in Table 12.

[13] Matsumoto et al. (1999) do not mention the name of the corpus they used in the manual.

Table 12: Types of Compounds and their Distribution

1 \ 2	Deverbal	Regular Noun	De-adjectival
Deverbal	400	409	112
Regular Noun	400	408	86
De-adjectival	21	22	2

1 = Head, 2 = First Constituent

This study used the first half of the Utiyama Corpus, which contains approximately 130,000 translation pairs. Because de-adjectival compounds are very rare, it was hard to obtain the same number of compounds from each group. Other types of compounds such as deverbal compounds and regular noun compounds are relatively easy to find. In this study, compounds were randomly collected until the number of compounds reached approximately 400. There were less than 400 compounds that contained one or two de-adjectival nouns in the first half of the corpus. For these compounds, the compound collection was discontinued when the first half of the corpus was exhausted because considering the number of these compounds in the first half of the corpus, it was not likely that there were 400 of them in the entire corpus.

4.5 Justification of the Present Classification

In the present classification system, Sino-Japanese compounds are first grouped according to the part of speech of their head. The head can be a regular noun, a deverbal noun, or a de-adjectival noun. Compounds are classified this way because, firstly, the translation of the Japanese component of the corpus shows that regular-noun heads tend to be translated as a noun in English, and deverbal heads are translated with a deverbal element such as a tensed verb, progressive, gerundive, or a noun that is derived from a verb as shown in Example 14. Likewise, de-adjectival heads are often translated with an adjective in English.

Example 14: Examples of Translation in the Utiyama Corpus

Head	Compound	English Translation in the Corpus
Deverbal	tanzyun-sikou simple-think/thinking	simple thinking
	ziyuu-kyousou free-compete/competition	free competition
	heiwa-izi peace-maintain/maintenance	maintaining peace
Regular noun	hituyou-zyouken necessary-condition	necessary condition
	shotoku-suizyun income-level	income level
	kokumin-kanzyou citizen-feeling	the sentiment of the people
De-adjectival	zizoku-kanou sustain-possible	sustainable
	teigi-konnan define-hard	hard to define
	ikuzi-huan raise a child-anxiety	anxious to raise a child

Secondly, compounds with a deverbal head demonstrate some characteristics that are not found in compounds with a regular noun head or a de-adjectival head. Specifically, some compounds with a deverbal head show striking similarities with noun incorporation

(Baker, 1988; Mithun, 1984) or with coordinated VP construction while this is not the case for compounds with a regular noun head or a de-adjectival head.

De-adjectival compounds may have a subject-predicate relation. Some de-adjectival compounds whose head is an emotion word may have the non-head constituent of the compound functioning as the theme of the emotion word. Grouping the compounds according to the part of speech of their head elicits a clear picture of the properties of Sino-Japanese compounds. Within each group, compounds are further classified according to the part of speech of the non-head constituent to observe if the properties of the non-head constituent affect the overall properties of the compounds in any way.

4.5.1 Structure of Compounds with a Deverbal Head

4.5.1.1 Noun Incorporation

As mentioned in the previous section, some deverbal compounds resemble noun incorporation. Noun incorporation is an operation in which an argument of a verb co-occurs with the verb, and the argument and the verb appear as a compound word while retaining the argument function. Mithun (1984: 847) considers noun incorporation to be a morphological operation because she states that "noun incorporation is perhaps the most nearly syntactic of all morphological processes". Mithun (1984) identifies four types of noun incorporation across languages. The noun incorporation types are provided in Table 13. Although some consider noun incorporation to be a morphological operation, other researchers such as Baker (1988) consider it to be a syntactic operation. Syntactically, noun incorporation is an operation whereby a theta-role bearing noun head, which can canonically appear as an internal argument of a verb, adjoins with a verb head to form a larger word while retaining the referential property of the theta-role bearing noun (Baker, 1988; Massam, 2001; Mithun, 1984). Here, theta roles are semantic relationships an entity can have with regard to an action or a state (Chomsky, 1981; Gruber, 1965; Jackendoff, 1972, 1976). Some examples of theta roles are agent (the doer of an action) and patient (the undergoer of an action).

One of the characteristics of noun incorporation is that a sentence with an incorporated noun can also be paraphrased without the use of an incorporated noun as

shown in Example 15 although normally they are not in free variation (Gerdts, 1998). The examples are from Onondaga, an American Indian language of the Iroquoian family. As can be seen, in Example 15 (1a. and b.), the theme, *hwist*, 'money' appears separately from the verb in (1b) while it is attached to the verb in (1a). Likewise, the theme is not incorporated in the b. examples in (2) and (3), and it is incorporated in the a. examples. In the sentences in Example 15, there is a difference in specificity, but as Baker states that a. and b. are 'thematic paraphrases' of one another.

Another characteristic of noun incorporation is that there is an animacy hierarchy on the nouns that can be incorporated such that inanimate nouns are more frequently incorporated than animate nouns, and nonhuman animate nouns are more frequently incorporated than human animate nouns (Gerdts, 1998).

Table 13: Noun Incorporation Types (Mithun, 1984)
The table is taken from Baker et al. (2004)

Type	Characteristic Properties
I: Lexical Compounding	• Incorporated Noun (IN) is generic, non-referential. • N + V is a conventional, institutionalized activity.
II: Manipulation of Case	• IN loses argument status. • Another NP takes on the grammatical function it vacates.
III: Manipulation of Discourse	• Noun Incorporation is used productively for discourse purpose, e.g. to background known information.
IV: Classificatory Noun Incorporation	• An IN can be supplemented by more specific NP material external to the complex verb.

Example 15: Paraphrasable Property of Noun Incorporation (Baker 1988: 76-77, originally taken from Woodbury (1975))

(1)a. pet wa?-ha-hwist-ahtu-?t-a?.

Pat PAST-3MS-money-lost-CAUS-ASP
'Pat lost money'

b. pet wa?-ha-htu-?t-a? ne? o-hwist-a?.

Pat PAST-3MS/3N-lost-CAUS-ASP the PRE-money-SUF
'Pat lost the money'

(2)a. (pro) wa?-ha-yv?kw-ahni:nu-?.

PAST-3MS/3N-tobacco-buy-ASP
'He bought tobacco.'

b. (pro) wa?-ha-hninu-? ne? o-yv?kw-a?.

PAST-3MS/3N-buy-ASP the PRE-tobacco-SUF
'He bought the tobacco.'

(3)a. (pro) t-a-shako-?ahs-v:- ?.

CS-PAST-3MS/3F-basket-give-ASP
'He handed a basket to her.'

b. (pro) t-a-shaka-u-? (pro) ka-?ahsœ:- ?

CS-PAST-3MS/3F-give-ASP PRE-basket-SUF
'He gave her a basket.'

Some English noun-verb compounds can be, according to Mithun, a type of noun incorporation. For example, *berry-picking* has an incorporated noun, *berry*, attaching to the verb *picking*. The incorporated noun does not refer to a specific berry, but rather generic berries. Also, the verb *picking*, which is transitive by default, becomes intransitive. Therefore, this compound fits the description of Mithun's Type I noun incorporation, which states that the incorporated noun is generic and non-referential. The other types are different from Type I because the incorporated noun retains its syntactic

identity, and can be definite or referential, meaning that the incorporated noun can be referred to by a pronoun.

The properties of noun incorporation vary cross-linguistically as shown in Table 13. However, some properties are common across languages. For example, it is almost always the case that the direct object of a transitive verb or the theme can undergo noun incorporation, but the subject of a transitive verb or the agent frequently cannot (Baker, 1988; Baker et al., 2004; Mithun, 1984). Mithun states that Instrument, Benefactive, or Location may undergo noun incorporation in some languages but it is not common.

4.5.1.2 Pseudo-Noun Incorporation in Sino-Japanese Compounds with a Deverbal Head

Japanese does not generally have noun incorporation. However, some Sino-Japanese compounds with a deverbal head share characteristics that are found in noun incorporation. It is generally agreed that deverbal nouns have an argument structure (Manning, 1993; Shimada & Kordoni, 2003; Tsujimura, 1992). Observing four-character Sino-Japanese compounds with a deverbal noun head, one notices that the first constituent sometimes serves as the direct object of the deverbal noun, and the thematic role of the direct object is usually Theme or Patient as shown in Example 16. As the translation shows, the first constituent in Example 16 is always the direct object of the deverbal head.

Example 16: Examples of Sino-Japanese Compounds with a Deverbal Head

(1) koyou-soushutu
employment-create
'to create employment' or 'employment creation'

(2) iken-koukan
opinion-exchange
'to exchange opinions' or 'opinion exchange'

(3) enjo-keizoku
aid-continue
'to continue (the) aid' or 'aid continuation'

Another characteristic of Sino-Japanese compounds with a deverbal head is that the compounds can be rephrased by breaking the compound into the non-head constituent and the head, and attaching the accusative case to the non-head constituent as shown in Example 17. In Example 17, the two sentences are a paraphrase of each other. The theme *keizai* 'economy' has an option of taking the accusative case particle or adjoining to the verb *kaikaku* 'reform' without affecting the argument structure and thematic role requirement of the verb *kaikaku* 'reform'.

Example 17: Paraphrase of Sino-Japanese Compounds

(1) seihu-wa soukyuu-ni keizai-**o** kaikaku si-ta
government-TOP urgent-ADV economy-**ACC** reform do-PAST
'The government urgently reformed the economy.'

(2) seihu-wa soukhyuu-ni keizai-kaikaku si-ta
government-TOP urgent-ADV economy-reform do-PAST
'The government reformed the economy.'

Also, the transitive verb behaves as an intransitive when it incorporates a noun. For instance, in Example 17 (1) and Example 18, the noun *keizai* is canonically the direct object of the verb. As Example 18 (1) shows, when the verb incorporates the direct object, the verb can no longer take additional objects because the argument requirement of the verb is already fulfilled by the incorporated noun, which retains its syntactic role. If an additional object, for instance, *nihon-o* in Example 18 (1), needs to be inserted, the deverbal noun *kaikaku* needs to incorporate its direct object *keizai* to form a compound, and the whole compound takes the accusative case. In this case, the dummy verb becomes a main verb, taking the compound *keizai-kaikaku* as its direct object. After the incorporation, an additional "object", *nihon* can be inserted before the whole compound, modifying the whole compound *keizai-kaikaku* as shown in (2). The status of the compound in (2) is discussed later in this section.

Example 18: Change of Transitivity after Incorporating a Noun

(1) *seihu-wa soukyuu-ni nihon-o keizai-kaikaku si-ta
government-TOP urgent-ADV Japan-ACC economy-reform do-PAST

(2) seihu-wa soukyuu-ni nihon-no keizai-kaikaku-o si-ta
government-TOP urgent-ADV Japan-GEN economy-reform-ACC do-PAST
'The government urgently reformed Japan's economy.'

Another interesting property of Sino-Japanese compounding is that the incorporated noun can be referential. For instance, in Example 19, the incorporated noun appears with a topic marker in the following sentence. One of the functions of the topic marker *–wa* is to express discourse-old information (Bond, 2004; Kuno, 1973; Shibatani, 1990). Therefore, *keizai-wa* in Example 19 can be considered referential.

Example 19: Referential Property of the Compounded Noun

(1) seihu-wa keizai-kaikaku si-ta-ga
government-TOP economy-reform do-PAST-however

keizai-wa kaihuku si-nakat-ta
economy-TOP recover do-NEG-PAST

'The government reformed the economy, but the economy did not recover.'

The referential property of the incorporated noun, and the fact that the direct object co-occurs with a deverbal noun in Sino-Japanese compounding are pieces of evidence that Sino-Japanese compounding has strikingly similar characteristics to Mithun's Type III noun incorporation. What is different is that Type III noun incorporation is only found in polysynthetic languages, and the operation of Type III noun incorporation itself is a device to signal discourse-old information. Also, the incorporated noun can be a pronoun. On the other hand, in Sino-Japanese compounding, the compounding is not a device to express discourse-old information. Further, the compounded noun can be referred to by a noun that takes a topic marker, but it cannot be a pronoun when it appears inside the compound. This is maybe because unlike European languages, it is not common to use

pronouns in Japanese although Japanese has a set of personal pronouns (Shibatani, 1990). The information that is understood by the interactants is conveyed through the use of pronoun in English, but it does not overtly appear on the surface in Japanese. The topic marker is used for discourse entities that are already mentioned. Therefore, the topic marker may function similarly to pronouns in English. But in any case, Sino-Japanese compounding itself is not used to express discourse-old information. This property is different from Type III noun incorporation.

Assuming that noun incorporation is a syntactic operation, although it is debatable, this section presents a syntactic analysis of Sino-Japanese compounds with a deverbal head. This paper assumes that the Uniformity of Theta Role Assignment Hypothesis (UTAH) (Baker, 1988) holds in Japanese. The definition of UTAH is provided in (2). Assuming UTAH, I suggest that the sentences in Example 17 are derived from the same underlying form.

(2) Uniformity of Theta Role Assignment (Baker 1988: 46)

> Identical thematic relationships between items are represented by identical structural relationships between those items at the level of D-structure[14].

If we assume that Sino-Japanese compounding is a syntactic operation, we can easily explain why an additional object cannot be inserted after compounding. First, we need to change our assumption on deverbal nouns. Traditionally, deverbal nouns[15] are viewed as nouns that can undergo verbalization by taking the dummy verb *–suru* 'do'. However, there is no reason to reject the hypothesis that these deverbal nouns are originally verbs that can undergo nominalization to become nouns. In fact, Takahashi (2000) argues that this is the case, observing the syntactic as well as phonological behaviour of Sino-Japanese deverbal nouns. She claims that traditionally so-called Sino-Japanese deverbal nouns are in fact verbs, and these verbs undergo zero conversion when functioning as nouns. She argues that regular verbs are bound morphemes that must attach to auxiliaries such as tense, although in the standard view verbs are considered to be free morphemes.

[14] Baker (1988) states that "D-structure ("deep" or underlying structure) [is] a formal syntactic representation at which the thematic relations among items and phrases are directly represented". It is a level at which theta roles are assigned to theta-role receivers.

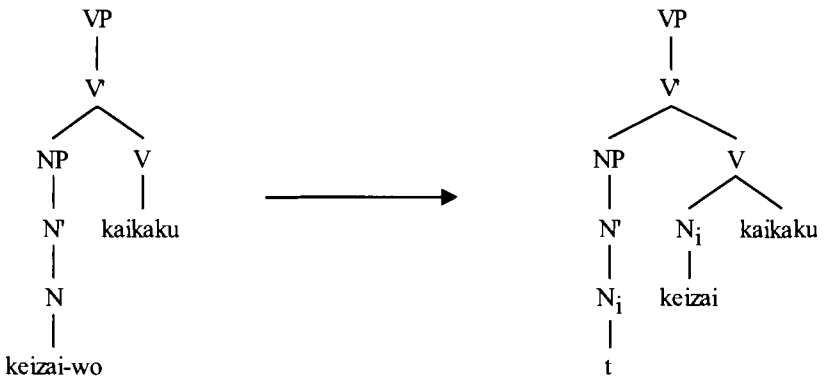
[15] Many researchers use the term 'verbal nouns' if they assume that deverbal nouns are actually nouns that can function as verbs.

On the other hand, Sino-Japanese deverbal nouns are free morphemes and cannot directly attach to verbal particles. She claims that the status of the deverbal nouns is the reason why deverbal nouns must take the dummy verb *-suru.*

If we assume that deverbal nouns are verbs, Sino-Japanese compounds with a deverbal head can be analyzed as follows illustrated in Figure 3. In the tree, the direct object *keizai* gets its theta role from the verb. The theme, which is occupying the canonical direct object position, merges with the verb stem leaving a trace in the position in which it originates. Because the trace or the incorporated noun is fulfilling the argument-structure requirement of the verb, nothing can be inserted after the operation of noun incorporation, forcing the verb to behave like an intransitive verb.

Another motivation for my account is that the deverbal noun and its direct object when appearing as a compound, maintain the thematic relation without the help of the dummy verb *–suru*. This is especially evident in newspaper headlines.

Figure 3: Syntactic Structure of Noun Incorporation in Japanese



In the canonical form, an adverb phrase can be inserted between the direct object and the verb. However, the adverb cannot be incorporated because it is prohibited by the First Sister Principle (Roeper & Siegel, 1978) which states that a transitive verb is allowed to combine only with its first sister noun.

Deverbal Sino-Japanese compounds can take the accusative case marker *–o* to function as a noun as shown in Example 18 (2). In such case, I propose that the verb incorporates the direct object, and the whole compound undergoes nominalization as proposed by Takahashi (2000) to become a noun. The evidence comes from the fact that nothing can be inserted between the constituents of the compound. It may be possible to hypothesize that the NP in Figure 3 is not incorporated, but stays in the canonical position without any case marker because according to Shibatani (1990), case markers, especially the subject marker *–ga* and the accusative marker *–o*, are frequently omitted in Japanese. In such case, the structure can give the appearance of a compound. If we suppose that the first constituent stays in the canonical position without any case marker, an adverbial phrase should be able to be inserted between the NP and V. However, this is not the case. This suggests the possibility that N is incorporated to V. Compounds like in Example 18 are relatively common. The number of such compounds in the present data is provided in Chapter 5.

Sino-Japanese compounds demonstrate some similarities with noun incorporation. However, none of the definition provided by Mithun in Table 13 seems to fit the characteristics of Sino-Japanese compounds. I consider Sino-Japanese compounds that have an internal relation to be situated somewhere between Mithun's Type II and Type III noun incorporation.

When the verb is intransitive or unaccusative, the only available element that can fulfil the argument requirement of the verb is the subject. In fact, when the head of a Sino-Japanese compound is intransitive or unaccusative, the first constituent, if there is an internal relation between the constituents, is the subject, as shown in Example 20.

Example 20: Compounds with an Intransitive Head

(1)a. kinou-ga kaihuku suru
function-NOM recover do
'The function recovers'

b. kinou-kaihuku
function-recover
'the recovery of the function'

(2)a. ninsiki-ga husoku suru
recognition-NOM lack do
'(someone) is ignorant (of something)'

b. ninsiki-busoku
recognition-lack
'lack of understanding' 'ignorance'[16]

Because intransitive verbs do not have a sister, the First Sister Principle does not apply. As mentioned earlier, nouns that can undergo noun incorporation are normally the direct object of a verb. Nonetheless, according to Baker (1988), the incorporation of the subject has been attested. When the subject and the verb in Example 20 appear as a compound, nothing can be inserted between the constituents as shown in Example 21. The inability to insert any adverbial phrases between the constituents of the compound suggests some mechanism that prohibits the insertion of the adverbial phrases. I suggest that it is the incorporation that disallows any element to be inserted between the constituents. The proposed analysis is as follows. The subject and the verb originate in their canonical position. The subject is incorporated into the verb, leaving a trace in the canonical position.

Example 21: No Insertion between the Constituents of the Compound

(1)a. kinou-ga umaku kaihuku suru
function-NOM well recover do
'The function recovers.'

[16] This particular compound is more common with the copula as opposed to the dummy verb although the compound in (1) can appear with both.

b. *kinou-umaku-kaihuku
function-well-recover

(2)a. ninsiki-ga kanari husoku suru
recognition-NOM much lack do
'(someone) is very ignorant (of something)'

b. *ninsiki-kanari-husoku
recognition-much-lack

A similar phenomenon is reported in Korean shown in Example 22. In (1) *hay* 'sun' functions as the subject of the verb *tot* 'rise', and in (2) and (3), *non* 'field' and *koki* 'fish' function as the object of the respective verb. Shi (1997) argues that these compounds are generated by noun incorporation at the level of syntax. He proposes that *hay*, *non*, and *koki* originate in the canonical subject or object position, and receives their theta role from the verb. These nouns then undergo noun incorporation and merge with the verb. The noun making suffix *–i* is attached after the noun incorporation.

Example 22: Korean Synthetic Compounds

(1) hay tot i
sun rise noun-making suffix
'sunrise'

(2) non kal i
rice field till noun-making suffix
'tilling a field'

(3) koki cap i
fish catch noun-making suffix
'fishing'

4.5.1.3 Deverbal Compounds with an Adjunct Relation

The previous section has dealt with compounds within which there is an internal relation such as verb and complement relation. However, there are a number of deverbal compounds that do not demonstrate any internal relation. For instance, *kaihatu* 'development' and *kyousei* 'compulsion' in Example 23 are not the subject or the object of the deverbal noun, but are rather an adjunct.

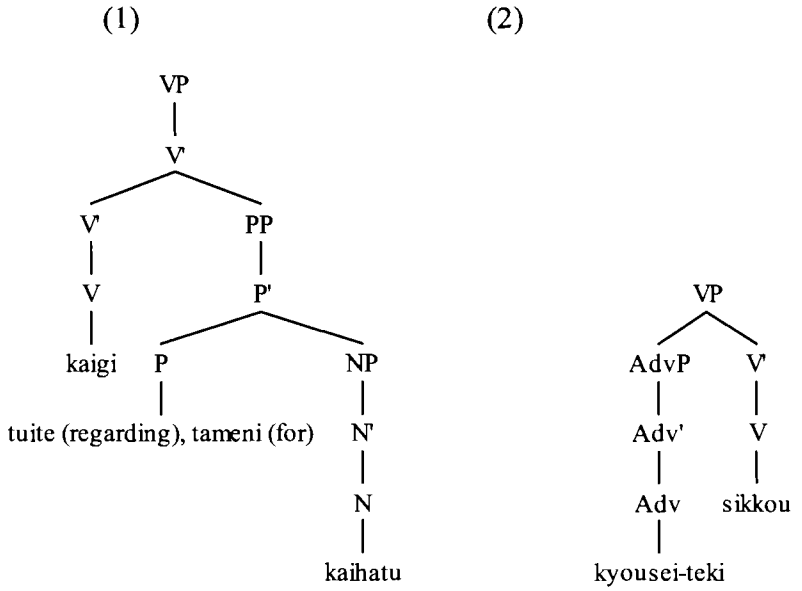
Example 23: Compounds with No Internal Relations

(1) kaihatu-kaigi
development-meeting
'a meeting regarding the development' or 'a meeting for development'

(2) kyousei-sikkou
compulsion-execute
'compulsory execution' or 'execute compulsorily'

These compounds sometimes have multiple meanings depending on the context as in Example 23 (1). If one were to paraphrase (2), the first constituent becomes an adverbial *kyousei-teki-ni* (*teki* 'ADV-making suffix'; *-ni* 'suffix that links an adverbial with a verb, an adverb, or adjective). The structure of the paraphrased compounds in Example 23 will be the following, illustrated in Figure 4.

Figure 4: Structure of the Compounds in Example 23



If we assume that the compounds in Example 23 are derived from their paraphrased counterparts in Figure 4, it may be possible to hypothesize that the preposition in (1) and the adverb-making suffix in (2) are deleted, and the N is incorporated into the V. However, what enforces the movement is not clear.

The compounds look as if the non-head constituent has an adjunct relation to the head. Similar relations are commonly found in compounds with a regular noun head, which will be explained in 4.5.2, that the constituent relation is always attributive. The difference between the compounds in question and regular noun compounds is that apart from the part of speech of the head, the non-head constituent cannot take the genitive case when the compounds in question are paraphrased, as illustrated in Example 24, whereas regular noun compounds can always be paraphrased with the genitive case attaching to the non-head constituent. In other words, the modifying constituent is adjectival. Regular noun compounds, as will be detailed in 4.5.2, have an attributive relation between the constituents. Since the deverbal compounds with no argument relation do not behave like regular noun compounds, and the non-head constituent behaves adverbially when the compound is paraphrased, the constituent relation seems to be adjunct.

Example 24: Inability to Take the Genitive Case in Compounds with a Regular Noun Head

(1) *kaihatu-no kaigi
development-GEN meeting

(2) *kyousei-no sikkou
compulsory-GEN execution

I have stated that incorporation is not likely. Also, the first constituent functions as an adverbial to the head. Regarding the locus of compounding, more evidence is necessary to claim whether the compounding occurs at the syntactic level or at the lexical level. Alternatively, it is possible to propose that deverbal nouns are syntactically ambiguous such that it can be inserted in a V slot and a N slot as suggested by Manning (1993). However, in this study, I assumed that deverbal nouns are derived from verbs. Therefore, I reject the analysis that deverbal nouns are syntactically ambiguous. A further study is necessary to determine at which level the compound is formed.

4.5.1.4 V+V Compounds

Section 4.5.1.2 discussed compounds with a deverbal head whose non-head constituent functions as a noun. Non-head constituents can be deverbal, regular, or de-adjectival. In the case in which the non-head constituent is deverbal, I assume as discussed in 4.5.1.2 that these deverbal nouns are originally verbs, but undergo nominalization when functioning as a noun. The majority of the deverbal compounds in the present data show that the non-head constituent functions as a noun. However, there are thirteen out of 400 compounds in which both of the constituents can function as a verb as shown in Example 25.

Example 25: Both Constituents Functioning as Verbs

(1) sindan tiryou
diagnose treat
'diagnose and treat'

(2) shoukyaku shobun
burn dispose
'burn and dispose'

(3) seiri tougou
readjust unify
'readjust and unify'

(4) hatumei hakken
invent discover
'invent and discover'

(5) shuryou saishuu
hunt collect
'hunt and gather'

(6) zoushuu zoueki
increase-of-income increase-of-profit
'increase of proceeds'

The above examples (1) to (3) express two successive events, and the order of the event is reflected by the surface linear order of the compounds. For instance, in (1), a doctor diagnoses a patient, and treats him/her, and the order is never the other way around. Therefore, **tiryou sindan* 'treat diagnose' is not permitted. These examples resemble V + V compounds which are commonly found in Japanese. The properties and the structures of the examples (1) to (3) are compared against those of V + V compounds in 4.5.1.4.

The examples (4) to (6) do not show any order of events, but rather two pragmatically similar entities are compounded although it is hard to measure how pragmatically similar these entities are. What we can say about these compounds is that the two entities have equal weight, and frequently co-occur. The combination of the

constituents is relatively fixed. Therefore, it is possible to hypothesize that these compounds are lexicalized in the Japanese language. However, the possibility of the lexicalization process needs to be investigated. In any case, because the constituents of these compounds have equal weight, and do not show any temporal sequence, I consider these as dvandva compounds.

4.5.1.5 Structure of Coordinated V + V Compounds

Japanese has a large stock of V + V compounds. Some examples are shown in Example 26. The examples are taken from Nishiyama (1998), originally taken from Kageyama (1993) and Li (1993).

Example 26: V + V Compounds (Nishiyama 1998:175)

(1) John-ga Bill-o **osi-taosi**-ta
-NOM -ACC push-topple-PAST
'John pushed Bill down.'

(2) John-ga niwatori-o **naguri-korosi**-ta
-NOM chicken-ACC beat-kill-PAST
'John beat and killed a chicken.'

In Example 26 (1) and (2), the compound verbs are in boldface. These verbs are, as the gloss shows, composed of two independent verbs, and the temporal sequence is reflected in the linear order of the constituents. Nishiyama (1998) compares the properties of these compounds with those of serial verbs found in West African languages and Caribbean creoles, and proposes that Japanese V + V compounds and serial verbs in those languages have the same underlying syntactic structure.

It is generally assumed that the definition of serial verb construction is as follows:

(3) Definition of Serialization

> Serialization refers to a phenomenon where a sentence contains a succession of verbs and their complements (if any) with one subject and one tense value that are not separate by any overt marker of coordination or subordination (Nishiyama 1998: 176, adapted from Collins (1997: 4)).

Another common phenomenon in the serial verb construction is that the internal argument is shared by two verbs (Collins, 1997). The V+V compounds in Example 25 (1) to (3) meet the description of serial verb construction in that the constituents of these compounds express two successive events, and the order of the events is the order of the constituents. Also, the internal argument is shared by two verbs as shown in Example 27.

Example 27: Internal Argument Sharing

(1) isha-ga kanja-o sindan-tiryou sita
doctor-NOM patient-ACC diagnose-treat do:PAST
'(The) doctor diagnosed (the) patient and treated him/her.'

(2) seihu-ga usi-o shoukyaku-shobun sita
government-NOM cow-ACC burn-dispose do:PAST
'(The) government burned up cows and disposed of them.

Taking into account the properties of these compounds, it is possible to view these compounds as serial verbs. However, in the following, I will show that these are not serial verbs, but rather coordination.

The compounds in Example 25 (1) to (3) can be paraphrased by attaching a tenseless dummy verb on the first constituent or attaching a dummy verb and a temporal adverb as shown in Example 28.

Example 28: Paraphrase of V + V Compounds

(1)a. isha-ga kanja-o sindan-chiryou sita
doctor-NOM patient-ACC diagnose-treat do:PAST
'(the) doctor diagnosed and treated (the) patient.

b. isha-ga kanja-o sindan si chiryou sita
doctor-NOM patient-ACC diagnose do treat do:PAST
'(the) doctor diagnosed and treated (the) patient.'

c. isha-ga kanja-o sindan site kara chiryou sita
doctor-NOM patient-ACC diagnose do:and then treat do:PAST
'(the) doctor diagnosed (the) patient, and then treated him/her."

(2)a. seihu-ga usi-o shoukyaku-shobun sita
government-NOM cow-ACC burn-dispose do:PAST
'(the) government burned and disposed of cows.'

b. seihu-ga usi-o shoukyaku si shobun sita
government-NOM cow-ACC burn do dispose do:PAST
'(the) government burned and disposed of cows.'

c. seihu-ga usi-o shoukyaku site kara shobun sita
government-NOM cow-ACC burn do:and then dispose do:PAST
'(the) government burned cows and then disposed of them.'

Since we assume that the UTAH holds in Japanese, I consider these sentences to be derived from the same syntactic structure. Both constituents of the compound can take a dummy verb, and these constituents can optionally be joined with a conjunctive. As Example 29 shows, the two different adverbs can modify the verb phrases *sindan suru* 'diagnose do' and *tiryou suru* 'treat do', it can be hypothesized that underlyingly, the constituents of V + V compounds are under VPs conjoined with a conjunction as illustrated in Figure 5.

There are two possibilities where the conjunction phrase is inserted; one possible position is immediately above two VPs as shown in Figure 5, and the other is that the conjunction phrase coordinates two V's. In the former, the deletion of the NP subject and NP object in the second VP can give the appearance of a compound. In the latter, the deletion of the NP object in the second VP can give the appearance of a compound. I propose that the latter is correct because when the compound appears as a noun phrase rather than two verb phrases as shown in Example 30, the theme of these two constituents can modify the compound by taking the genitive case. However, the subject cannot modify the compound. Therefore, I suggest that when V + V compounds are a result of being conjoined at the V' level, the second NP object is deleted. When the compound is

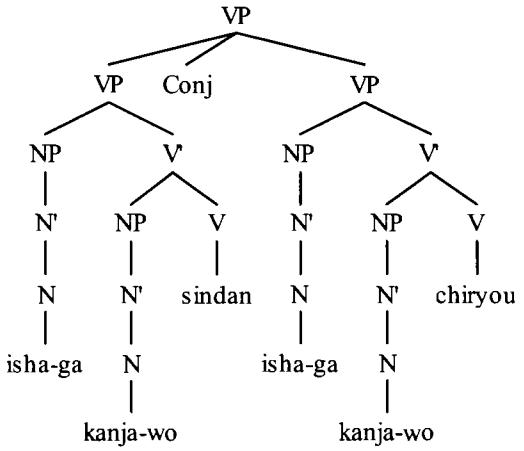
functioning as a noun phrase, as in Example 30, the whole V' can undergo nominalization, and become a noun compound.

Example 29: Insertion of Adverbs

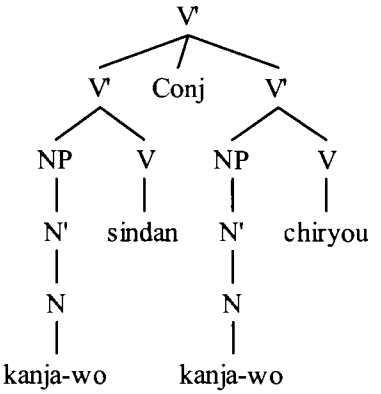
isha-ga	kanja-o	kinou	sindan	site	kyou	chiryou	sita
doctor-NOM	patient-ACC	yesterday	diagnose	do:and	today	treat	do:PAST

Figure 5: Structure of V + V compounds and their Paraphrase

(1)



(2)

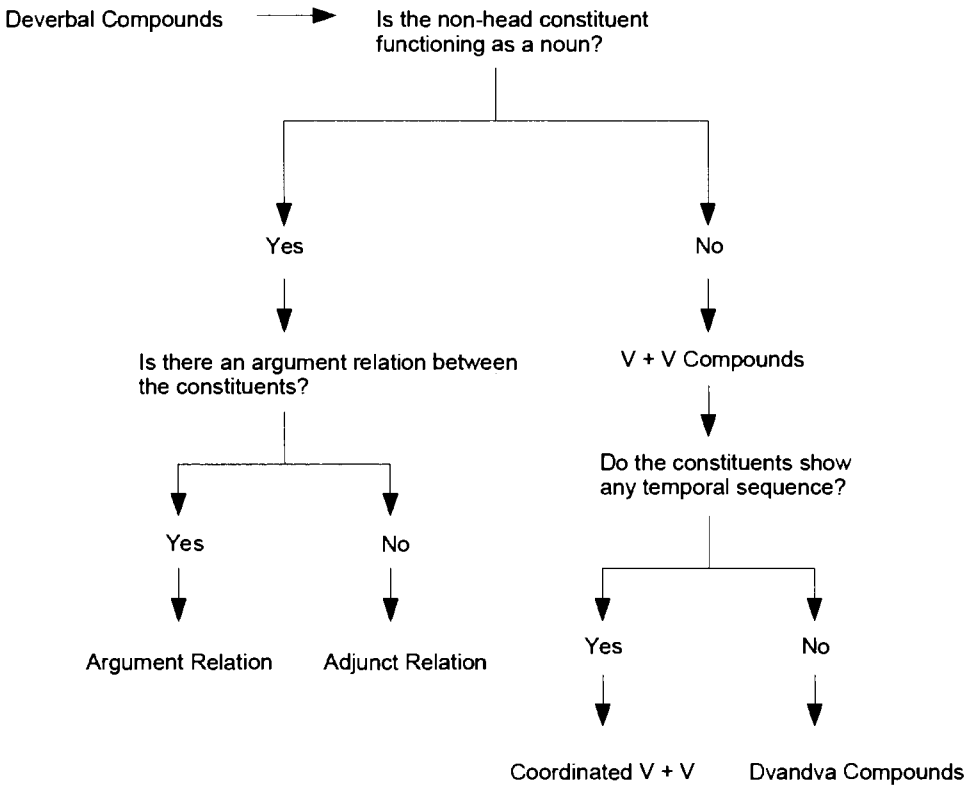


Example 30: Nominalization of V + V Compounds

(kanja-no) sindan-tiryou-ga okonaware-ta
(patient-GEN) diagnose-treat-NOM take.place-PAST
'The diagnosis and the treatment of the patient took place.'

Section 4.5.1 discussed the classification of deverbal compounds. A summary of the classification of deverbal compounds is provided in Figure 6.

Figure 6: Classification of Deverbal Compounds



4.5.2 Structure of Compounds with a Regular Noun Head

As discussed in the previous section, compounds with a deverbal head can be classified into four groups, one with an internal relation between the non-head constituent and the head, and one that does not hold any internal relation, and V + V compounds. Within V + V compounds, there are dvandva compounds and coordinated V + V compounds. Compounds with a regular noun head do not show any internal relation as do some deverbal compounds. Some examples of regular noun compounds are shown in Example 31. The first constituent of Example 31 (1) is a regular noun, that of (2) is a deverbal noun, and lastly, the first constituent in (3) is a de-adjectival noun. As Example 31 shows, the first constituent does not have any argument-structure relation to the head, but rather it is merely modifying the head noun regardless of its part of speech. In other words, the non-head constituent only has an attributive function. These compounds are so-called root compounds. Root compounds are ones whose second stem is not derived from verbs.

Example 31: Examples of Compounds with a Regular Noun Head

(1) zinkou-seisaku
population-policy
'population policy'

(2) housou-bunya
broadcasting-field
'the field of broadcasting'

(3) hituyou-zyouken
necessary-condition
'necessary condition'

One property of these compounds is that when paraphrased, the first constituent takes the genitive case *–no* if it is a deverbal noun or a regular noun, the suffix *–na* if it is de-adjectival as shown in Example 32.

Example 32: Paraphrase of Compounds with a Regular Noun Head

(1) zinkou-no seiksaku
population-GEN policy
'population policy'

(2) housou-no bunya
broadcasting-GEN field
'the field of broadcasting'

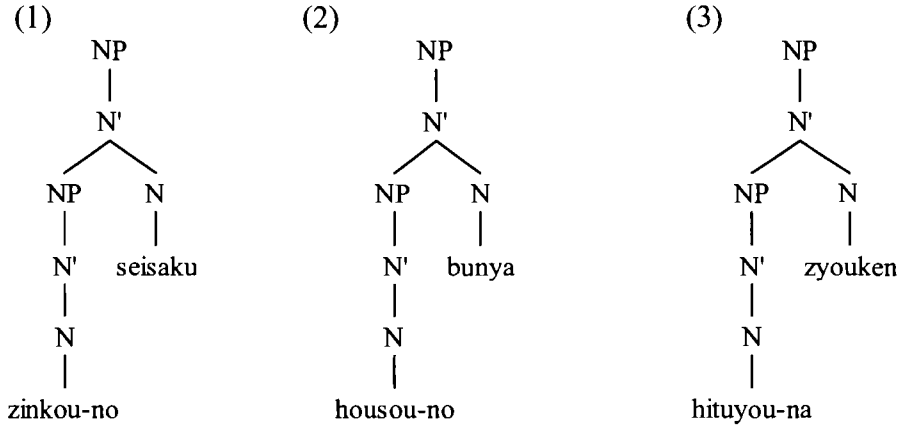
(3) hituyou-na zouken
necessary-ADJ condition
'necessary condition'

In both compounds and their paraphrased counterparts, the first constituent modifies the head, regardless of the part of speech of the first constituent. I suggest that the underlying structure of the paraphrase of the compound is the following, illustrated in Figure 7. It is possible to hypothesize that in compounds with a regular noun head, speakers have an option of dropping the case marker, giving the structure the appearance of a compound.

Alternatively, it is also possible to analyze that these compounds are formed at the lexical level. The locus in which the compounding takes place needs to be further investigated.

In Figure 7, the deverbal noun in (2) and the de-adjectival noun in (3) are treated as nouns because, first, they are modifying a regular noun, and second, there is no thematic relation between the head and the deverbal or de-adjectival head. In this study, it is assumed that deverbal and de-adjectival nouns undergo nominalization prior to the compounding operation.

Figure 7: Underlying Representation of Compounds with a Regular Noun Head



4.5.3 Structure of Compounds with a De-adjectival Head

Compounds with a de-adjectival head are rare. However, there are some in the Utiyama Corpus. As Table 12 shows, there are 21 compounds whose non-head constituent is a deverbal noun, and 22 compounds whose non-head constituent is a regular noun. Compounds whose non-head constituent and the head are both de-adjectival noun are extremely rare. There are only two instances of such compound in the present dataset collected from the Utiyama Corpus.

Regarding the structure of de-adjectival compounds whose non-head constituent is a deverbal noun, one notices that the head functions as the predicate of the first constituent. For instance, if one were to paraphrase the compounds in Example 33, the

head is the predicate adjectival noun of the non-head constituent as shown in Example 34.

Example 33: De-adjectival Compounds whose First Constituent is Deverbal

(1) zizoku kanou
sustain possible
'sustainable' or 'possible to sustain'

(2) teigi konnan
define difficulty
'hard to define' or 'difficult to define'

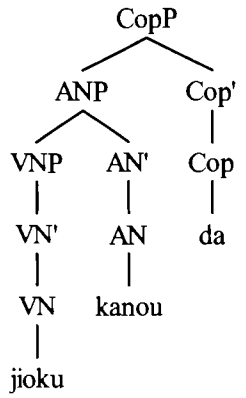
Example 34: Paraphrase of De-adjectival Compounds in Example 33

(1) zizoku-ga kanou da
sustaining-NOM possible COP
'sustaining is possible'

(2) teigi-ga (totemo) konnan da
defining-NOM (very) hard COP
'defining is (very) hard'

Based on the paraphrase of de-adjectival compounds, it is possible to hypothesize that underlyingly the first constituent occupies the subject position, and the head occupies the adjectival phrase, taking the copula, as shown in Figure 8. The compounds in Example 33 cannot take another element while their paraphrase can take an adverbial phrase before the second de-adjectival as shown in Example 34. Therefore, it is possible that when it appears as a compound, the noun phrase that is occupying the subject position is incorporated to the head of the second adjectival phrase. However, an alternative analysis would be that the compounds are produced at the lexical level.

Figure 8: Underlying Structure of De-adjectival Compounds (The non-head is deverbal)



There is a group of de-adjectival compounds in which the non-head constituent acts as the theme of the de-adjectival head. The head of these compounds are emotion or perception words as shown in Example 35.

Example 35: De-adjectival Compounds with Emotion or Perception Words

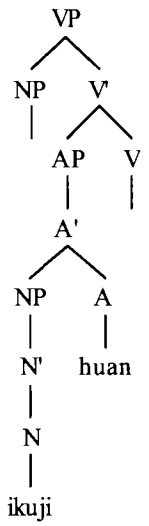
(1) ikuzi huan
raising a child anxiety or anxious
'anxious to raise a child'

(2) yokkyu human
desire dissatisfaction or dissatisfied
'frustration'

In the above examples, the heads are emotion words. These emotion words usually imply an experiencer and the situation or thing the experiencer has a feeling towards. Therefore, these emotion words have an argument structure; it requires an experiencer and a theme, which is the thing towards which the experiencer has a feeling. The analysis of the data reveals that the non-head constituent is normally the theme of the emotion head. Thus, it is possible to hypothesize that the non-head constituent, which can function as the theme of the emotion word, underlyingly originates as the sister of the emotion word as illustrated in Figure 9. It is worth mentioning that when the emotion word in Figure 9 is taking a copula, the theme and the emotion word do not appear as a compound, but rather

it is more common to insert the subject marker on the first constituent, and the topic marker on the experiencer. When the theme of the emotion word and the emotion word both appear as a compound, the whole compound can function as a noun phrase by taking a case marker such as the subject marker as shown in Example 36. In such case, the compound cannot take any additional elements. Therefore, I propose that the first constituent is incorporated to the head to form a compound, and the whole compound undergoes nominalization, as do deverbal nouns, to become a compound noun.

Figure 9: Structure of Compounds with Emotion Words



Example 36: Nominalization of De-adjectival Compounds

(1) ikuzi-huan-ga tuyoi
rasing:a:child-uneasy-NOM strong
'very anxious about raising a child'

(2) yokkyuu-human-ga bakuhatu-suru
desire-dissatisfied-NOM explode-do
'frustration explodes'

Regarding de-adjectival compounds whose non-head constituent is a regular noun, the paraphrase of these compounds can always take the genitive case as shown in Example

37. This suggests that these compounds have a similar structure to regular noun compounds. I propose that the structure of these compounds is the same as regular noun compounds.

Example 37: Paraphrase of De-adjectival Compounds (the non-head is a regular noun)

(1)a. sekai heiwa
world peace (peaceful)
'world peace'

b. sekai-no heiwa
world-GEN peace
'world peace'

(2)a. shourai huan
future anxiety (anxious)
'anxiety about one's future'

b. shourai-no huan
future-GEN anxiety (anxious)
'anxiety about one's future'

(3)a. naizyu husin
domestic demand inactivity (inactive)
'inactive domestic demand'

b. naizyu-no husin
domestic demand-GEN inactivity (inactive)
'inactive domestic demand'

In the present data, there are only two instances whose constituents are both de-adjectival, shown in Example 38. In Example 38 (1), the first constituent specifies the subcategory of the head. Therefore, its function is attributive. Example 38 (2) is an instance of coordination because the compound express an event or situation that is both harmful and futile. It could also be analyzed as a dvandva compound because the two constituents

seem to have equal weight, and frequently co-occur. In any case, because there are only two instances, and these two behave differently from each other, the identification of their structure is left for a future study.

Example 38: Examples of Compounds whose Constituents are Both De-Adjectival

(1) titeki ziyuu
intellectual freedom
'intellectual freedom'

(2) yugai mueki
harmful futile
'harmful and futile'

4.6 Summary

This chapter has introduced a classification system of Sino-Japanese compounds. Sino-Japanese nouns usually consist of two characters, each of which represents one morpheme. Although these morphemes have some semantic meaning, they are bound morphemes, and have to combine with another morpheme to express a concept. Two-character Sino-Japanese nouns are viewed as compounds by Shibatani (1990), Kageyama (1982) and Tsujimura (1996). However, they are considered simple nouns in this study, as ChaSen, a Japanese morphological analyzer, segments two-character Sino-Japanese nouns as one content word containing one semantic meaning.

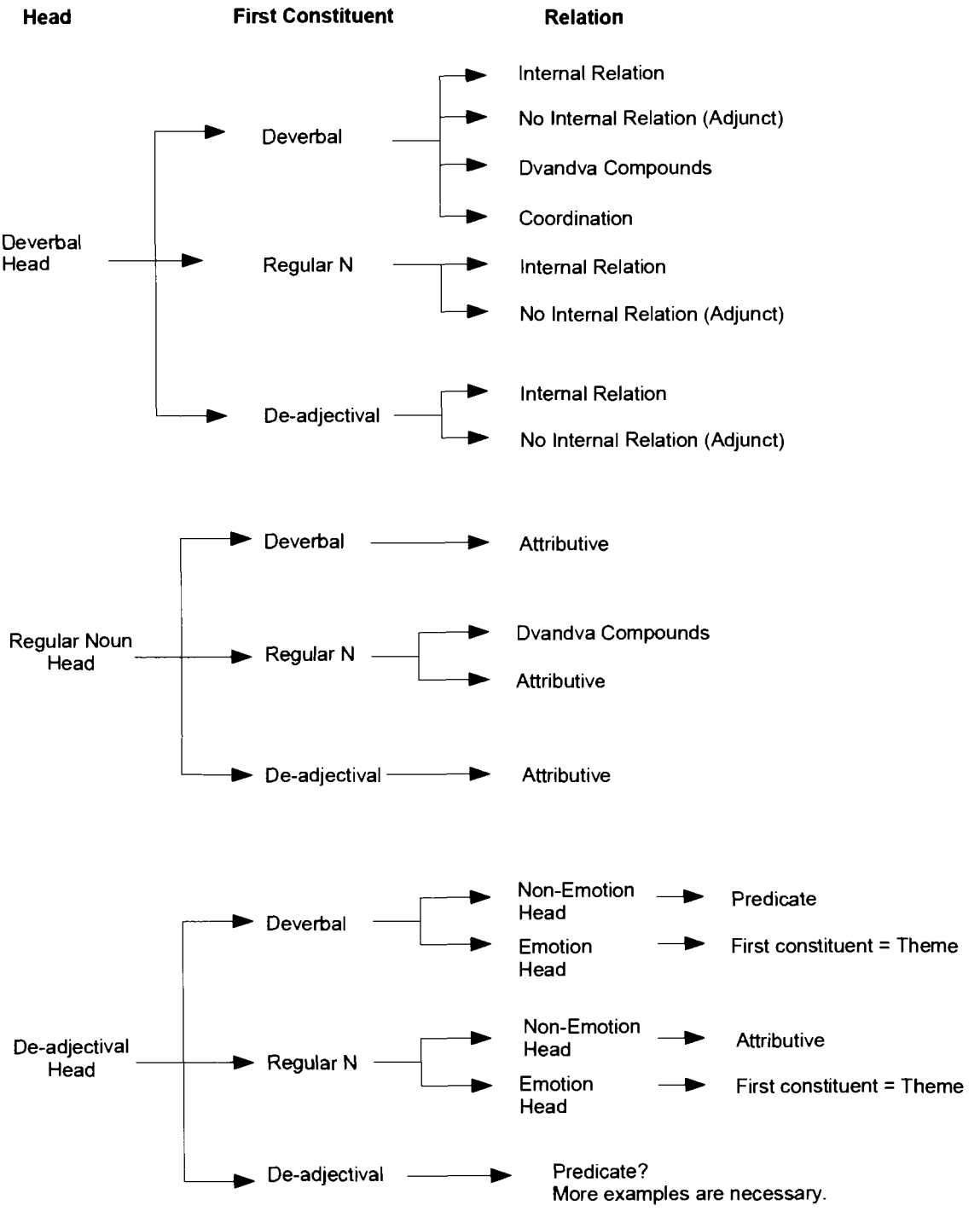
Sino-Japanese compounds are classified according to the part of speech of the head, which can be deverbal noun, regular noun, or de-adjectival noun. This classification is justified by the structure of Sino-Japanese compounds. The summary of the present classification is provided in Figure 10. The non-head constituent of some deverbal compounds holds an argument relation to the head. These are considered syntactic compounding because of the syntactic structural resemblance to noun incorporation. Regarding compounds whose non-head constituent has an adjunct relation to the head, the non-head element may be base-generated at the level of syntax or the compounding

occurs at the lexical level. Some deverbal compounds are V + V compounds. Among these V + V compounds, dvandva compounds show that the constituents frequently co-occur, have equal weight, but have no temporal sequence while coordinations do have a temporal sequence. Coordinations rather express two successive events, the order of which is reflected by the order of the constituents.

Compounds with a regular noun head only have an attributive function to the head. As discussed earlier, the locus of the compounding can either be at the syntactic or at the lexical level. It needs to be further investigated. Theoretically, dvandva compounds are possible if both constituents are regular nouns. However, regular noun dvandva compounds are not found in the present dataset.

De-adjectival compounds behave differently depending on the part of speech of the first constituent. For the majority of de-adjectival compounds with a deverbal non-head constituent, the head functions as the predicate adjectival of the first constituent. However, when the head is an emotion or a perception word, the non-head constituent acts as the theme of the head. De-adjectival compounds whose first constituent is a regular noun behave similarly to regular noun compounds because the first constituent takes the genitive case when they are paraphrased, and its sole function is to specify the subclass of the head.

Figure 10: Classification Summary



CHAPTER 5:LEXICAL SEMANTIC ANALYSIS OF SINO-JAPANESE COMPOUNDS

5.1 Introduction

Sino-Japanese compounds have been studied extensively in Japanese Linguistics. However, the previous studies focus on the morphological structure of Sino-Japanese compounds, and little has been done on their lexical semantic structure. Further, the previous studies focus on one type of Sino-Japanese compounds, and there has not been a unified analysis of all types of Sino-Japanese compounds. This chapter presents an analysis for each type of compound classified in Chapter 4 using Lexical Semantics. This study assumes that, as Selkirk (1982) argues, noun compounds can be generated by a context free grammar. A context free grammar is a formal grammar in which the desired output is generated by the rule A → B C, or A → α, B and C expressing non-terminal symbols and α representing the terminal symbol. The rule A → α states that A becomes α whenever A occurs. In the present study, it is assumed that Sino-Japanese compounds can be analyzed and translated into English by context free grammar. Also, this study assumes that word meaning and the internal structure of a word can be represented by primitives and structural templates as assumed by many Lexical Semanticists (Jackendoff, 1972, 1983, 1987a, 1990; Lieber, 2004; Pustejovsky, 1992; Szymanek, 1988; Wierzbicka, 1972, 1980, 1985, 1988, 1996). As briefly mentioned in 3.3.1, there are a number of lexical semantic frameworks one can choose from. Jackendoff's Lexical Conceptual Structure is concerned largely with verb meanings, and it may not be suitable for nouns and adjectives. Takeuchi et al. (2003a, 2003b, 2001) use Jackendoff's lexical semantic framework (1990) to analyze the structure of deverbal Sino-Japanese compounds. Takeuchi et al. identify twelve possible lexical semantic structures for verbs.[17] However, some of their structures are a slightly different version of others, and can be reduced to

[17] Takeuchi et al. (p.c. via e-mail) identified more structures, and now have 18 different structures at the time of writing.

one. Reducing the number of possible structures is important when programming because it possibly helps minimize the number of errors that could occur during the actual programming. Because Jackendoff's Lexical Conceptual Structure is not suitable for nouns, and Takeuchi et al. have not succeeded in analyzing deverbal Sino-Japanese compounds using Jackendoff's framework, it is not used in this study. Wierzbicka's system uses fifty-six features, which is too many for the purpose of the present study. Further, it is not so clear how her features can be applied to languages that have not been considered in her studies. Szymanek's framework largely focuses on the morpho-syntactic aspects of compounds. In the case of this study, the structure is quite clear. What this study focuses on is rather the internal semantic structure of these compounds. Therefore, I consider his framework inappropriate for the present study. In Pustejovsky's framework, syntax and lexical semantics are tightly intertwined, and lexical semantics alone cannot be isolated. This is not suitable in this study because Sino-Japanese compounds need to be analyzed independent of the syntactic context. In the present study, Lieber's framework is used to reveal the structure and the properties of Sino-Japanese compounds because her framework accounts for verbs as well as other parts of speech such as nouns, adjectives and prepositions. Another advantage of her framework is that there are only a limited number of primitives and these primitives are cross-categorial. Her framework was developed to solve the problems of English derivational morphology, compounding, and conversion. Since the present study deals with compounding, her framework may be best suited for Sino-Japanese compound analysis. A brief description of Lieber's Lexical Semantic classification system is provided in 5.1.1. Section 5.2 provides the analysis of the compounds whose head is a deverbal noun, and Section 5.3 presents the analysis of the compounds whose head is a regular noun. The analysis of compounds with a de-adjectival head is provided in Section 5.4.

5.1.1 Lieber's Lexical Semantics

As Figure 11 and Figure 12 show, Lieber's basic framework consists of two semantic categories, SITUATIONS and SUBSTANCES/THINGS/ESSENCES. The former accounts for verbs and adjectives, and the latter accounts for nouns. With another set of features, Lieber also presents a taxonomy for prepositions which are not introduced in this section.

As Figure 11 shows, if the situation expressed by a word is dynamic, the feature [+dynamic] is assigned, and if it is not dynamic, [-dynamic] is assigned to the word. [-dynamic] verbs are stative. Adjectives are also characterized as [-dynamic] because Lieber claims that adjectives are conceptually identical to stative verbs although adjectives and stative verbs are syntactically different. [+dynamic] verbs are further classified as shown in the figure. If an event is a simple activity, only [+dynamic] is assigned to the word. For example, the verb *to kiss* is a simple dynamic activity. If a dynamic event involves changing from one state to another, the feature [+/- IEPS], which stands for Inferable Eventual Position or State, is used to distinguish words that are unaccusative or inchoative and ones that express the manner of change. The feature [+/- IEPS] accounts for the change of state from one point to another. For instance, the verb *descend* is [+dynamic, +IEPS] while the verb *walk* is [+dynamic, -IEPS] (Lieber, 2004).

Figure 12 shows the classification of nouns. Lieber states that major categories of nouns are concrete nouns and abstract nouns. These are distinguished by the feature [material]. [+material] is used for concrete nouns, and [-material] for abstract nouns. These two categories are further classified by the use of feature [dynamic]. When this feature is used for [+material] nouns, it indicates that there is a specific activity that is associated with the noun. For instance, *author* implies writing, *chef* implies cooking, and *mother* implies parenting. When the [dynamic] feature is used for [-material] nouns, it signals that the nouns are eventive. One may argue that some words are ambiguous between [+material] and [-material]. for the purpose of computation, these words are considered as having two senses, one which can be represented as, for example, [+material], and the other represented as [-material]. Words are represented using the features in Figure 11 and Figure 12, along with the number of arguments the word can take, represented by the number of empty square brackets. For instance, the verb *to know* is represented as [-dynamic ([ ], [ ])] because it is a stative, transitive verb.

Figure 11: Lieber's Set of Lexical Semantic Features for Verbs and Adjectives (Lieber 2004: 30)

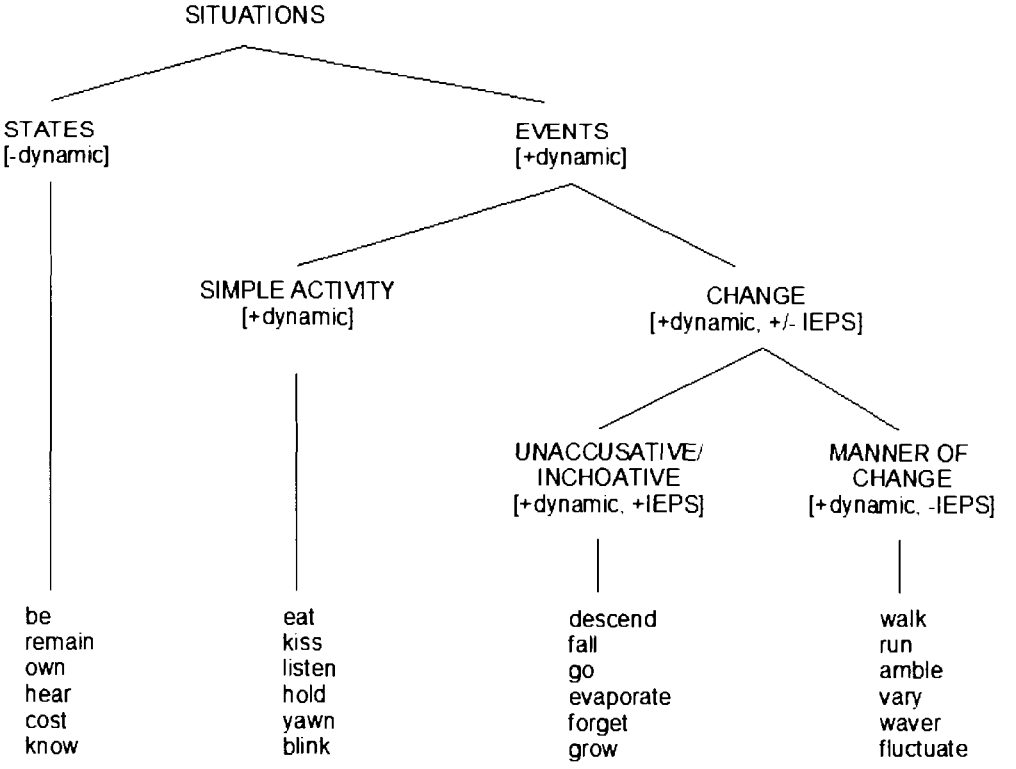
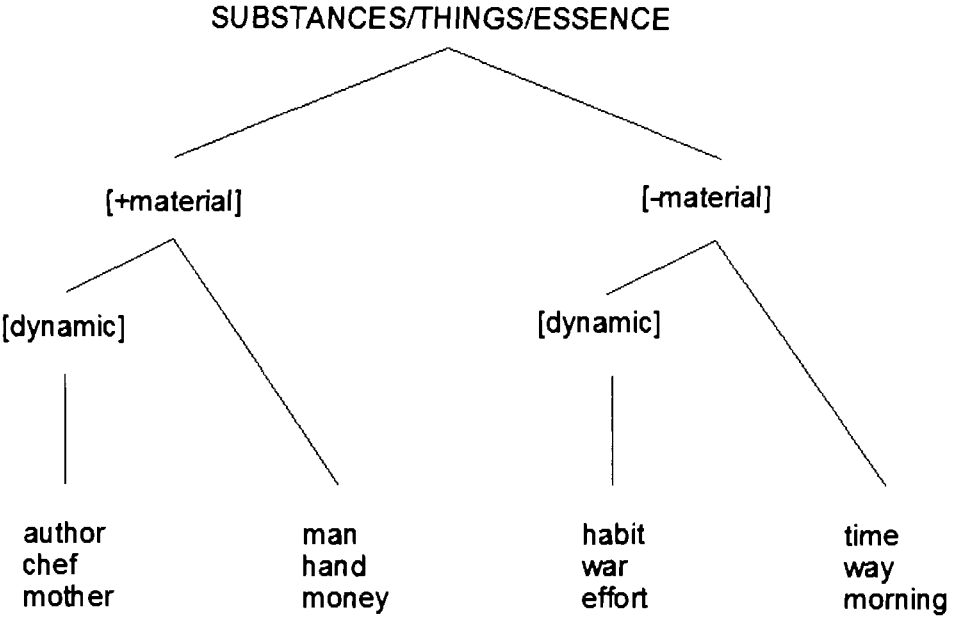


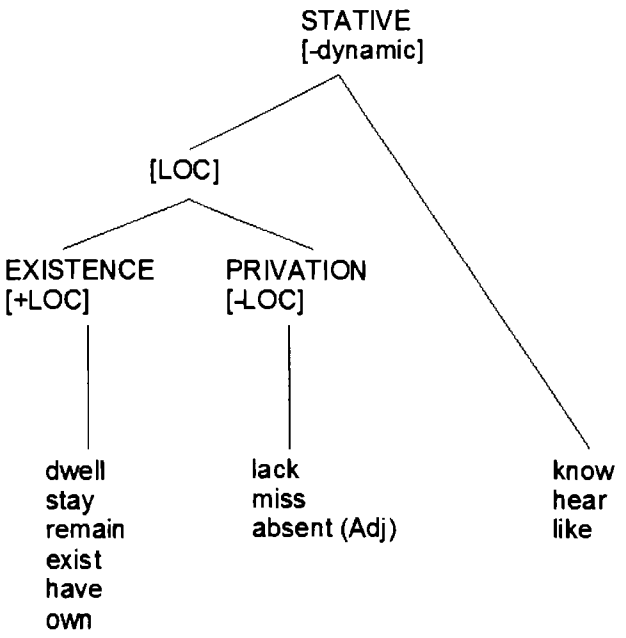
Figure 12: Lieber's Set of Lexical Semantic Features for Nouns (Lieber 2004: 27)



Lieber uses additional features [LOC], [B], and [CI] to refine her framework. Lieber (2004: 99) states that the feature [LOC] "asserts the relevance of place or position in the conceptual make-up of a lexical item". The difference between [IEPS] and [LOC] is that [IEPS] "adds a PATH component of meaning in a semantic skeleton" (Lieber, 2004: 99) while [LOC] does not. Lieber demonstrates the use of [LOC] in stative verbs. For instance, some stative verbs such as *dwell*, *stay*, *remain*, and *exist*, are, as Lieber calls, "verbs of existence". These verbs are specified as [+LOC] as shown in Figure 13.

The features [B] and [CI] are proposed to characterize the notion of quantity, duration, internal individuation, and boundaries, that are encoded in the lexicon. The feature [B], which stands for "bounded", asserts spatial and temporal boundaries that are manifested within the lexical item. The feature [CI], which stands for "composed of individuals", asserts the relevance of spatial or temporal units manifested in the meaning of a lexical item. These two features are relevant to both the category of SITUATION and SUBSTANCE/THINGS/ESSENCE. Lieber demonstrates the use of the feature [B] and [CI] in the noun category, shown in Example 39.

Figure 13: Use of the Feature [LOC] (Lieber 2004: 100)



Example 39: Application of Quantitative Features to SUBSTANCE/THINGS/ESSENCES (Lieber 2004: 137)

[+B, -CI]	Singular Count Nouns	*person, pig, fact*
[-B, -CI]	Mass Nouns	*furniture, water*
[+B, +CI]	Group Nouns	*committee, herd*
[-B, +CI]	Plural Nouns	*cattle, sheep*

When the features [B] and [CI] are applied to the category of SUBSTANCES/THINGS/ESSENCES, these features distinguish mass and count nouns if the noun is singular. Plural count nouns consist of individuals, and do not have intrinsic boundaries. Therefore, they are [+B, +CI]. Mass nouns that are [+B] are composed of individuals, but have intrinsic boundaries.

When the features [B] and [CI] are used in the category of SITUATION, these features characterize the quantitative and temporal aspects such as durative/instantaneous or repetition of action in the meaning of a lexical item, which is closely related to the issue of telicity. Comrie (1976: 4) defines a telic situation as "one that involves a process that leads up to a well-defined temporal point". Pustejovsky (1992) points out that the telicity of a verb may change depending on where it occurs in a sentence. Example 40 shows that the verb *walk* is atelic or telic depending on the adverbial phrase that modifies the verb.

Example 40: Telicity (Pustejovsky 1993: 49)

(1) Mary walked.

(2) Mary walked to the store.

(3) Mary walked for 30 minutes.

Because the telicity of a verb is influenced by factors other than the verb itself, Lieber considers telicity to be "an aspectual property which is not purely or even primarily lexical" (Lieber, 2004: 141). Nonetheless, durative/instantaneous and repetitive aspects are in fact encoded in the lexicon as shown in Example 41.

Example 41: Durative/Instantaneous and Repetitive Aspects of Verbs (Lieber 2004: 138)

a. *The train arrived for an hour.
b. *The bomb exploded for an hour.
c. The prisoner tapped for an hour. (iterative reading)
d. The student sneezed for an hour. (iterative reading)
e. We walked for an hour.
f. They studied the map for an hour.

As can be seen in Example 41, some verbs are felicitous with durative adverbs while some are not. Lieber considers ones that can take durative adverbs as durative, and ones that cannot as instantaneous. These verbs can be distinguished by the use of quantitative features [B] as shown in Example 42. As shown in Example 41 c. and d., some verbs express repeated actions. The repetitive aspect can be characterized by the use of [+CI] illustrated in Example 42.

Example 42: Application of Quantitative Features to SITUATIONS (Lieber 2004: 139)

[+B, -CI]	Non-repetitive Punctuals	*explode, jump, flash*
[-B, -CI]	Non-repetitive Duratives	*descend, walk, draw*
[+B, +CI]	<logically impossible>	
[-B, +CI]	Repetitive Duratives	*totter, pummel, wiggle*

Pustejovsky points out that the same verb can express different aspects depending on how the verb is used as shown in Example 43. The verb *close* in Example 43 (1) has a stative meaning. The same verb in (2) is intransitive, expressing an event in which the door closed by itself while in (3), the verb is used transitively conveying an event in which John closed the door such that the door is closed. As in English, there are many words in Japanese that can express different aspects depending on the context in which they occur. Lieber does not explicitly mention this issue. However, these words are considered to have multiple senses, each of which denotes one meaning that is related but slightly different from one another in this study.

Apart from the features used in Lieber's framework, I also used [animate] because this feature is necessary to determine the relation between the constituents of the compound. This feature turns out to be useful in distinguishing deverbal nouns that only take an animate noun as their argument.

Example 43: Different Senses of the Verb *close* (Pustejovsky 1992: 53, originally taken from Lakoff (1970)

(1) The door is closed.

(2) The door closed.

(3) John closed the door.

5.2 Procedure of the Analysis

5.2.1 Compounds with a Deverbal Head

Compounds whose head is identified as deverbal by ChaSen belong to this group. There are four types of deverbal compounds: ones with an internal relation, ones that possess an adjunct relation, coordination, and dvandva compounds. First, dvandva compounds and coordination must be isolated from others. The part of speech of the constituents are the same, and both the non-head constituent and the head can function as verbs in dvandva and coordination compounds while it is not the case for others; the non-head constituent always functions as a noun. Some disambiguation methods have been proposed for serial verb-like constructions (see 4.5.1.3) (Uchiyama et al., 2005; Uchiyama & Ishizaki, 2003). However, no disambiguation method for V + V compounds has yet been proposed. V + V compounds are not frequent; they constituted only 3.3% of the deverbal compounds in the present data, as will be shown in Table 15. Therefore, isolating V + V compounds from other deverbal compounds is quite trivial. Nonetheless, it is necessary in order to get the correct translation because when they are translated in English, the two constituents are frequently joined with a conjunctive *and* while it is not the case for non-V + V compounds.

As discussed in 4.5.1.2, some deverbal compounds have an argument relation between the constituents while others do not. In order to find what distinguishes these two types of compounds, the following steps are taken. First, the transitivity of the head is identified. As discussed earlier, the First Sister Principle is assumed to hold in Japanese. If there is an internal relation, the transitive deverbal head is expected to take its theme when forming a compound, and the intransitive deverbal noun is expected to take its agent[18]. Second, the lexical semantic structure of the first constituent and that of the default arguments of the deverbal nouns are identified to see if there is any difference between compounds that have an internal relation and ones that have an adjunct relation.

5.2.2 Compounds with a Regular Noun Head

Compounds whose head is identified as a regular noun by ChaSen (see 4.3.2) belong to this type. Because the non-head constituent only has an attributive function regardless of its part of speech as discussed in 4.5.2, all the compounds with a regular noun head follow the same procedure. In order to analyze the relation of the modifier to the head, the lexical semantic structure of the modifier and the head is identified.

5.2.3 Compounds with a De-adjectival Head

Compounds whose head is identified as de-adjectival belong to this group. Because there are only two instances of compound whose constituents are both de-adjectival as shown in Example 38 in Chapter 4, this group is not analyzed in the present study because of lack of examples. Ones whose non-head constituent is a deverbal or regular noun are grouped into three, according to their constituent relation, emotion-theme, subject-predicate, and attributive. For ones whose head is an emotion word, the lexical semantic structure of the default theme of the emotion word and that of the non-head constituent of the compound is identified and compared. Regarding ones whose head functions as the predicate of the non-head constituent, the lexical semantic structure of the default subject of the head and that of the non-head constituent are identified and compared against each other. De-adjectival compounds that have an attributive relation behave like regular noun

[18] If the verb is unaccusative, the theme is incorporated.

compounds as mentioned in Chapter 4. Therefore, these compounds follow the procedures for regular noun compounds.

5.3 Analysis

5.3.1 Compounds with a Deverbal Head

A total of 921 deverbal compounds were collected from the Utiyama Corpus; 400 with a deverbal non-head constituent, 409 with a regular noun non-head constituent, and 112 with a de-adjectival non-head constituent, as shown in Table 14.

Table 14: Distribution of Deverbal Compounds

Part of Speech of the Non-head Constituent	Number
Deverbal	400
Regular noun	409
De-adjectival	112

Among deverbal compounds whose non-head constituent is a deverbal noun, there are 31 errors made by ChaSen. These compounds were identified as having a deverbal head by ChaSen, but the head was actually not a deverbal head. It was either a regular noun or a suffix. 233 compounds hold an internal relation between the constituents; 214 compounds have the first constituent acting as the theme or patient of the deverbal noun, 19 of them have the first constituent acting as the agent, and 121 of them have an adjunct relation. There are ten coordinated V + V compounds, and three dvandva compounds. There are two compounds whose non-head constituent can either function as the agent or the theme/patient of the deverbal noun. The summary of the distribution of these compounds is shown in Table 15.

Table 15: Deverbal Compounds whose First Constituent is Deverbal

Theme	214
Agent[19]	19
Adjunct	121
Theme & Agent	2
Coordination	10
Dvandva	3
ChaSen Errors	31
Total	400

Table 16 shows the detail of the deverbal compounds whose non-head constituent is a regular noun. The number of compounds collected is 409, among which 22 were errors made by ChaSen. As the table shows, 182 of them hold a theme relation, 49 of them function as the agent of the deverbal head, three of them can be either the theme or the agent. 153 compounds have the first constituent functioning as adjunct to the head.

Table 16: Deverbal Compounds whose First Constituent is a Regular Noun

Theme	182
Agent	49
Theme & Agent	3
Adjunct	153
ChaSen Errors	22
Total	409

Deverbal compounds whose non-head constituent is de-adjectival are relatively rare. Therefore, only 112 compounds were collected. The detail is shown in Table 17. As the table shows, there are 20 compounds whose non-head constituent functions as the theme of the deverbal head, and 88 compounds have an adjunct relation. The number of errors made by ChaSen is 4.

[19] 'Agent' here includes the theme of unaccusative verbs.

Table 17: Deverbal Compounds whose First Constituent is De-adjectival

Theme	20
Adjunct	88
Errors	4
Total	112

Regarding the transitivity of the heads[20], there are 587 transitive heads, 172 intransitive heads, and 88 can be either transitive or intransitive depending on the context. As mentioned earlier, theoretically, if the deverbal head is transitive, and the constituents have an internal relation, the first constituent has to be the theme or patient because of the First Sister Principle. If the head is intransitive, the first constituent is expected to be the agent unless we are dealing with unaccusative deverbal heads. If the head is unaccusative, its theme is expected to be the first constituent of the compound, provided that there is an internal relation between the constituents. All deverbal compounds that have an internal relation in the data behave as expected. In cases in which the head is either transitive or intransitive, the first constituent can be either the theme or the agent of the head. Because there are 88 deverbal heads that can be either transitive or intransitive in the present data, the transitivity of the compound head needs to be disambiguated. The heads in Example 44 can be either transitive or intransitive. It appears that the transitivity of these heads cannot be determined without any context. For instance, in Example 44 (1), the head can be transitive or intransitive shown in (2) and (3). When only the compounds are available without any context, the non-head constituent of the compound alone cannot seem to disambiguate the transitivity of the head. Transitivity disambiguation is beyond the focus of this study. However, one can disambiguate the transitivity of the deverbal head when taking into account the surrounding elements of the compound.

Example 44: Deverbal Heads that can be Transitive or Intransitive

(1) kigyou heisa
cooperation closure
'(a) winding-up'

[20] Here, dvandva compounds, coordination, and errors are excluded. Therefore, the total number is 847.

(2) kigyou-ga heisa-sita
cooperation-NOM closure-do:PAST
'(the) cooperation closed'

(3) shatyou-ga kigyou-o heisa-sita
the chief executive-NOM cooperation-ACC closure-do:PAST
'The chief executive closed (the) cooperation'

When compounds hold an internal relation, the transitivity of the head is a good indicator of the precise relation of the constituents. However, one of the goals of this section is to distinguish compounds with an internal relation from ones that have an adjunct relation. Transitivity does not seem to help distinguish these two types. In the next section, the lexical semantic properties of these compounds are analyzed to see if there is any element that can distinguish the two types of compounds.

5.3.1.1 Lexical Semantic Analysis of the Deverbal Compounds

The lexical semantic structure of the non-head constituent and that of the default complement of the deverbal head are identified in the present data. The default complement means the complement of the deverbal head when the deverbal head appears independently in a sentence in any context. The lexical semantic structure of the non-head constituent and the default complement of the deverbal head are analyzed because if they hold an internal relation, the lexical semantic properties of the first constituent should be compatible with those of the head's default complement. In other words, if the lexical semantic properties of the non-head constituent are not compatible with those of the head's default complement, it is possible that the non-head constituent modifies the head.

In the present data excluding errors, coordination, and dvandva compounds, there are 868 compounds. Among these, 853 compounds obey the hypothesis that if the lexical semantic properties of the first constituent match those of the default complement, the constituents hold an internal relation. If they do not, the relation is adjunct as shown in Example 45 and Example 46. Note that as shown in Figure 12, the [dynamic] feature in nouns was below the level of [+/- material]. Unless the head does not specify that the

complement must have the [dynamic] feature, [+/- material] subsumes nouns with the [dynamic] feature. For instance, the head of the last example in Example 45 requires a [-material] to be its complement. The lexical semantic structure of the first constituent is compatible with the head requirement because [-material] subsumes [-material, dynamic].

The head of the second example in Example 46, *katudou*, which is roughly translated as 'act' in English, requires a concrete noun. In English, the verb 'to act' can take an abstract noun. However, in Japanese, it only takes a concrete noun, often a concrete animate object, because according to the Daily Concise Japanese Dictionary (Satake, 2000), it means "actively move or work" (translated by the author). One may argue that it can take a noun that denotes an organization or an institution in Japanese. These normally consist of a group of people. Therefore, they are not considered an abstract noun, but rather a group of people.

Although the majority of the compounds comply with the hypothesis, there are fifteen compounds that did not. Example 47 shows an example. In the example, the head *sengen* 'declare' normally requires its theme to be an abstract concept. Since the first constituent *keizai* 'economy' is an abstract noun, it could potentially be the theme of the head. However, the constituent relation is adjunct because it is not that someone declares economy but rather declares something regarding the economy. It is worth mentioning that there is no compound that has an internal relation, but the lexical semantic requirement of the head is not met by the non-head constituent.

The analysis seems to suggest that Lexical Semantics can help distinguish the two types of compounds, ones in which the constituent relation is argument relation and the others in which the relation is adjunct. When the lexical semantic structure of the non-head constituent is compatible with that of the head's default complement, it is likely that the constituents have an argument relation while the constituents have an adjunct relation when the structure is not compatible.

Example 45: Lexical Semantic Properties of Compounds with an Internal Relation

Compound	LS of the head's default complement	LS of the modifier
kouzou-kaikaku structure-reform 'structural reform' or 'to reform the structure'	[-material ([ ])]]	[-material ([ ])]]
mondai-kaiketu issue-solve 'issue solving'	[-material ([ ])]]	[-material ([ ])]]
sishutu-keikaku spending-plan 'plan the spending'	[-material ([ ])]]	[-material, dynamic ([ ])]]

LS = Lexical semantic structure

Example 46: Lexical Semantic Properties of Compounds with No Internal Relation

Compound	LS of the head's default complement	LS of the modifier
souki-teiketu early-conclude 'early conclusion'	[-material ([ ])]	[-dynamic ([ ])]
keizai-katudou economy-act 'economic activities'	[+material ([ ])]	[-material ([ ])]

Example 47: Compounds that Do Not Comply with the Hypothesis

Compound	LS of the head's complement	LS of the modifier
keizai-sengen economy-declare 'economic declaration'	[-material ([ ])]	[-material ([ ])]

5.3.1.2 Lexical Semantic Analysis of V + V Compounds

Dvandva compounds contain two elements that are similar or identical in their semantic properties, and these elements carry the same weight. Therefore, it can be hypothesized that the lexical semantic structure of their constituents is the same or at least similar. On the other hand, the constituents of coordinated V + V compounds do not necessarily have to have an identical or similar lexical semantic structure because these constituents are

simply coordinated by a null conjunctive as discussed in 4.5.1.4 although the occurrence of the two successive events must be pragmatically possible. As expected, the constituents of the dvandva compounds in the present data have an identical lexical semantic structure, while those of coordinated V + V compounds may not, as shown in Example 48 and Example 49. It is interesting to note that a characteristic shared by the two types of V + V compounds is that the transitivity of the constituents is usually the same. Therefore, both constituents have the same argument structure. This is expected at least in coordinated V + V compounds because as shown in 4.5.1.4, coordinated V + V demonstrate internal argument sharing. If the argument structure of both constituents is different, it is impossible to have the internal argument sharing.

The analysis shows that lexical semantics alone is not sufficient to differentiate these two groups. What is different between the two groups is that dvandva compounds do not denote successive events while coordinations do. Dvandva compounds contain two entities that carry the same weight, but this weight is largely determined by pragmatics. For instance, Example 48 (2) *shuryou-saishuu* 'hunting and gathering' consist of *shuryou* 'hunting' and *saishuu* 'collecting'. The relation between these two constituents is not so clear unless we refer to the hunting-gathering period. The order of the events in coordination is also determined pragmatically. In Example 49 (2), it is clear that doctors diagnose their patient first, and treat the patient's symptoms after the diagnosis, but this knowledge comes from our pragmatic knowledge. It may be possible to distinguish dvandva compounds from coordinated V + V compounds using a WordNet (Miller & Fellbaum, 1992)[21]-type thesaurus, which arranges words according to hierarchically arranged concepts. However, this is left for future study.

[21] A brief explanation of WordNet is provided later in this section.

Example 48: Lexical Semantic Structure of Dvandva Compounds

(1) hakken hatumei

discover invent

'discovery and invention' or 'to discover and invent'

hakken discover [+dynamic ([ ],[ ])]

hatumei invent [+dynamic ([ ],[ ])]

(2) shuryou saishuu

hunt collect

'hunting and gathering'

shurou hunt [+dynamic ([ ],[$_i$]);+IEPS([$_i$][$_{PATH}$])]

saishuu collect [+dynamic ([ ],[$_i$]);+IEPS([$_i$][$_{PATH}$])]

(3) zoueki zoushuu

increase of profit increase of earnings

'increase of proceeds'

zoueki increase-of-profit [+dynamic ([$_i$]);+IEPS([$_i$][$_{PATH}$])]

zoushuu increase-of-earnings [+dynamic ([$_i$]);+IEPS([$_i$][$_{PATH}$])]

Example 49: Lexical Semantic Structure of Coordinated V + V Compounds

(1) kenkyuu kaihatu

investigate develop

'investigate and develop'

kenkyuu investigate [+dynamic ([ ],[ ])]

kaihatu develop [+dynamic ([ ],[$_i$]);+IEPS([$_i$][$_{PATH}$])]

(2) sindan tiryou

diagnose treat

'diagnosis and treatment' or 'to diagnose and treat'

sindan 'diagnose' [+dynamic ([ ],[ ])]

tiryou 'treat' [+dynamic ([ ],[$_i$]);+IEPS([$_i$][$_{PATH}$])]

5.3.1.3 Lexical Disambiguation

In the above sections, I presented my analysis without mentioning that a word may have more than one meaning or sense, and each sense may have a different lexical semantic requirement. In fact, there are a number of words that have multiple senses. For instance, *seityou* 'to grow' has three senses, according to the Daily Concise Japanese Dictionary (Satake, 2000). One sense is used for an object that physically grows. Another sense is used for an abstract entity that can develop. The last sense is used as a biological term that denotes a quantitative increase in the number or mass of an organ or cell. For the first two senses, the former requires a concrete noun, which is [+material ([ ])], and the latter requires an abstract noun, [-material ([ ])]. The biological term requires an organ or cell.

A dictionary must contain lexical semantic information for each sense, and any machine translation engine needs to be able to disambiguate which sense of a word is appropriate in a particular context. Lexical disambiguation is another major issue in Natural Language Processing. Jurafsky and Martin (2000) introduce two major approaches to word-sense disambiguation: the machine learning approach and the dictionary-based approach. Since one of the primary purposes of this study is to analyze the lexical semantic properties of Sino-Japanese compounds, the issue of word-sense disambiguation is left for a future study. However, it needs to be pointed out that in order to apply the present analysis to Natural Language Processing, a word-sense disambiguation algorithm is essential.

5.3.2 Compounds with a Regular Noun Head

A total of 894 regular noun compounds are collected. The details are provided in Table 18.

Table 18: Compounds with a Regular Noun Head

Deverbal Modifier	Number	Regular Noun Modifier	Number	De-adjectival Modifier	Number
Attributive	400	Attributive	407	Attributive	84
ChaSen Error	0	ChaSen Error	1	ChaSen Error	2
Total	400	Total	408	Total	86

As mentioned in Chapter 4, the first constituent always functions attributively to the head regardless of its part of speech. If the two constituents are both regular nouns, theoretically, we should be able to find dvandva compounds. However, these were not found in this particular dataset. Some of the types of constituent relations are shown in Example 50. Compounds with a regular noun demonstrate a variety of relations. In fact, more relations are expected to be found in other compounds not present in the data because as Marchand (1960: 22) states "it is no use trying to exhaust the possibilities of relationship; many combinations defy an indisputable analysis [. . .]. We will always try to classify, but it should be borne in mind that the category of compounding is not one that fills the need for classification [. . .]. In forming compounds we are not guided by logic, but by associations". Also, one compound can have multiple relations depending on the context as shown in Example 51.

The lexical semantic structure of the modifier and the head was identified to see if there is any pattern associated with any relation. However, it turns out that there was no association between a particular relation and the lexical semantic structure of the modifier and the head as shown in Example 52. In these examples, the compounds share the same head. It appears that there is no restriction on the lexical semantic structure of the modifier because there are a variety of lexical semantic structures that can be associated with the modifier.

Example 50: Types of Relation between the Modifier and the Head[22]

Compounds	Relations between N1 and N2
minkan-koukuu private aviation 'private aviation'	'operated by'
kagaku-heiki chemical weapon 'chemical weapon'	'made of'
keizai-mondai economic-issues 'economic issues'	'regarding'
zinkou-seisaku population-policy 'population policy'	'for'

Example 51: Compounds with Multiple Relations

Compounds	Possible Relations
seiyou-ongaku Western-music 'Western music	• produced in • came from
toumei-toryou Transparent-paint 'transparent paint'	• Paint that is transparent • Paint that becomes transparent when dry

[22] The table is not meant to be exhaustive. There are more relations found in naturally occurring discourse.

Example 52: No Association between the Relation and the Lexical Semantic Structure of the Constituents

Head and LS	Modifier	Compound
mondai 'issue' [-material ([ ])]	keizai economy [-material ([ ])]	keizai-mondai economy-issue 'economic issue'
	koyou employment/employ [-material, dynamic ([ ])]	koyou-mondai employ-issue 'issue of employment'
	kensetu build, construct [-material, dynamic ([ ])]	kensetu-mondai construct-issue 'issue of construction'
	bouei defense/defend [-material, dynamic ([ ])]	bouei-mondai defense-issue 'defense problem'
seisaku policy [-material ([ ])]	kyousou compete/competition [-material, dynamic ([ ])]	kyousou-seisaku compete-policy 'policy for (the) competition'
	wakai settle/settlement [-material, dynamic ([ ])]	wakai-seisaku settle-policy 'policy for settlement'
	koukyou public [-dynamic ([ ])]	koukyou-seisaku public-policy 'public policy'
	zinkou population [-material ([ ])]	jinkou-seisaku polulation-policy 'population policy'

Lieber (2004) examines English root compounds, ones in which the head is not derived from a verb as shown in Example 53, and states that the lexical semantics alone cannot identify the relation between the modifier and the head. Lieber cites Selkirk (1982: 22) as follows:

> The semantic relation obtaining between the head constituent and its sister nonhead constituent can vary considerably, though, and a general characterization of the relation is probably impossible . . . it would seem that virtually any relation between head and nonhead is possible – within pragmatic limits, of course.

Example 53: English Root Compounds (Lieber 2004: 52)

Lexical Semantic Structure of *dog bed*

Skeleton	[+material ([$_i$])]	[+material ([$_i$])]
	dog	*bed*
Body	<natural>	<artifact>
	<animate>	<furniture>
	<canine>	<horizontal surface>
		<for sleeping>

Arehart (2003) states that "there are no linguistic constraints on noun compounding meaning, and that interpretation is wholly a matter of pragmatics and real-world knowledge". He proposes a method to identify possible head-modifier relations using statistical techniques, Description Logic formalism and WordNet (Miller & Fellbaum, 1992). WordNet is a lexical database in which words are hierarchically arranged based on semantic concepts. He generates all possible denotations from denotation templates, which are expressed in Description Logic, and eliminates redundant or implausible denotations using some heuristics. The rest of the candidates undergo statistical procedures and are ranked. He claims his method is language-independent. Therefore, it can be applied to Sino-Japanese root compounds as well as other types of compounds to identify possible constituent relations, assuming that a lexical database like WordNet exists in Japanese.

Although Lexical Semantics alone does not seem to help us identify the head-modifier relations of compounds with a regular noun head, what can be said is that the modifier is always attributive to the head constituent and the whole compound denotes a subcategory of the compound head. In fact, the translation of Japanese compounds also shows that when the compounds are translated compositionally the modifier specifies the subcategory of the head as shown in Example 54. Therefore, it seems that the identification of the head's part of speech in root compounds alone is useful if one were to translate Japanese root compounds into English.

Example 54: English Translation of Japanese Root Compounds

(1) zinkou seisaku
population policy
'**population** policy'

(2) kokusai shakai
international society
'**international** society'

(3) kokumin sinri
citizens psychology
'the psychology **of the people**'

- The translation of the modifier also modifies the head.

5.3.3 Compounds with a De-adjectival Head

There are three types of relations found in de-adjectival compounds as discussed in Chapter 4. The three relations are emotion-theme, subject-predicate, and attributive relation. There are five compounds whose head is an emotion word. These emotion words are expected to have an experiencer, which is typically an animate object, and the thing the experiencer has a feeling towards. Among the five compounds, the experiencer never appears within the compound. The situation or thing towards which the experiencer has a feeling tends to be an abstract concept rather than a concrete object. In reality a person can have a feeling towards a concrete animate object although when the experiencer has a feeling towards an animate object, it is usually the case that the experiencer has a feeling towards the attitude, state, or condition of the animate object, rather than the animate object itself.

Nonetheless, it is worth noting that when these emotion words take a copula, the thing towards which the experiencer has a feeling tends to be an abstract concept. When one searches *–ga human-da*[23] '-NOM dissatisfied-COP', using the search engine Google to try to elicit what the experiencer has a feeling towards, the search engine returns 56,900 sites. Among the first 100 sites, there are 86 instances in which the thing the

[23] The experiencer tends to appear with the topic marker *–wa* and the theme tends to appear with the subject marker *–ga* in Japanese.

experiencer has a feeling towards is an abstract concept, 6 instances in which the subject marker is attached to the experiencer, and it is not clear what the experiencer is dissatisfied with, and the rest could be either an abstract concept or a concrete object because it was an interrogative pronoun or demonstrative (performed on May 24, 2006, www.google.co.jp). It is possible to hypothesize that emotion words, at least when taking a copula, prefer an abstract concept to be their theme. However, it is impossible to draw any conclusion from a small set of data.

There are fifteen compounds whose head functions as the predicate adjectival noun of the non-head constituent. The non-head constituent is always deverbal in the data analyzed as shown in Example 55. Many of these compounds share the same head because there are only four different words that appear as the head. Therefore, the variety of the head is limited. In the examples below, the lexical semantic structure of the non-head constituent and the lexical semantic requirement of the default subject of the predicate do seem to be compatible, but any firm conclusion on this type of compounds cannot be drawn based on the properties of four heads.

Example 55: Lexical Semantic Analysis of Compounds whose First Constituent is Deverbal

Compounds	LS requirement of the head	LS of the first constituent
zizoku-kanou sustaining-possible 'sustainable'	[-material, dynamic ([ ])]	[-material, dynamic ([ ])]
teigi-konnan defining-hard 'hard to define'	[-material, dynamic ([ ])]	[-material, dynamic ([ ])]
tousen-mukou be elected-invalid 'the nullification of the election'	[-material, dynamic ([ ])][24]	[-material, dynamic ([ ])]
senkyo-husei election-illegal 'electoral fraud'	[-material, dynamic ([ ])]	[-material, dynamic ([ ])]

[24] *Mukou* 'invalid' has two senses, according to the Concise Japanese dictionary (Satake, 2000), one in which an object such as a subway pass becomes invalid, and the other in which things that were decided in an event become invalid. In the case of the compound 'tousen mukou', the latter definition is more suitable because *tousen* 'be elected' is eventive.

When the non-head constituent is a regular noun, de-adjectival compounds behave like regular noun compounds, as discussed in 4.5.3. Like regular noun compounds, the modifier-head relation is not obvious, shown in Example 56. Arehart's (2003) method, mentioned in 5.3.2, may be able to identify possible interpretations of the constituent relation. In any case, lexical semantic structure does not seem to play an important role in determining the modifier-head relations.

Example 56: Modifier-Head Relation of De-adjectival Compounds whose First Constituent is a Regular Noun

Compounds	Relation
koukyuu-heiwa eternal-peace 'eternal peace'	Subcategory of the head
sekai-hukyou world-slump(economy) 'global slump'	Subcategory of the head

5.4 Summary

In this chapter, the lexical semantic analysis of four-character Sino-Japanese compounds has been presented using Lieber's framework. There are four different relations found in deverbal compounds: argument relation, adjunct relation, dvandva and coordination. For dvandva compounds and coordinations, the analysis reveals that Lexical Semantics alone does not help distinguish these two types because the constituent relations are pragmatically constrained. The present data suggest that the two constituents have an identical lexical semantic structure in dvandva compounds while it may not be the case for coordinated V + V compounds. For non-V + V compounds, if the lexical semantic requirement of the head is fulfilled by the lexical semantic structure of the non-head constituent, the constituents tend to have an argument relation while it has an adjunct relation if it is not fulfilled.

The present analysis suggests that the constituent relations of regular compounds is solely pragmatic, and lexical semantics does not help identify possible constituent

relations. Arehart (2003) attempts to identify possible modifier-head relations in English root compounds using WordNet. It may be possible to apply his algorithm to Japanese root compounds, provided that Japanese has a WordNet-like lexical database because he claims that his method is language-independent. However, to my knowledge, Japanese does not have a WordNet-like thesaurus that is freely available. For de-adjectival compounds, it seems that compounds behave differently depending on the part of speech or the type of the non-head constituent. When the head is an emotion or perception word, the lexical semantic requirement of the head seems to be fulfilled by the non-head constituent. When the head is not an emotion word, and a deverbal noun, the head behaves as the predicate of the non-head constituent. The lexical semantic requirement of the head seems to be satisfied by the first constituent. When the non-head constituent is a regular noun, the constituent relation is attributive. There are only two instances in which both of the constituents are de-adjectival. Because the constituent relation in these two compounds is different, it is impossible to generalize what the relations this type of compounds can have. In general, de-adjectival compounds are very rare. More compounds are necessary to make a generalization of the properties of these compounds.

CHAPTER 6:
APPLICATION OF THE PRESENT STUDY TO MACHINE TRANSLATION

6.1 Introduction

This chapter suggests a classification algorithm of Sino-Japanese compounds, excluding V + V coordinations, V+V or N+N dvandva compounds, based on the analysis in Chapters 4 and 5. After the classification algorithm, a translation algorithm is suggested for deverbal compounds. Because Lexical Semantics alone does not help us determine the relationship between the constituents of regular noun compounds, dvandva compounds, and V + V compounds, translation algorithms for these compounds are left for future research. De-adjectival compounds are rare, and more compounds are necessary to generalize the characteristics of these compounds. Hence, no translation algorithm for these compounds is suggested. In the present study, the algorithm is suggested, but implementation and testing are left for a future study.

The assumption in this chapter is that the Japanese dictionary implemented in the machine translation engines encodes all possible word senses, and has all lexical semantic information encoded in a machine readable form.

6.2 Classification Algorithm

Chapter 4 demonstrated that the part of speech of the head and the non-head constituent plays a key role in classifying Sino-Japanese compounds. Taking advantage of ChaSen, which tags Japanese text with parts of speech, one can classify Sino-Japanese compounds using the parts of speech provided by ChaSen. The present study assumes that Sino-Japanese compounds can be extracted from a text using an extraction algorithm such as the one provided by Baldwin and Tanaka (2004), which is briefly explained in Chapter 3. As mentioned earlier, coordinated V + V compounds and dvandva compounds, which include deverbal and regular noun dvandva compounds are excluded from this

classification algorithm because these behave differently from other compounds in that both of the constituents in these compounds have a head while others only have one head. No algorithm that can isolate these compounds from others has been proposed, but this thesis assumes that they can be isolated.

ChaSen segments four-character Sino-Japanese compounds into two two-character compounds as shown in Example 57.

Example 57: ChaSen Segmentation of Sino-Japanese Compounds

First Constituent:	研究 'investigation' 名詞-サ変接続 'deverbal noun'
Head:	成果 'result' 名詞-一般 'regular noun'

Compounds can be first grouped according to the part of speech of the second constituent, which is the head, because the properties of the whole compound are determined by the head. Then compounds can be further grouped according to the part of speech of the non-head constituent. Regarding regular noun compounds, the constituent relation was attributive regardless of the part of speech of the non-head constituent. Therefore, the identification of the head's part of speech helps determine the relation of regular noun compounds. For de-adjectival compounds, if the head is an emotion word, the non-head constituent acts as the theme of the emotion word. If the head is a non-emotion word, and the non-head constituent is a regular noun, the relation seems to be attributive. Finally, the constituents have a predicate relation if the first constituent is a deverbal noun. However, as mentioned in Chapter 5, only a few examples of de-adjectival compounds appear in the corpus, and the behaviour of these compounds alone should not be extended to all de-adjectival compounds.

With respect to deverbal compounds, the analysis in Chapter 5 suggests that compounds that hold an argument relation can be distinguished from ones that have an adjunct relation if the lexical semantic structure of the head and non-head constituents are taken into consideration. Deverbal heads have an argument structure. The analysis in Chapter 5 shows that if the lexical semantic structure of the non-head constituent matches

the head's lexical semantic requirement on its argument, the constituents tend to have an argument relation while they have an adjunct relation when it does not. In order to automatically classify deverbal compounds, I propose that lexical semantic information, for instance, as presented in Example 58 be encoded in the dictionary. In Example 58 (1), the head is transitive. Therefore, theoretically, only the theme can be part of the compound. In the example, the theme requirement of the head matches the lexical semantic structure of the non-head constituent. Therefore, it is possible that the compound has an argument relation. In fact, it does have an argument relation, as can be seen in the gloss. On the other hand, in (2), the theme requirement of the head is not the same as the lexical semantic structure of the non-head constituent. Therefore, it is likely that the compound has an adjunct relation. Indeed, it is the case that the compound relation is adjunct.

When a deverbal noun appears as the non-head constituent, it never has an argument structure. It may fulfil the argument requirement of the head. However, it does not itself have an argument requirement. This suggests that one entry requires two lexical semantic representations, one in which the deverbal noun functions as a noun, and the other in which it functions as a verb. When it functions as a noun, the deverbal noun has no argument requirements.

Example 58: Lexical Semantic Information in the Dictionary

(1) keizai kaikaku

economy reform
'economic reform'

Lexical Entry: kaikaku 'reform'

TRANSITIVITY: transitive
AGENT: [+material, +animate ([ ])]
THEME: [-material ([ ])]

Lexical Entry: keizai 'economy'
Structure: [-material ([ ])]

(2) kyoudou teian

joint propose
'joint poposal'

Lexical Entry: teian 'propose'

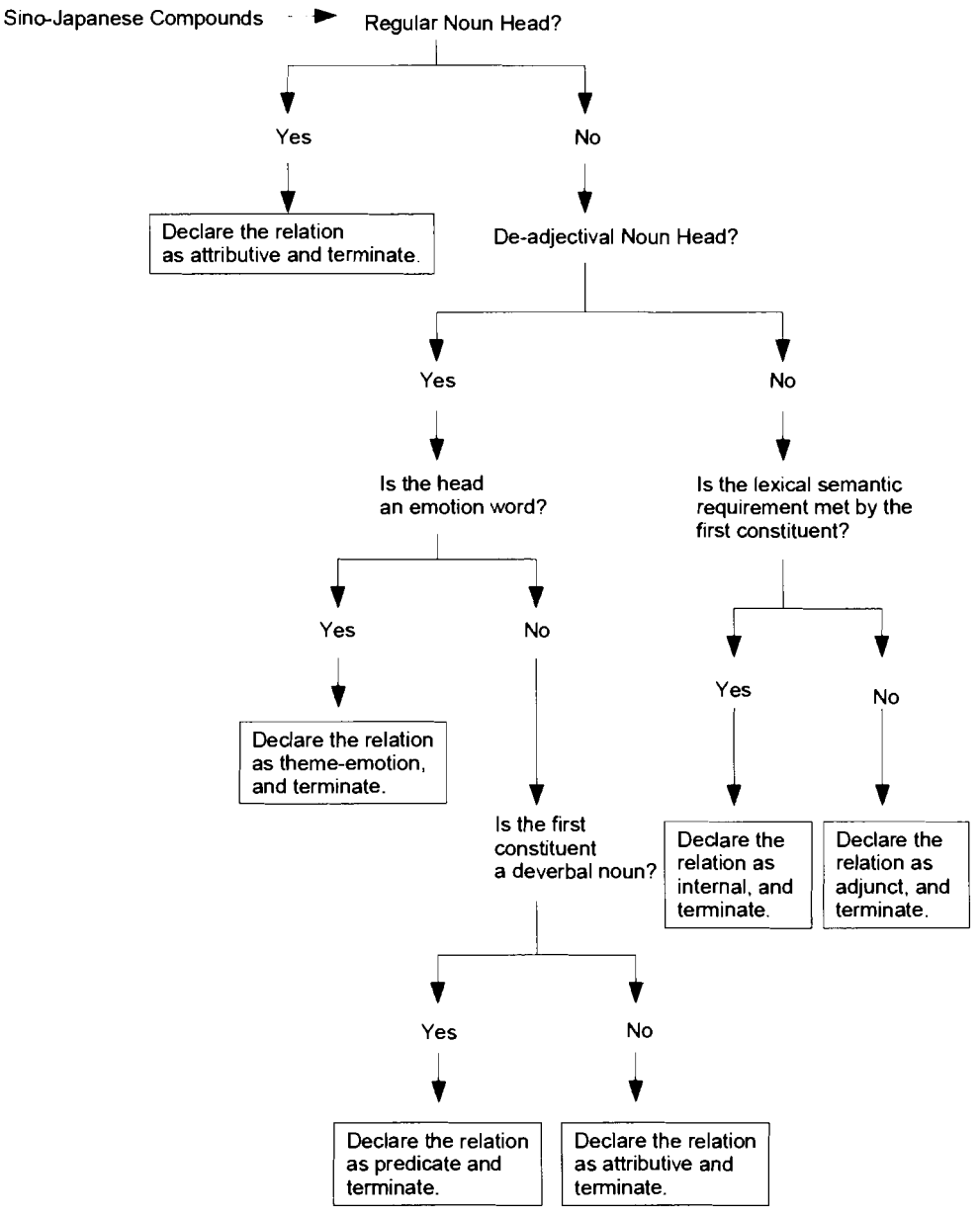
TRANSITIVITY: transitive
AGENT: [+material, +animate ([ ])]
THEME: [-material ([ ])]

Lexical Entry: kyoudou
Structure: [-dynamic ([ ])]

Figure 14 shows the summary of the classification algorithm for Sino-Japanese compounds, excluding dvandva compounds and coordinated V + V compounds. It is assumed that, when the deverbal head is transitive, the machine is instructed to compare the structure of the theme or the patient of the deverbal head and the non-head constituent. Likewise, if the deverbal head is intransitive, it is instructed to look at the structure of the agent of the intransitive head and the non-head constituent. When the head is unaccusative, it looks at the theme requirement of the head and the structure of the non-head constituent. Also, it needs to be mentioned that machine needs to be able to distinguish emotion words from non-emotion words.

The classification algorithm suggested in Figure 14 is strictly for Natural Language Processing. It does not in any way model how Sino-Japanese compounds are acquired.

Figure 14: Classification Algorithm (Excluding V + V and dvandva compounds)



6.3 Translation of Deverbal Compounds

The procedure of deverbal compound translation is as follows. A text that needs to be translated goes through ChaSen (see 4.3.2.) and is segmented into content words, inflections and case markers along with their part of speech. Then, Sino-Japanese compounds are extracted from the text by an extraction algorithm. This thesis does not provide any suggested extraction algorithm, but the one suggested by Baldwin and Tanaka (2003a) was described in 3.3.2.1. Once Sino-Japanese compounds are extracted from the text, these compounds are sorted out by the classifier outlined in the previous section. The classification was motivated by the syntactic behaviour of each group.

After the classification, the text undergoes a translation algorithm. Based on the observation of the parallel Japanese-English corpus, the syntactic behaviour of these compounds is reflected by the translation. For instance, when the head is deverbal, it tends to be translated into a verbal element in English. Likewise, when the head is a regular noun or a de-adjectival noun, their part of speech tends to be preserved in the translation. In this section, a translation algorithm for deverbal compounds is suggested, based on the translation patterns found in the Utiyama Corpus.

The classification algorithm suggested in the previous section categorizes the compounds based on their syntactic behaviour. Therefore, the algorithm is language-independent; it can be used in translation engines that deal with Japanese and a language other than English. However, the proposed translation algorithm uses translation patterns of compounds as translation templates. Therefore, it is only compatible with translation engines that deal with Japanese-English pair; it cannot be used for other language pairs such as English-Japanese because the translation templates may vary depending on the direction of the translation.

Deverbal compounds, excluding dvandva compounds and coordinations, are translated in many ways. For instance, the head of deverbal compounds can be translated into a finite verb, gerund, derived noun, infinitive, phrasal verb or passive. Some examples of English translation are provided in Example 14 in 4.5. In this section, only compositional compounds (see 3.3.2.1) are discussed. Non-compositional compounds are discussed in the next section. A comparison between the Japanese text and its English

translation in the Utiyama Corpus reveals that there seem to be translation patterns in deverbal compounds, and some differences are observed between the two types of compounds, ones that have an argument relation and ones that have an adjunct relation. The patterns found in the Utiyama Corpus are shown in Figure 15. In the figure, fourteen patterns are identified. However, more patterns are possible, since I did not analyze the entire corpus.

Some translation patterns in Figure 15 reflect the structure of the compounds. For ones that have an internal relation, the first member acts as the theme or patient of the deverbal head if the head is transitive, and it acts as the agent if the head is intransitive[25]. In English, theme follows the verb if it is an object. Therefore, we can expect that the head of a deverbal compound may be translated into a verbal element and the first constituent is translated into a noun occupying the object position. As a matter of fact, this pattern is found as can be seen in Figure 15 (1). In (1), the head is translated into a verb in English, but the form of the verb depends on the position in which the compound occurs. For instance, if it is used in a main clause, it may be translated into a finite verb.

When a transitive verb is passivized, the theme occupies the subject position, and the verb appears in the past participle form in English. This pattern is also found in (3) under Internal Relation. Further, the complement of a noun that is derived from a transitive verb may appear in the form of '*of* N' or just a noun phrase in English. This is also found in the corpus as can be seen in (4) and (5). If the head is intransitive, we can predict that the non-head constituent may appear as the agent, or the theme if the verb is unaccusative. This pattern corresponds to (2) under Internal Relation in Figure 15. However, oddly, this pattern was not common in the corpus. The patterns (6) to (8) under Internal Relation are unexpected considering the structure of the Sino-Japanese compounds because adjectives or prepositional phrases cannot be arguments of a verb. Nonetheless, these patterns are found in the corpus.

[25] If the verb is unaccusative, it is the theme that occupies the subject position.

Figure 15: Translation Patterns

Internal Relation	Adjunct/Attributive
(1) V_2 + N_1 (V can be finite, gerund, or infinitive depending on the position in which the compound appears.) e.g. katuryoku-iji vigor-maintain 'maintain the vigor'	(1) Adj_1 + N_2 e.g. keizai-kyouryoku economy-cooperate 'economic cooperation'
(2) (N_1 + V_2) Theoretically, it should be found, but this pattern did not occur frequently.	(2) N_1 + N_2 e.g. tousi-katudou investment-act 'investment activity'
(3) N_1 be V_2-ed e.g. riyou-sokusin use-promote 'the use of (something) is promoted.	(3) V_2 + Adv_1 or Adv_1 + V_2 e.g. kyoudou-teian joint-propose 'jointly propose' douji-tuuyaku simultaneous-interpret 'interpret simultaneously'
(4) N_1 + N_2 e.g. seisaku-handan policy-judge 'policy judgement'	(4) V_2 + PP_1 e.g. dantai-koudou group-behave 'go as a group'
(5) N_2 (of) N_1 e.g. jouhou-koukan information-exchange 'exchange of information'	(5) Adv_1 + V_2-ed e.g. kanzen-zissi complete-implement 'fully implemented'
(6) Adj_1 + N_2 e.g. kouzou-kaikaku structure-reform 'structural reform'	(6) Complex V or Complex N e.g. kashou-hyouka too.small-evaluate 'underestimate' kazyou-hannou too.much-react 'overreaction'
(7) Adj_2 + N_1 e.g. tousi-zoudai investment-increase 'increasing investment'	
(8) N_2 + PP_1 e.g. kabuka-geraku share price-drop 'drop in share prices'	

The subscripted number corresponds to the constituent of the compound. 1 corresponds to the first member, and 2 corresponds to the head.

With respect to deverbal compounds that have an adjunct relation, the non-head constituent of the compound is expected to appear as an adverbial phrase or an adjectival phrase because it only modifies the head, and holds no thematic relation to the head. The translation patterns (1) to (5) in Figure 15 match the prediction. The pattern (6), which is Complex V or Complex N, also reflects the structure of compounds with no internal relation. The examples of complex verb and complex noun in Figure 15 have a preposition incorporated into the verb or noun. In these compounds, the preposition component functions adverbially. In fact, the prepositional part of these complex words corresponds to the modifier of the compounds as shown in Example 59.

Example 59: Complex Words

(1) kazyou hannou
too.much react/reaction
'overreaction' or 'overreact'

(2) kashou hyouka
too.small evaluate/evaluation
'underestimate'

In some cases, translation is influenced by personal preference. For instance, the same sentence can be translated into multiple ways as shown in Example 60. However, translation does not need to be 'perfect' because translation is good enough as long as the semantic content conveyed by the compound is retained. Therefore, if we make use of the translation patterns found in Figure 15 one can approximate the translation. The suggested translation method is as follows. First, a set of translation templates is created for compounds with an internal relation and another set for compounds with an adjunct relation, based on the translation patterns. A translation engine generates possible translation candidates using the translation templates. Using a large English corpus such as the British National Corpus (Burnard, 2000), one can calculate the frequency of the generated translation candidates and rank them according to their frequency. The candidate that has the highest frequency can be considered the optimal translation

candidate. The summary of the proposed translation method is provided in Figure 16. It assumes that Sino-Japanese compounds are already extracted from the text that needs to be translated using an extraction method, such as the one presented by Baldwin and Tanaka (2004). Therefore, the heads the proposed algorithm deals with must be a regular noun, deverbal noun, or de-adjectival noun. Another assumption is that deverbal compounds that have an argument relation are isolated from ones that have an adjunct relation prior to running the translation algorithm using a classification algorithm such as the one outlined in the previous section.

Example 60: Translation Variation

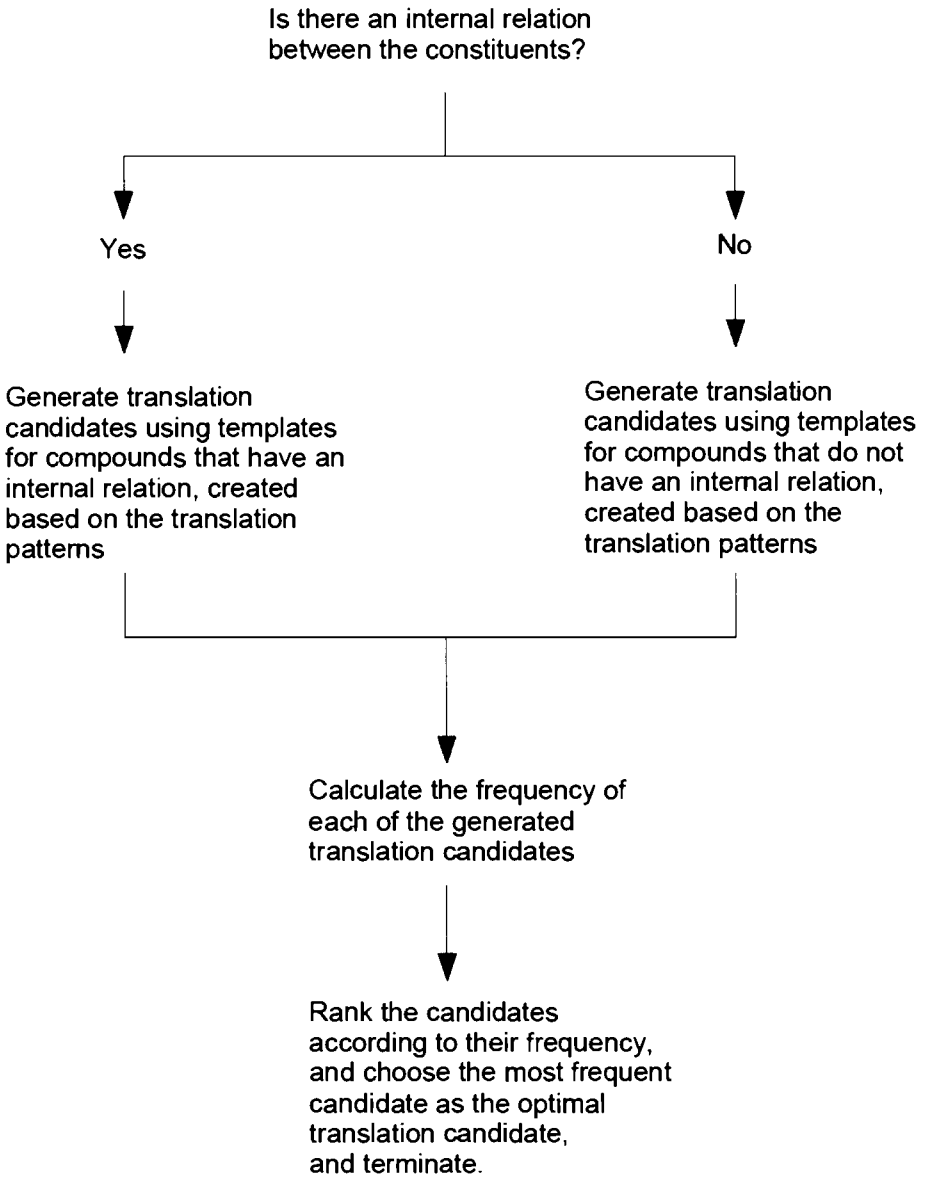
(1)	hutari-ga	kyoudou	site	**kagaku**	**kenkyuu**-o	okonat-ta
	2-people-NOM	joint	do	science	research-OBJ	carry-out-PAST

The words in bold face are a Sino-Japanese compound.

Translation of the Sino-Japanese compounds:

'Two people carried out a joint **scientific research**.'
'Two people carried out a joint **scientific investigation**.'
'Two people carried out a joint **scientific study**.'

Figure 16: Translation Algorithm



6.4 Translation of Non-Compositional Compounds

As mentioned in 3.3.2.1, compositional compounds refer to ones whose translation can be obtained by translating each constituent and combining them in a certain way, while non-compositional compounds refer to ones whose translation cannot be obtained by translating each constituent. The proposed translation method in 6.3 assumes that

compounds can be translated compositionally. However, there are some non-compositional compounds. Among non-compositional compounds, some can be translated compositionally without losing the meaning of the compound while others cannot. Example 61 (1) shows the former and (2) the latter.

Example 61: Non-Compositional Compounds

(1) hasan zyoutai
insolvency state
'insolvent'

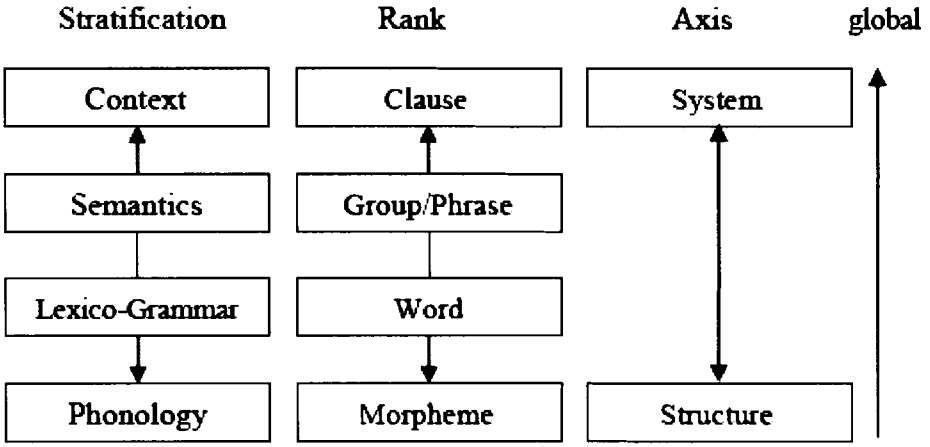
(2) kidou shuusei
orbit adjust
'make adjustment'

For ones that can be translated compositionally without losing their meaning, the translation method proposed in 6.3 can be used. However, the proposed translation method cannot be used for true non-compositional compounds, such as (2) in Example 61 because otherwise the meaning of the compound will be lost. Based on the dataset, it seems that there is a correlation between compositionality and idiomaticity. In Example 61, the second example is idiomatic in Japanese while the first example is not. When the degree of idiomaticity increases, the degree of compositionality also increases.

One possible solution to the translation of non-compositional compounds is to translate them at a more abstract level rather than translating them at the word level; in other words, one can translate the concept the compounds convey rather than translate the constituents of the compounds literally. Matthiessen (2001) states that language is a semiotic system, and is organized in three dimensions, all of which have hierarchically arranged components as shown in Figure 17. For instance, under Stratification, language is organized into context, semantics, lexico-grammar, and phonology. These are hierarchically ordered such that context is the most global and phonology is the most local. Matthiessen (2001) claims that changes may be required somewhere in the Rank scale during the process of translation because the rank scale varies across languages.

Also, he states that if translation cannot be found at a particular level, one should go up the hierarchy to find the equivalent translation in the target language because the degree of translation equivalence is higher if translation occurs in the more global environment.

Figure 17: Structure of Language (Matthiessen 2001: 81)



Compositional compounds can be translated at the word level. However, non-compositional compounds have to be translated above the word level if an equivalent translation cannot be found at the word level. In so doing, one needs to understand the concept or semantic meaning the compound conveys. Humans can quite easily express the same concept in various ways. However, it is a hard task for computers because computers cannot 'comprehend' what we call 'concept'. The only option for non-compositional compounds in Machine Translation is to encode their translation in the dictionary although it is inefficient and unrealistic because Sino-Japanese compounds are very productive as mentioned in 3.2, and therefore, the number of non-compositional compounds in the Japanese language is unknown. The estimate of non-compositional compounds is provided by Baldwin and Tanaka's (2004). For the compounds they collected for their study, 43.1% of the compounds are compositional, and the rest are non-compositional[26]. However, the estimate may change because sometimes compositionality is influenced by the translator's personal preference. For instance,

[26] Baldwin and Tanaka do not provide the exact percentage of non-compositional compounds in the paper. I assume that the remaining 56.9% are non-compositional.

zangyaku-koui (cruel-act/behaviour) 'atrocity' may seem non-compositional, but if it is translated as 'cruel act' or 'cruel behaviour', it can be considered compositional.

6.5 Discussion

This chapter assumed that the constituents of the compounds are all found in the dictionary. However, in a real life situation, there are cases in which the constituent is not listed in the dictionary. In such case, a machine translation engine may be able to guess the part of speech of the unknown word by looking at the surrounding elements, but identifying the exact word meaning, part of speech and translation is impossible because there is simply no resource in the machine translation engine to handle unknown words. Because of this reason, handling unknown words is an issue for any current machine translation engine.

The proposed translation algorithm was based on the translation patterns found in the Japanese-English parallel corpus. Therefore, the translation patterns may be different if the language pair a machine translation engine deals with is not Japanese-English. For instance, if the language pair was English to Japanese, as opposed to Japanese to English, a different set of translation patterns may be found. Since the translation patterns identified in this chapter is language-dependent, the proposed translation algorithm is only suitable for Japanese-English translation engines that use the transfer approach, which requires language-dependent rules to translate one language to another as mentioned in 3.1.1. The proposed translation algorithm cannot be implemented in translation engines that use the interlingua approach because it does not convert the structure of Sino-Japanese compounds into a language-neutral semantic representation.

In Chapter 3, several approaches to Sino-Japanese compound translation are introduced. Although these approaches have applications to automatic translation of Sino-Japanese compounds in mind, most of them present a segmentation or classification method. The only approach that presents translation is the one suggested by Baldwin and Tanaka (2003a, 2004). They propose a template method for compound translation. This approach is similar to my approach. However, what is different is that Baldwin and Tanaka did not take into account the morpho-syntactic structure and the lexical semantic

properties of the compounds while in my approach, compounds are grouped according to their morpho-syntactic and lexical semantic behaviour. Also, the dataset Baldwin and Tanaka used contains 500 compounds, which is a very small dataset. In my study, I collected 1860 compounds.

Another approach that is close to my approach is the one proposed by Takeuchi et al. (2003b). Their study did not propose a translation method. Their main focus was to classify deverbal compounds into ones that have an argument relation and ones that have an adjunct relation. In their approach, the lexical semantic structure of the head of the compound is identified using Jackendoff's (1990) lexical semantic framework. They state that there are more than eighteen possible lexical semantic structures at the time of writing. They are still trying to find other possible structures. Therefore, the number of possible lexical semantic structures may increase. After identifying the structure of the head, the first constituent of the compound is analyzed whether it can take the accusative case when it appears alone in a sentence regardless of context. However, this procedure alone is not sufficient to classify the deverbal compounds. The first constituent is then analyzed and labelled according to what type of verb it is used in a sentence when it appears alone. The label of the first constituent is then checked against the lexical structure of the head, whether the first constituent is compatible with the lexical semantic requirement of the head. Therefore, there are four steps involved in classifying deverbal compounds. In my approach, the lexical semantic structure of the head's default complement and the first constituent is identified using Lieber's (2004) framework. Her framework was chosen because the number of possible lexical structures of deverbal nouns is only four. Also, her framework accounts for not only verbs but also nouns, which cannot be accounted for by Jackendoff's framework because he is largely concerned with the lexical semantic structure of verbs. After identifying the structure of the head's default complement and the first constituent, their lexical semantic properties are checked against each other to see if they match. If they match, the compound is likely to have an argument relation while if they do not match, it is likely to have an adjunct relation. In my approach, there are three steps to classify deverbal compounds as opposed to four steps that were required in Takeuchi et al.'s approach.

6.6 Summary

In this chapter, I first proposed a Sino-Japanese compound classification algorithm. This classification was based on the syntactic behaviour of the compounds in Chapter 4. Therefore, it can be used in machine translation engines that deal with language pairs other than Japanese-English. Once compounds are classified by the proposed classifier, the compounds are sent to a translation algorithm. In this chapter, only a translation algorithm for deverbal compounds is suggested. This translation algorithm uses translation patterns found in the Utiyama Corpus as translation templates. The constituents of the compounds are translated at the word level, and translation candidates are generated using the translation templates. The translation candidates are then ranked according to their frequency in an English corpus, and the most frequent candidate is chosen as the optimal translation. Because this translation algorithm uses language-pair-dependent translation templates, it is only compatible with translation engines that deal with Japanese-English pair.

The suggested algorithm assumes that compounds can be translated compositionally. However, according to Baldwin and Tanaka (2004)'s study, 56.9% of Sino-Japanese compounds are non-compositional. One possible solution to the translation of these compounds is to translate the concept the compounds convey. However, it is impossible for a machine translation engine because it simply does not understand 'concepts'. Finally, the proposed translation algorithm is compatible with translation engines that use the transfer approach because it uses language-dependent translation templates. It is not compatible with the interlingua approach because it does not convert the compounds into a language-neutral representation.

CHAPTER 7: CONCLUSION

The goals of this thesis were to:

- Develop a classification system of four-character Sino-Japanese compounds, based on a corpus study,
- Identify the lexical semantic structure of four-character Sino-Japanese compounds,
- Suggest an automatic classification algorithm for Sino-Japanese compounds, and
- Propose a translation algorithm for Sino-Japanese compounds whose head is deverbal.

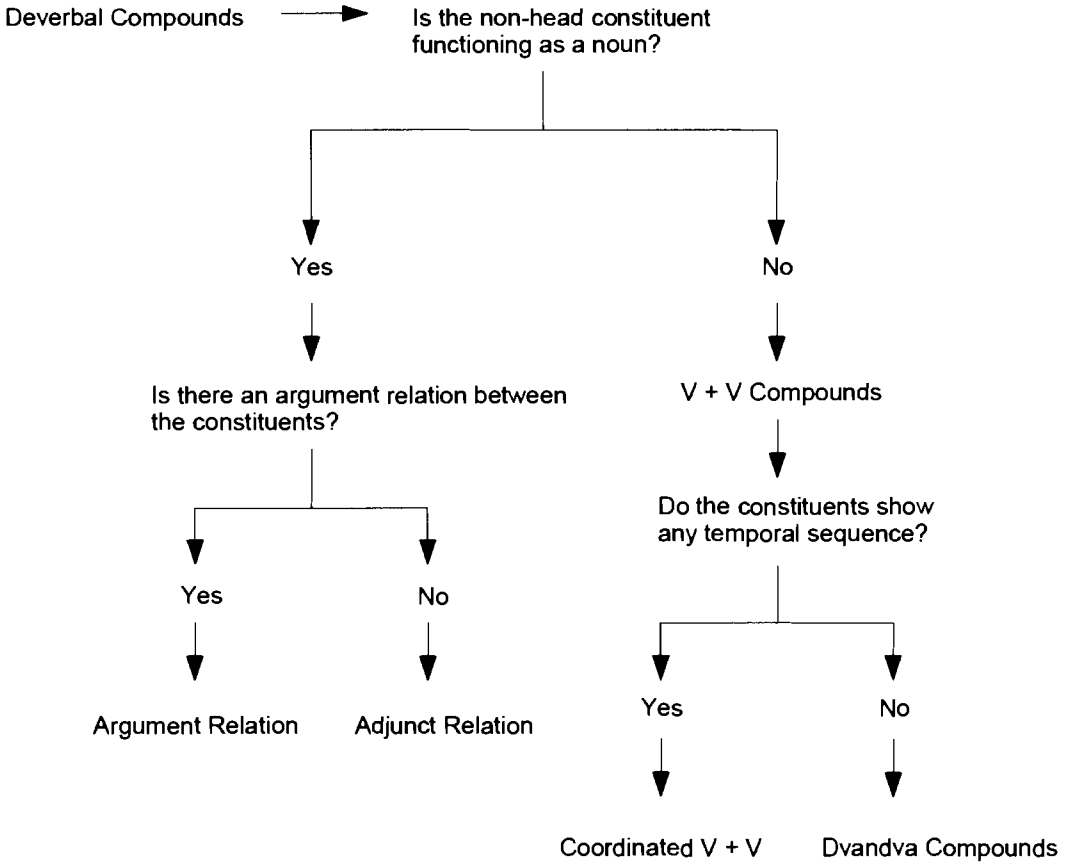
This thesis has provided a unified classification system for four-character Sino-Japanese compounds. This classification system uses the part of speech of the compound head and the non-head constituent. The compounds are first grouped according to the head's part of speech because the head's part of speech determines the properties of the whole compound. This complies with William's (1981b) Righthand Head rule which is introduced in 2.4.1. The classification is as follows. First, Sino-Japanese compounds are grouped according to the part of speech of the head. Then, they are further classified according to the part of speech of the non-head constituent. Another motivation for this classification system is that the head's part of speech tends to signal its part of speech in the English translation. For instance, if the head is a regular noun, it is likely to be a regular noun in the translation. Likewise, a deverbal noun is translated into a verbal element, and a de-adjectival noun tends to be translated into an adjective in English. This classification thus helps arrive at the correct translation of the Sino-Japanese compounds.

After the compounds are classified according to their head, they are further grouped according to the part of speech of the non-head constituent, which is the first constituent of the compound. Although the part of speech of the non-head constituent does not seem to be useful in accounting for the behavioural differences in deverbal and

regular noun compounds, it may be useful in de-adjectival compounds because the part of speech of the non-head constituent seems to divide de-adjectival compounds in a unified way. Specifically, if the non-head constituent is a deverbal noun, its head tends to be the predicate of the non-head constituent. If it is a regular noun, the head and the non-head constituent have an attributive relation. Apart from the part of speech of the head, another criterion that needs to be added to de-adjectival compounds' heads is sentiment because if the head of the compound is an emotion or perception word that expresses sentiment, the non-head constituent is usually the theme of the head.

After the compounds are classified according to the constituents' part of speech, each group is analyzed in terms of syntactic behavioural differences. Compounds whose head is a deverbal noun demonstrate four distinctive types. If the non-head constituent is functioning as a noun, compounds may have an argument relation or an adjunct relation between the constituents. It has been discussed that compounds that have an argument relation resemble Mithun's (1984) Type III noun incorporation. The syntactic properties and differences between the Sino-Japanese compounds and Type III noun incorporation is discussed in 4.5. If the non-head constituent is functioning as a verb, the compounds are considered V + V compounds. Among these V + V compounds, there are two types, which are distinguished by the temporal sequence of the constituents. If there is a temporal sequence, the compounds are considered coordinated V + V compounds, and the event sequence is reflected by the order of the constituents. If there is no temporal sequence, the compounds are dvandva compounds. The figure below summarizes the classification of deverbal compounds.

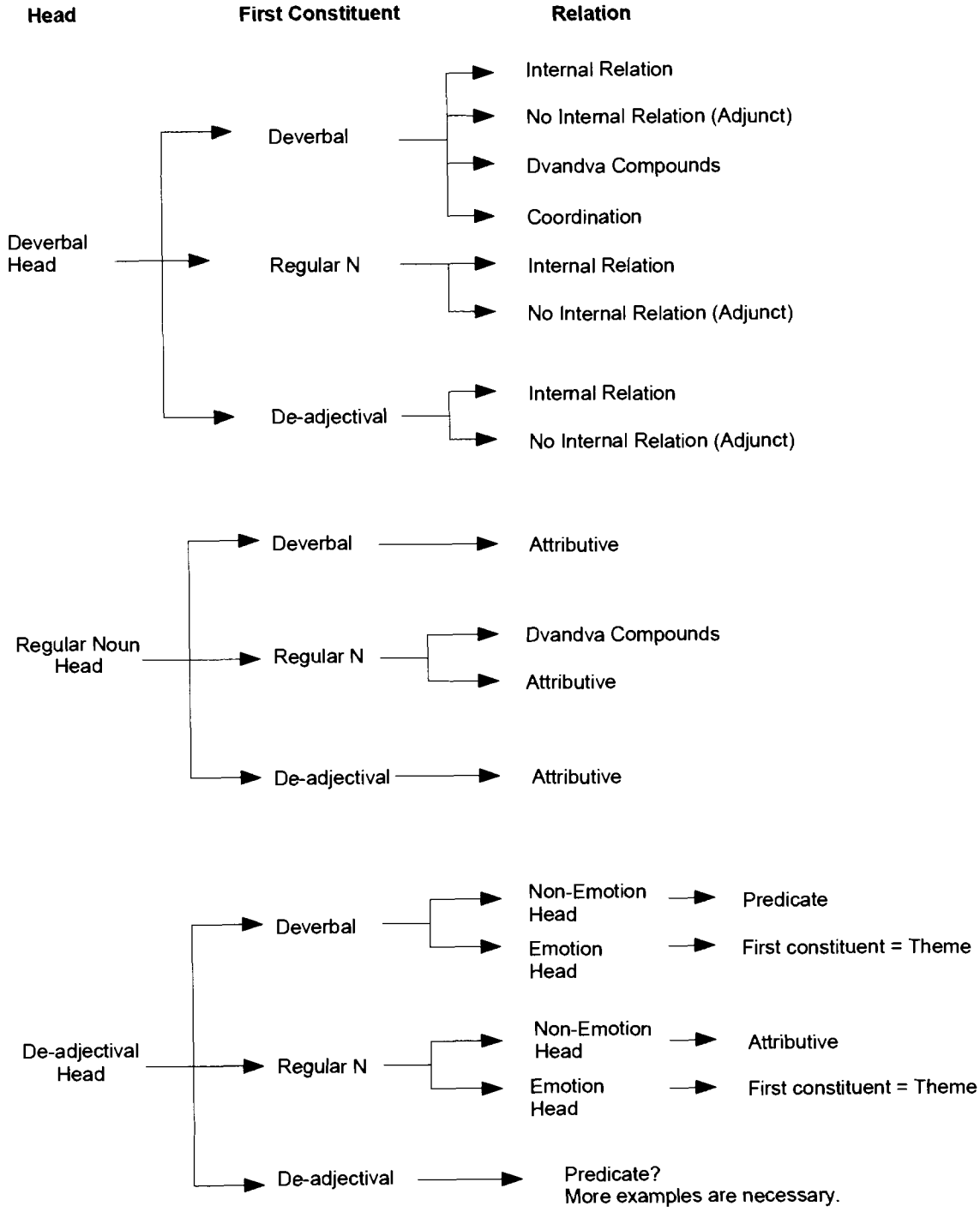
Figure 18: Classification of Deverbal Compounds



Regarding compounds whose head is a regular noun, the non-head constituent is attributive to the head regardless of the part of speech of the non-head constituent. Theoretically, dvandva compounds are possible if both the head and the non-head constituents are a regular noun. However, these were not found in the dataset analyzed in this study. De-adjectival compounds have three kinds, although this needs to be further investigated because there are fewer instances of these compounds in the corpus. Among de-adjectival compounds, if the head is an emotion word, the non-head constituent acts as the theme of the emotion word as mentioned earlier. When the non-head constituent is a deverbal noun, the head constituent functions as the predicate of the non-head constituent, whereas when the non-head constituent is a regular noun, the compounds behave the same as regular noun compounds that the constituent relation is always attributive. However, more examples of de-adjectival compounds are necessary to draw

firm conclusions on the properties of these compounds. The final classification system of four-character Sino-Japanese compounds is presented in 4.6, repeated in Figure 19.

Figure 19: Classification Summary



In this thesis, the lexical semantic properties of four-character compounds are examined using Lieber's (2004) framework. The analysis reveals that for deverbal compounds except V + V compoundsn and dvandva compounds, the constituent relation can be determined by the lexical semantic structure of the non-head constituent and the head's default complement. If the lexical semantic structure of the non-head constituent and that of the default complement of the head are compatible, the constituents tend to have an argument relation while the relation is adjunct if they are not compatible. The constituent relations of coordinated V + V and dvandva compounds cannot seem to be determined by Lexical Semantics because the co-occurrence of the two constituents is solely constrained by pragmatics. It has been argued that at least in English the constituent relations of regular noun compounds (root compounds) is constrained by pragmatics and lexical semantics does not seem to be useful in determining the constituent relation (Arehart, 2003; Lieber, 2004; Marchand, 1960; Selkirk, 1982). The present lexical semantic analysis reveals that it is also the case in Sino-Japanese compounds whose head is a regular noun. Lexical Semantics alone does not help identify possible constituent relations of regular noun compounds. A possible solution to identifying possible relations of regular noun compounds is the algorithm proposed by Arehart (2003), which uses WordNet (see 5.3.2). His algorithm was developed for English root compounds. However, he claims that his algorithm is language-independent, and can be applied to other languages, provided that there is a machine-readable lexical database like WordNet.

Regarding compounds whose head is de-adjectival, the lexical semantic structures of the non-head and the default complement of the head seem to be the same if the head is an emotion word. However, it cannot be confirmed due to the lack of examples. De-adjectival compounds whose non-head constituent is a regular noun behave as regular noun compounds. Like regular noun compounds, the constituent relation cannot be determined by Lexical Semantics alone as expected.

Chapter 6 provides a classification algorithm that was based on the classification system proposed in Chapter 4. This classification system can be applied to any projects that deal with four-character Sino-Japanese compounds because it classifies compounds

based on the properties and syntactic behaviour of these compounds. In this study, a translation algorithm for deverbal compounds is also proposed based on the lexical semantic analysis provided in Chapter 5. Translation algorithms for regular noun compounds and adjectival compounds are left for future research because lexical semantics alone cannot provide useful resources for translation. The translation algorithm suggested in Chapter 6 uses the lexical semantic analysis in Chapter 5 as well as translation patterns found in the Utiyama Corpus. Since the translation patterns used in this algorithm are language-pair-dependent, the translation algorithm cannot be applied to machine translation pairs other than Japanese-English. This translation algorithm assumes that compounds can be translated compositionally. In other words, these compounds are translated at the word level. However, there are compounds that cannot be translated compositionally. A possible solution to non-compositional compound translation is to translate these compounds at a more abstract level (Matthiessen, 2001), as discussed in Chapter 6. The concept the compounds convey needs to be translated. However, this is not possible in Machine Translation because machine translation algorithms do not understand concepts.

The focus of this thesis is the lexical semantic analysis of four-character Sino-Japanese compounds and its application to machine translation. I provided a lexical semantics analysis of these compounds in a way that it can be used in Natural Language Processing. However, it can also be applied to other areas of Linguistics. For instance, one possible application of this study is second language acquisition. Four-character Sino-Japanese compounds are not likely to be taught at the beginner or intermediate level because Sino-Japanese nouns themselves express abstract concepts and tend to be highly technical and academic. Understanding the structure of Sino-Japanese compounds may facilitate advanced learners' comprehension of written texts.

There has been debate over whether grammar should be taught (Mohamed, 2004: 160). Some researchers such as Krashen (1982) argue that grammar should be acquired naturally and should not be taught. On the other hand, White (1987) claims that grammar should be taught because some grammar points cannot be acquired naturally by just being exposed to target language input. Mohamed (2004) claims that grammar instruction can enhance and speed up learners' grammar acquisition.

There are two major approaches to teaching grammar. One is the deductive approach, in which the rules of the target language are explicitly taught. The other is the inductive approach, in which learners are encouraged to observe particular language input and find regularities. Learners formulate rules for themselves based on their observation. Although teaching grammar is a controversial issue, some researchers seem to agree that raising consciousness on particular grammar points can facilitate language learning. Tasks that are designed to raise learners' awareness of particular grammar points are called consciousness-raising (CR) tasks. The definition of CR tasks by Ellis (1997: 160) is provided below.

> A pedagogic activity where the learners are provided with L2 data in some form and required to perform some operation on or with it, the purpose of which is to arrive at an explicit understanding of some linguistic property or properties of the TL(Target Language).

Mohamed (2004) states that while CR tasks need to be further investigated, inductive CR tasks are as effective as deductive CR tasks.

There is not much literature on compound learning or vocabulary learning in general. As mentioned throughout this thesis, Sino-Japanese compounding is a productive word formation process. The number of Sino-Japanese compounds is virtually infinite. Therefore, it is not practical nor recommendable that Japanese language learners memorize each Sino-Japanese compound. However, being aware of different types of Sino-Japanese compounding may facilitate language learning such as reading comprehension since Sino-Japanese compounds are frequently used in newspaper articles and academic texts. To my knowledge, there is no literature on the acquisition of Sino-Japanese compounds. A few suggestions are as follows. The types of Sino-Japanese compounds can be taught inductively or deductively. For inductive teaching, the instructor can present some Sino-Japanese compounds and ask the students to identify the constituent relation. At the end of the exercise, the instructor can show what types of Sino-Japanese compounds are found in Japanese using a classification chart such as the one presented in Figure 19. If the instructor were to teach Sino-Japanese compounds deductively, the classification can be presented at the beginning of the exercise, and

present some examples. The suggested pedagogic techniques are intended for fairly advanced learners who can distinguish different part of speech of Sino-Japanese nouns.

REFERENCE LIST

Arehart, Mark David. (2003). *Noun Compound Semantics: Linguistics and General-Purpose Reasoning.* Unpublished Dissertation, University of Michigan.

Baker, Mark C. (1988). *Incorporation: A Theory of Grammatical Functio Changing.* Chicago and London: The University of Chicago Press.

Baker, Mark C., Aranovich, Roberto, & Golluscio, Lucia A. (2004). Two Types of Syntactic Noun Incorporation: Noun Incorporation in Mapudungun and its Typological Implications. *Language, 81*(1), 138-176.

Baldwin, Timothy. (2004). Making Sense of Japanese Relative Clause Construction, *Proceedings of the 2nd Workshop on Text Meaning and Interpretation* (pp. 49-56). Barcelona, Spain.

Baldwin, Timothy, & Tanaka, Takaaki. (2003a). Translation Selection for Japanese-English Noun-Noun Compounds, *Proceedings of Machine Translation Summit IX* (pp. 378-385). New Orleans.

Baldwin, Timothy, & Tanaka, Takaaki. (2003b). Translation Selection for Japanese-English Noun-Noun Compounds, *Proceedings of Machine Translation Summit IX* (pp. 378-385). New Orleans.

Baldwin, Timothy, & Tanaka, Takaaki. (2004). Translation by Machine of Complex Nominals: Getting it Right, *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing.* Barcelona, Spain.

Berger, Adam L., Della Pietra, Stephen A., & Della Pietra, Vincent J. (1996). A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics, 22*(1), 39-71.

Bond, Francis. (2004). *Translating the untranslatable: A solution to the problem of generating English determiners.* Stanford, California: CSLI Publications.

Breen, Jim. (1995). Building an Electronic Japanese-English Dictionary. Japanese Studies Association of Australia Conference.

Burnard, Lou. (2000). User Reference Guide for the British National Corpus: Oxford University Computing Service.

Chomsky, Noam. (1981). *Lectures on Government and Binding.* Foris: Dordrecht.

Collins, Chris. (1997). Argument Sharing in Serial Verb Construction. *Linguistic Inquiry, 28*, 461-497.

Comrie, Bernard. (1976). *Aspect.* Cambridge: Cambridge University Press.

Ellis, Rod. (1997). *SLA Research and Language Teaching.* Oxford: Oxford University Press.

Gerdts, Donna B. (1998). Incorporation. In A. Spencer & A. M. Zwicky (Eds.), *The Handbook of Morphology* (pp. 84-100). Oxford: Blackwell.

Greenberg, Joseph Harold. (1963). Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In J. H. Greenberg (Ed.), *Universals of languages.* Cambridge, Mass.: MIT Press.

Gruber, Jeffrey. (1965). *Studies in Lexical Relations.* MIT, Cambridge, MA.

Hayashi, Shiro. (1992). The Semantic Head of Two-character Sino-Japanese Words, *Computer and Language: Special Issues in Applied Linguistics* (Vol. 5). Tokyo: Meiji Shoin (in Japanese).

Hisamitsu, Toru, & Nitta, Yoshihiko. (1996). Analysis of Japanese Compound Nouns by Direct Text Scanning, *Proceedings of the 16th Conference on Computational Linguistics* (Vol. 1, pp. 550-555). Copenhagen, Denmark.

Hopper, Paul J., & Thompson, Sandra A. (1984). The Discourse Basis for Lexical Categories in Universal Grammar. *Language, 60*(4), 703-752.

Iida, Masayo. (1987). Case Assignment by Nominals in Japanese. In M. Iida, S. Wechsler & D. Zec (Eds.), *Working Papers in Grammatical Theory and Discourse Structure* (pp. 93-138). Stanford: CSLI.

Ikehara, Satoru, Shirai, Satoshi, Yokoo, Akio, & Nakaiwa, Hiromi. (1991). Toward an MT system without pre-editing-effects of new methods in ALT-J/E-, *Proceedings of the Third Machine Translation Summit (MT Summit III)* (Vol. 101-106). Washington DC, USA.

Jackendoff, Ray. (1972). *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press.

Jackendoff, Ray. (1976). Toward an Explanatory Semantic Representation. *Linguistic Inquiry, 7*, 89-150.

Jackendoff, Ray. (1983). *Semantics and Cognition*. Cambridge, MA: MIT Press.

Jackendoff, Ray. (1987a). The Status of Thematic Relations in Linguistic Theory. *Linguistic Inquiry, 18*, 369-412.

Jackendoff, Ray. (1987b). The Status of Thematic Relations in Linguistic Theory. *Linguistic Inquiry, 18*, 369-412.

Jackendoff, Ray. (1990). *Semantic Structures*. Cambridge MA: MIT Press.

Jackendoff, Ray. (1991). Parts and Boundaries. *Cognition, 41*, 9-45.

Jackendoff, Ray. (1992). Parts and Boundaries. In B. Levin & S. Pinker (Eds.), *Lexical and Conceptual Semantics* (pp. 9-46). Oxford: Blackwell Publishers.

Jackendoff, Ray. (1996a). The Architecture of the Linguistic-Spatial Interference. In P. Bloom, M. Peterson, L. Nadel & M. Garrett (Eds.), *Language and Space* (pp. 1-30). Cambridge, MA: MIT Press.

Jackendoff, Ray. (1996b). The proper treatment of measuring out, telicity, and perhaps even quantification in English. *Natural Language & Linguistic Theory, 14*, 305-354.

Jurafsky, Daniel, & Martin, James H. (2000). *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, New Jersey: Prentice Hall.

Kageyama, Taro. (1982). Word Formation in Japanese. *Lingua, 57*, 215-258.

Kageyama, Taro. (1993). *Grammar and Word Formation*: Hitsujishobo (in Japanese).

Kobayashi, Yosiyuki, Tokunaga, Takenobu, & Tanaka, Hozumi. (1994). Analysis of Japanese Compound Nouns using Collocational Information, *Proceedings of the 15th Conference on Computational Linguistics* (Vol. 2). Kyoto, Japan.

Krashen, Stephen. (1982). *Principles and Practice in Second Language Acquisition.* Oxford: Oxford University Press.

Kratzer, Angelika. (1996). Serving the External Argument from its Verb. In J. Rooryck & L. Zaring (Eds.), *Phrase Structure and the Lexicon* (pp. 109-137). Dordrecht, Boston, London: Kluwer Academic Publishers.

Kuno, Susumu. (1973). *The Structure of the Japanese Language.* Cambridge, MA: MIT Press.

Lakoff, George. (1970). *Integrity in Syntax*. New York: Holt, Rinehart, and Winston.

Levin, Beth, & Hovav, Malka Rappaport. (1995). *Unaccusativity*. Cambridge MA: MIT Press.

Li, Yafei. (1993). Structural Head and Aspectuality. *Language, 69*, 480-504.

Lieber, Rochelle. (2004). *Morphology and Lexical Semantics*. Cambridge: Cambridge University Press.

Mainichi Newspaper Co. (2001). *Mainichi Shimbun CD-ROM 2001*

Manning, Christopher D. (1993). Analyzing the Verbal Noun: Internal and External Constraints. *Japanese/Korean Linguistics, 3*, 236-253.

Marchand, Hans. (1960). *The categories and types of present-day English word-formation; a synchronic-diachronic approach.* Wiesbaden: O. Harrassowitz.

Massam, Diane. (2001). Pseudo Noun Incorporation in Niuean. *Natural Language & Linguistic Theory, 19*, 153-197.

Matsumoto, Yuji, Kitauchi, Akira, Yamashita, Tatsuo, Hirano, Yoshitaka, Matsuda, Hiroshi, & Asahara, Masayuki. (1999). Japanese Morphological Analysis System ChaSen version 2.0. Nara Institute of Science and Technology.

Matthiessen, Christian M.I.M. (2001). The Environments of Translation. In E. Steiner & C. Yallop (Eds.), *Exploring Translation and Multilingual Text Production: Beyond Content* (pp. 41-124). Berlin, New York: Mouton de Gruyter.

Miller, George A., & Fellbaum, Christiane. (1992). Semantic Networks of English. In B. P. Levin, Steven (Ed.), *Lexical and Conceptual Semantics* (pp. 197-230). Oxford: Blackwell Publishers.

Mithun, Marianne. (1984). The Evolution of Noun Incorporation. *Language, 60*, 847-894.

Miyazaki, Masahiro, Ikehara, Satoru, & Yokoo, Akio. (1993). Combined Word Retrieval for Bilingual Dictionary based on the Analysis of Compound Words. *Natural Language Processing, 34*(4), 743-754 (in Japanese).

Mohamed, Naashia. (2004). Consciousness-Raising Tasks: A Learner Perspective. *ELT Journal, 58*(3), 228-2237.

Namiki, Takayasu. (2001). Further Evidence in Support of the Righthand Head Rule in Japanese. In J. van de Weijer & T. Nishihara (Eds.), *Issues in Japanese Phonology and Morphology* (pp. 277-297). Berlin: Mouton de Gruyter.

The National Institute for the Japanese Language. (1991). *Bunrui Goi Hyo (The Classification Chart of Words)*

Nishiyama, Kunio. (1998). V-V Compounds as Serialization. *Journal of East Asian Linguistics, 7*, 175-217.

Nomura, M. (1973). Hukujiketugougo no Kouzou. *Kokuritukokugokenkyushohoukoku*(49), 72-93.

Pustejovsky, James. (1992). The Syntax of Event Structure. In B. Levin & S. Pinker (Eds.), *Lexical and Conceptual Semantics* (pp. 47-82). Oxford: Blackwell Publishers.

Pustejovsky, James. (1995). *The Generative Lexicon.* Cambridge MA: MIT Press.

Roeper, Thomas, & Siegel, Muffy. (1978). A Lexical Transformation for Verbal Compounds. *Linguistic Inquiry, 9*, 197-260.

Rose, Tony, Stevenson, Mark, & Whitehead, Miles. (2002). The Reuters Corpus volume 1 - from yesterday's news to tomorrow's language resources, *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (pp. 827-833). Las Palmas, Canary Island.

Satake, Hideo (Ed.). (2000). *Daily Concise Japanese Dictionary*. Tokyo: Sanseido.

Selkirk, Elizabeth O. (1982). *The Syntax of Words*. Cambridge, MA: MIT Press.

Shi, Chung-Kon. (1997). Two Types of Synthetic Compounds and Move-Affix in Korean. *Japanese/Korean Linguistics, 6*, 369-380.

Shibatani, Masayoshi. (1990). *The Languages of Japan.* Cambridge: Cambridge University Press.

Shimada, Atsuko, & Kordoni, Valia. (2003). Japanese "verbal Noun and *suru*" constructions, *Proceedings of the Workship on Multi-Verb Constructions.* Trondheim: Norwegian University of Science and Technology, Trondheim.

Szymanek, Bogdan. (1988). *Categories and Categorization in Morphology*. Lublin: Catholic University Press.

Takahashi, Mari. (2000). The Syntax and Morphology of Japanese Verbal Nouns: Ph D Dissertation. University of Massachusetts Amherst.

Takano, Yuji. (2003). Why Japanese is Different: Nominalization and Verbalization in Syntax and the Distribution of Arguments. *Linguistic Variation Yearbook, 3*, 179-211.

Takeuchi, Koichi, Kageura, Kyo, & Koyama, Teruo. (2003a). Building Disambiguation System for Compound Noun Analysis Based on Lexical Conceptual Structure, *Proceedings of the second International Workshop on Generative Approaches to the Lexicon, (GL2003)* (pp. 146-153). University of Geneva: Geneva, Switzerland.

Takeuchi, Koichi, Kageura, Kyo, & Koyama, Teruo. (2003b). Deverbal Compound Noun Analysis Based on Lexical Conceptual Structure, *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 2 ACL '03* (pp. 181-184): Association for Computational Linguistics.

Takeuchi, Koichi, Uchiyama, Kiyoko, Yoshioka, Masaharu, Kageura, Kyo, & Koyama, Teruo. (2001). Categorising Deverbal Nouns Based on Lexical Conceptual Structure for Analysing Japanese Compounds, *Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference* (pp. 904-909).

Tanaka, Y. (1993). The Acquisition of Knowledge Data by Analyzing Natural Language - using 4 Chinese characters -. *Natural Language Processing, 94*(9), 63-70 (in Japanese).

Trujillo, Arturo. (1999). *Translation Engines: Techniques for Machine Translation*. London: Springer.

Tsujimura, Natsuko. (1992). Licensing Nominal Clauses: the Case of Deverbal Nominals in Japanese. *Natural Language & Linguistic Theory, 10*(3), 477-522.

Tsujimura, Natsuko. (1996). *An Introduction to Japanese Linguistics*. Cambridge: Blackwell Publishing.

Uchiyama, Kiyoko, Baldwin, Timothy, & Ishizaki, Shun. (2005). Disambiguating Japanese Compound Verbs. *Computer Speech and Language, 19*, 497-512.

Uchiyama, Kiyoko, & Ishizaki, Shun. (2003). A Disambiguation Method for Japanese Compound Verbs, *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (pp. 81-88). Sapporo, Japan.

Ueno, K. (1980). *Eigo-goi no kenkyu*. Tokyo: Kenkyusha.

Utiyama, Masao, & Isahara, Hitoshi. (2003). Reliable Measures for Aligning Japanese-English News Articles and Sentences, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)* (pp. 72-79).

Vapnik, Vladimir N. (2000). *The Nature of Statistical Learning Theory 2nd Edition*. New York: Springer.

White, Lydia. (1987). Against Comprehensible Input: The Input Hypothesis and the Development of Second Language Competence. *Applied Linguistics, 8*(2), 95-110.

Wierzbicka, Anna. (1972). Semantic Primitives. *Linguistische Forschungen, 22.*

Wierzbicka, Anna. (1980). *Lingua Mentalis: The Semantics of Natural Language.* Sydney: Academic Press.

Wierzbicka, Anna. (1985). *Lexicography and Conceptual Analysis.* Ann Arbor: Benjamins.

Wierzbicka, Anna. (1988). *The Semantics of Grammar.* Amsterdam: John Benjamins.

Wierzbicka, Anna. (1996). *Semantics: Primes and Universals.* Oxford: Oxford University Press.

Williams, Edwin. (1981a). Argument Structure and Morphology. *The Linguistic Review, 1*, 81-114.

Williams, Edwin. (1981b). On the Notions 'Lexically Related' and 'Head of a Word'. *Linguistic Inquiry, 12*, 245-274.

Woodbury, Hanni. (1975). *Noun Incorporation in Onondaga.* Unpublished PhD Dissertation, Yale University, New Haven, Conn.

Yokoyama, Shoichi, & Sakuma, Kazuhiro. (1996). Analysis by Synthesis of Compound Nouns Using Semantic Features. *Mathematical Linguistics, 20*(7), 304-331 (in Japanese).