NAIST-IS-DT0361008

Doctor's Thesis

Constructing, Refining and Exploiting Rich Linguistic Resources

Sanae Fujita (Kawai)

March 24, 2009

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

Doctor's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
DOCTOR of ENGINEERING

Sanae Fujita (Kawai)

Thesis committee: Yuji Matsumoto, Professor
Kiyohiro Shikano, Professor
Kentaro Inui, Associate Professor
Francis Bond, Doctor

Constructing, Refining and Exploiting Rich Linguistic Resources*

Sanae Fujita (Kawai)

Abstract

Linguistic resources, such as corpora, thesauruses, and (machine readable) dictionaries, are important as training data and knowledge sources in Natural Language Processing (NLP). These resources can take various forms. For example, corpora can be annotated with a variety of information; part-of-speech tags, syntax trees, and word sense information; to name a few, or none at all in the case of raw corpora.

Recently, the target of natural language processing becomes deeper and deeper, shifting from surface to sense, from morphological analysis to syntactic analysis, then to semantic analysis. Therefore, importance of linguistic resources with rich syntactic and semantic information increases.

There are several methods to construct resources, for example, hand-construction, automatic-construction, and semi-automatic-construction. With the increasing the amount of machine-readable data, automatically-constructed resources have become more popular. Generally, automatically-constructed resources are easy to expand and have high topicality, but unfortunately, they are relatively shallowly analyzed and include errors. Moreover, the richer resources we try to construct, the more difficult automatic-construction becomes. On the other hand, hand-construction is both time consuming and costly, but can provide much richer resources.

In this thesis, we focus on such rich resources, and describe the methods of constructing, refining and exploiting them. First, we describe the background of our research, and the thesis outline in Chapter 1. Then, in Chapter 2, we introduce the resources related with our research; the Japanese Ontology **Goi-Taikei**, bilingual valency

*Doctor's Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT0361008, March 24, 2009.

(pattern) dictionary, the **Hinoki** Corpus, and the **Lexeed** Dictionary. These resources are related to each other, and have been constructed by hand. Then, we propose some methods to extend them effectively (Chapter 3 and 4), and prove their usefulness through several task-based evaluations (Chapter 5 and 6). Finally, in Chapter 7, we reconfirm the importance of studies on constructing, refining and exploiting rich linguistic resources.

Keywords:

natural language processing, valency dictionary, ontology, thesaurus, treebank, sensebank, alternation, parse ranking, word sense disambiguation, Japanese, English, HPSG

Acknowledgements

I am deeply grateful to Professor Yuji Matumoto of Nara Institute of Science and Technology for his patient support and supervising this thesis.

I am also deeply grateful to Dr. Francis Bond (Present affiliation: National Institute of Information and Communications Technology (NiCT)), who supervised me at NTT Communication Science Laboratory, and gave me valuable, constructive and fruitful suggestions and comments.

I am also grateful to Associate Professor Kentaro Inui and Professor Kiyohiro Shikano of Nara Institute of Science and Technology for helpful comments.

I wish to express my gratitude to NTT Communication Laboratory that gave me a chance to take a doctor's degree.

I would like to thank the other members of the NTT Natural Language Research Group, and it's predecessor Machine Translation Research Group for their supports. Especially Takaaki Tanaka (Present affiliation: NTT West Corporation), Dr. Hiromi Nakaiwa and Dr. Satoshi Shirai (Present affiliation: NTT Advanced Technology Corporation (NTT-AT)).

A part of this research was supported by the research collaborations between the NTT Communication Science Laboratory and CSLI, Kyoto University, or Melborune University. I would like to thank the members of these laboratories, especially Dr. Timothy Baldwin and Dr. Shigeko Nariyama, for helpful comments.

I would also like to thank Eric Nichols and Dr. Franklin Chang for their volunteer correction for my English.

Then, I would like to express my gratitude to people who evaluated translations, annotated corpora, refined resources, and so on. Especially Takayuki Kuribayashi, Tomoko Tanaka (Hirata) and members of NTT-AT and IR-ALT.

Finally, I wish to thank my family, my partner Takuma Kawai for his understand-

ing, my children Azumi, Ryuto and Karin for their encouraging smiles, and my parents, Toru and Miyoko Fujita, for their hearty encouragements.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Importance of Linguistic Resources

Linguistic resources, such as corpora, thesauruses, and (machine readable) dictionaries, are important as training data and knowledge sources in Natural Language Processing (NLP). Almost all NLP tools and applications use at least one or more resources. For example, morphological analyzers (part-of-speech taggers) typically use lexicons and have been improved using tagged corpora as training and test data.

Such linguistic resources are important and indispensable in every field of NLP. Figure 1.1 shows a simplified illustration of the roles of linguistic resources in NLP. Linguistic resources are the foundation, and all tools and applications (such as information retrieval, machine translation, and question answering) depend on these resources. Therefore, linguistic resources serve as an important part of all NLP tools and applications. For this reason, it is important to efficiently and accurately construct, refine and exploit linguistic resources.

1.2 Aims of This Thesis

The ultimate aim of our research is to make machines capable of understanding natural language (or make it behave as if it understood).

Most recent research is based on statistical models and/or machine learning methods. Generally, statistical methods are stronger for in-domain data, especially for fre-

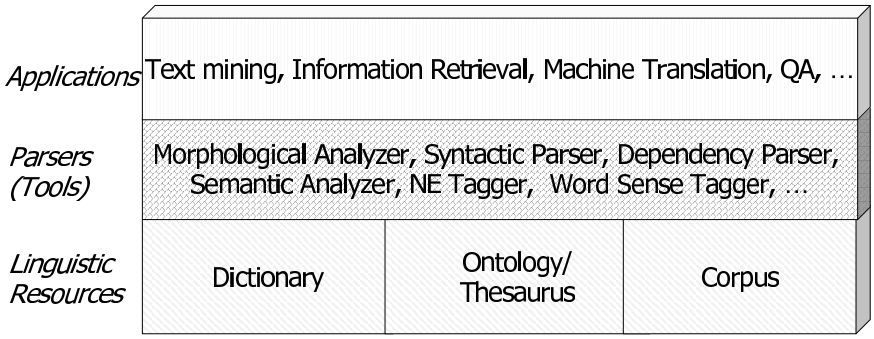


Figure 1.1: Roles of Linguistic Resources in Natural Language Processing

quent sentences, words, word senses, and so on. But they are relatively weak for out-of-domain-data, especially for infrequent sentences, words or word senses. Because language is infinitely creative and variable, we still need to handle semantics for infrequent data.

The rich resources needed to handle semantics are hard to construct. So the aims of this thesis are to construct and provide such rich resources, to prepare a framework for facilitating the construction of further rich resources. Furthermore, we also aim to show the effectivity of semantics in some NLP tasks.

1.3 Methods of Constructing Linguistic Resources

As described above, Linguistic resources, such as corpora, thesauruses, and (machine readable) dictionaries, can take various forms. For example, corpora can be annotated with a variety of information; part-of-speech tags, syntax trees, and word sense information; to name a few, or none at all in the case of raw corpora. Machine readable

dictionaries can encode many kinds of information about a lexical entry; monolingual definitions, foreign language translations, syntactic categories, case frame information, or word sense information to name a few. Then, there are several thesauruses (ontologies) constructed from several different points of view.

We also classify these resources based on methods for construction: that is, hand-constructed resources, automatically-constructed resources, and semi-automatically constructed resources.

Recently, with the increasing amount of digital data, automatically-constructed resources have become more prevalent. For example, Japanese n-gram data (Kudo and Kazawa, 2007) and Case Frames (Kawahara and Kurohashi, 2006) are automatically constructed from an enormous amount of automatically collected web data. Generally, automatically-constructed resources are easy to expand and have high topicality, but unfortunately, they are relatively shallowly analyzed and tend to be noisy (they include errors). Moreover, the richer the resources we try to construct, the more difficult automatic-construction becomes.

On the other hand, while methods of hand-construction are both time consuming and costly, they can provide more complex and reliable resources. Especially, to treat meaning, we still need to construct rich resources such as sense tagged corpora, thesauruses, either by hand or semi-automatically.

Semi-automatic methods combine the advantages of hand-construction and automatic-construction. That is, for example, we can extend resources efficiently using hand-constructed rich resources to bootstrap the process. Or, we can manually correct errors in resources which were constructed automatically at first. Generally, semi-automatic methods can provide more complex and reliable resources than fully-automatic methods at lower cost and in less time than hand-construction.

In this thesis, we show how to initially construct rich resources by hand, then expand them semi-automatically.

1.4 Contributions

In this thesis, we construct valuable and unprecedented rich resources. Our aim is for these resources to help make a breakthrough in NLP possible.

The resources we have constructed are already in use in various tasks. For example,

dictionary **Lexeed** and corpus **Hinoki** are used to evaluate several methods of word sense disambiguation; in specific, Lesk based method (Baldwin et al., 2008), a method using both syntactic and semantic features (Tanaka et al., 2007), and a method using superordinate semantic classes (See Chapter 6). The **Lexeed** dictionary also provides a basis for constructing verb dictionaries based on Lexical Conceptual Structure (LCS) by Takeuchi (2004).

We also construct the **Hinoki** treebank based on successful methods from the DELPH-IN Project (Deep Linguistic Processing with HPSG)[1], which builds and provides a shared format, tools and rigid scheme of evaluation for many different languages (currently including English, Japanese, French, Norwegian, Spainish, and so on). Thus there are several language resources (treebanks and grammars for parsing) which have the same format as the **Hinoki** treebank; therefore the **Hinoki** treebank has high inter-availability with multiple languages.

In this thesis, although we construct Japanese (and bilingual Japanese and English) resources, the proposed methods are general and not tied to any particular language pair or resources. For example, the method to expand bilingual valency patterns (in Chapter 3) inspired Hong et al. (2004) to use the same method to expand Korean-Chinese patterns. This shows that our method works for different systems and for different language pairs.

Until recently, the effectiveness of semantic information was under question. However, we showed that semantic information (in particular, superordinate semantic classes) works well for parse selection. It is especially effect when the training data size is relatively small. We thus expect that semantic information will make adaptation to new domains and language easy. This proposed method also can be expanded to other languages using **WordNet**, **EuroWordNet**, and other similar resources. (See e.g., Agirre et al. (2008)).

We also propose an easy-to-use method to estimate superordinate semantic classes. We hope that this method will provide a basic tool for estimation of superordinate semantic classes, and will be used for not only parse selection but also other NLP tasks such as Semantic Role Labeling (SRL).

[1] http://www.delph-in.net/

1.5 Thesis Outline

As described above, we believe that resources with rich information are important and useful even for statistical natural language processing. Therefore, we have been constructing various resources, such as an ontology, valency (pattern) dictionary, treebank and sensebank. In this thesis, first we introduce the features of these resources, then we investigate the usage and effectiveness of these resources.

In the following chapters, we introduce the resources related to our research (Chapter 2). Then, we propose some methods to extend them effectively (Chapter 3, 4), and prove their usefulness through several task-based evaluations (Chapter 5, 6).

In more detail, in Chapter 2, we introduce the resources we will use later on; the Japanese Ontology **Goi-Taikei** and its bilingual valency (pattern) dictionary, the **Hinoki** Treebank and Sensebank, and the **Lexeed** Dictionary. We also compare these resources with other resources.

In Chapter 3, we present a method of extending the coverage of the bilingual valency (pattern) dictionary, by assigning valency information and selectional restrictions to entries in a bilingual dictionary. The method exploits existing bilingual valency dictionaries and is based on two basic assumptions: words with similar meaning have similar subcategorization frames and selectional restrictions; and words with the same translations have similar meanings. In this chapter, we evaluate our methods through translation based evaluation and hand-evaluation.

In Chapter 4, first, we investigate the alternation features, then we present a method that uses alternation data to add new entries to an existing bilingual valency dictionary based on the features. We automatically created new valency entries using the causative/inchoative alternation data. If the existing lexicon has only one half of the alternation, then our method constructs the other half of the alternation. The created entries were hand evaluated.

In Chapter 5, we investigate the effectiveness of rich information by applying it to parse selection (ranking). In this chapter, we show that sense-based semantic features combined with ontological information are effective for parse selection.

In Chapter 6, to get the sense information automatically, we propose a method for word sense disambiguation (WSD) using superordinate semantic classes. We separate this method into two stages. At the first stage, we estimate superordinate semantic classes, then at the second stage we estimate word senses using the results of the first

stage.

Finally, we conclude with a summary and discussion of future work in Chapter 7.

Chapter 2

Resources

In this Chapter, we introduce some rich information resources which we are using or we have built, and then compare these resources with other resources. The resources which we introduce are Japanese Ontology **Goi-Taikei** (Section 2.1), Valency (Semantic Pattern) Dictionary (Section 2.2), Japanese Semantic Database **Lexeed** (Section 2.3), and Japanese Treebank **Hinoki** (Section 2.4).

We show the overview of our resources in Figure 2.1. As shown in the figure, these resources are related to each other and we are constructing deep parsers based on these resources.

2.1 Japanese Ontology: Goi-Taikei

NTT has developed the Japanese-to-English Machine Translation System: **ALT-J/E**. For **ALT-J/E**, several resources have been developed: that is, Japanese Semantic Word Dictionary, Japanese-to-English Dictionary, Valency (Semantic Pattern) Dictionary, and Japanese Ontology (**Goi-Taikei**, Ikehara et al. (1997)). In this section, we introduce **Goi-Taikei**, which is related to the other resources we introduce in the following sections.

According to Gruber (2008), ontologies are typically specified in languages that allow abstraction. To treat the huge variety of language usage, abstraction is important. Abstraction is effective to alleviate the data sparseness problem. In the case of **ALT-J/E**, to allow abstraction, several information and restrictions are written by using defined classes in the ontology.

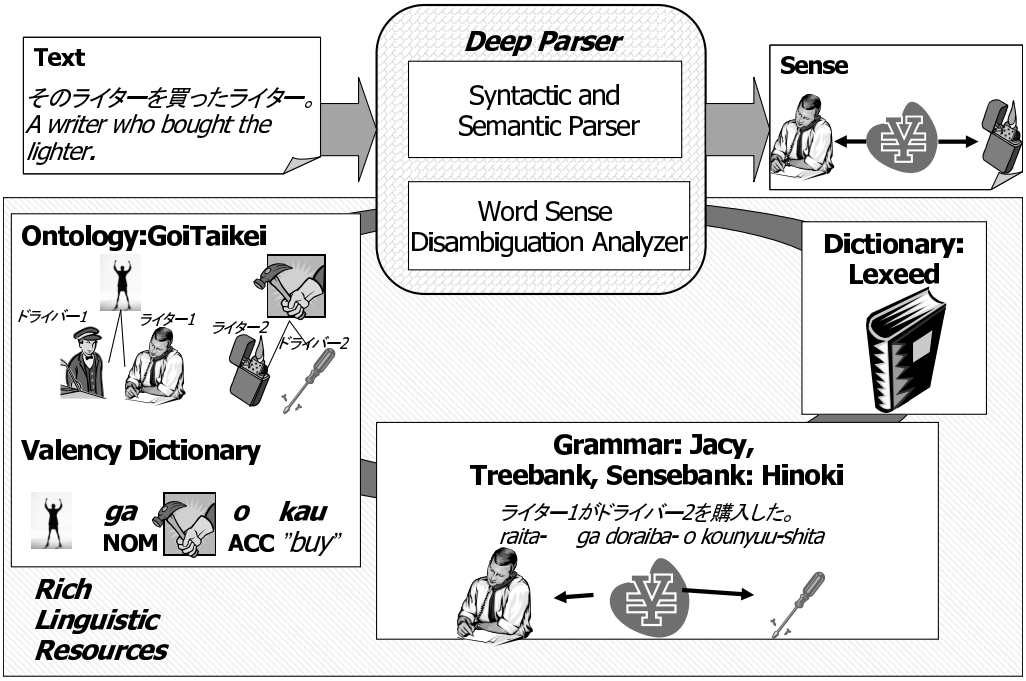


Figure 2.1: Overview of our Resources and Aims

Goi-Taikei is originally developed for the purpose of Japanese text-to-speech, but has been used extensively for Japanese-to-English machine translation. Because it has very wide coverage, it is applicable to versatile applications and systems. For other ontologies, see Section 2.1.1.

The **Goi-Taikei** Ontology consists of a hierarchy consisting of 2,710 semantic classes, defined for over 264,312 nouns, with a maximum depth of 12 (from Level 0 to Level 11). We show the top 4 levels of the **Goi-Taikei** Common Noun Ontology in Figure 2.2. The more specific classes are at deeper levels.

Table 2.1 shows the number of classes at each level of **Goi-Taikei**. It shows that Level 7 gives the largest number of semantic classes.

Because many words have multiple senses, in **Goi-Taikei**'s Word Dictionary, each word can be linked to up to 5 classes. For example, the word 牛 *ushi* "beef/cow" is linked to 2 classes; ⟨`537:beast`⟩ and ⟨`843:meat and eggs`⟩. The first category has the higher priority. On average, 1.2 classes are linked per word. We show some more

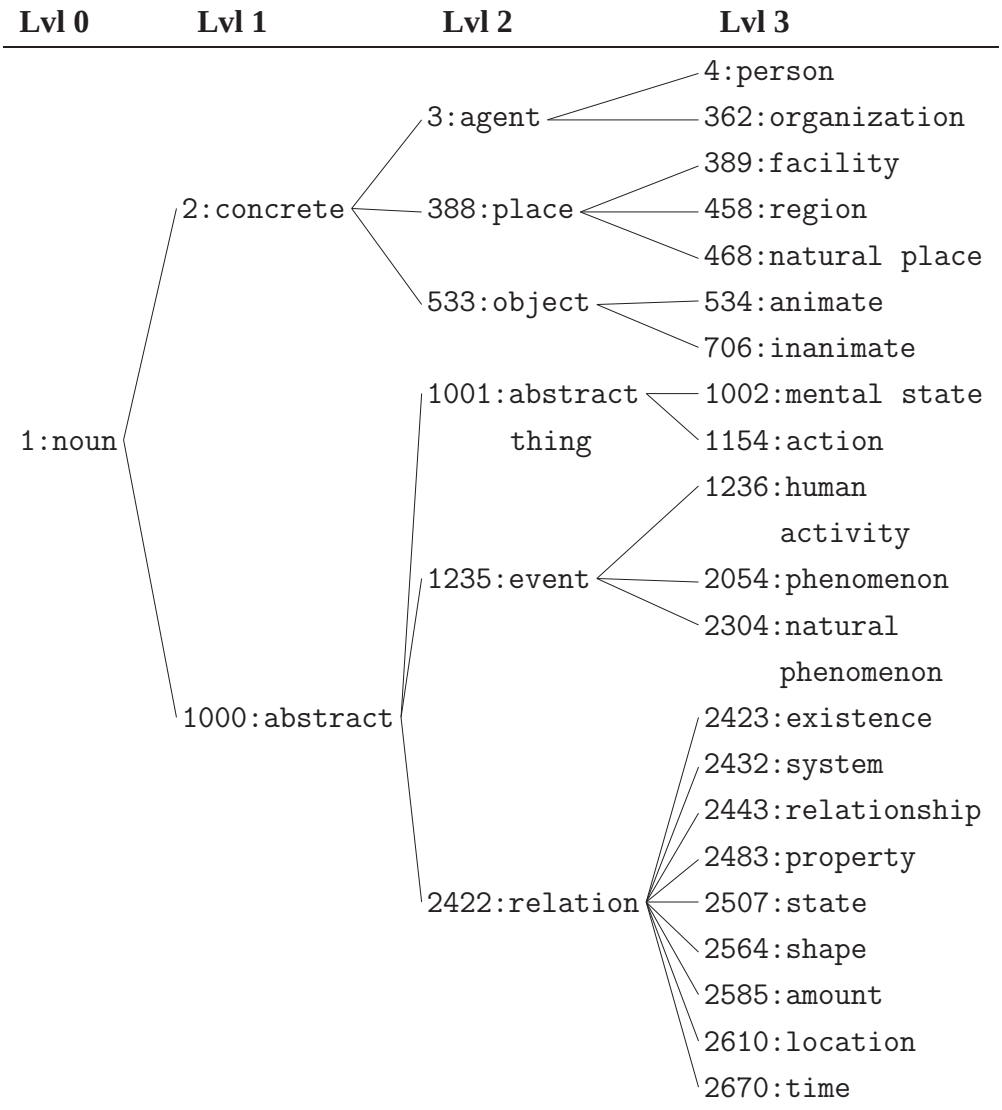


Figure 2.2: Top four levels (Lvl 0-3) of the **Goi-Taikei** Common Noun Ontology

Table 2.1: Number of Classes at each Levels of **Goi-Taikei**

Lvl	No. of Classes	Sample of Semantic Classes
0	1	⟨1:common noun⟩
1	2	⟨2:concrete⟩,⟨1000:abstract⟩
2	10	⟨3:agent⟩,⟨1001:abstract thing⟩
3	21	⟨4:person⟩,⟨1002:mental thing⟩
4	106	⟨5:human⟩,⟨1003:intellectual product⟩
5	256	⟨6:grammatical person⟩,⟨1004:study, department⟩
6	536	⟨7:1st person⟩,⟨1005:general field of study⟩
7	828	⟨8:1st person single⟩,⟨1010:commentary⟩
8	687	⟨9:1st person single male⟩,⟨1011:discussion/dispute⟩
9	211	⟨115:colleague/friend⟩,⟨1274:regret⟩
10	40	⟨116:associate/comrade/pal⟩,⟨1950:painting⟩
11	16	⟨1960:cultivation⟩

data about **Goi-Taikei**, in Appendix A.

2.1.1 Comparison with Other Resources

There are several thesauruses and ontologies besides **Goi-Taikei**. For English, there is the most popular ontology, **WordNet** (Fellbaum, 1998). In **WordNet**, words (nouns, verbs, adjectives and adverbs) are grouped into sets of cognitive synonyms called synsets, each of which represents a distinct concept. The synsets are interlinked by means of the concepts. **WordNet** is separated by POS. Based on **WordNet** 1.5, **EuroWordNet**[1] which is a multilingual database, is constructed. **EuroWordNet** includes several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian, etc.). **EuroWordNet** and **WordNet** are linked each other. Construction of Japanese **WordNet** project is now on going in NiCT (National Institute of Information and Communications Technology) (Bond et al., 2008).

[1]http://www.illc.uva.nl/EuroWordNet/

For Japanese, besides **Goi-Taikei**, there are famous ontologies, **EDR** (EDR, 1990) and **Bunrui-Goihyo** (Kokken, 2004). The **EDR** (EDR, 1990) is composed of five types of dictionaries (Word, Bilingual, Concept, Co-occurrence, and Technical Terminology), as well as the **EDR** Corpus. In the case of **EDR**, we can regard the Concept dictionary as an ontology. This ontology allows multiple inheritance, and is not necessary in the form of a tree. As with **Goi-Taikei**, **EDR** has Word dictionaries and Japanese-English Bilingual dictionary. **Goi-Taikei** is developed for the purpose of Japanese-to-English machine translation. Therefore, **Goi-Taikei**'s semantic classes are defined by comparison with English translation. But in contrast, **EDR** is a general purpose dictionary not depending on specific application and algorithms. Therefore, it's concepts are distinguished word by word.

Bunrui-Goihyo is a collection of words classified and arranged by their meanings. **Bunrui-Goihyo** is a thesaurus of 5 level tree structure. All of the words are classified into classes at the leaves of the tree.

To compare thesauruses/ontologies, we show the size and features of them in Tables 2.2 and 2.3. We used following version for Comparison: **Goi-Taikei** is Common Noun Ontology[2], **EDR** is ver. 1.5., **Bunrui-Goihyo** is revised and enlarged edition, and **WordNet** is ver. 2.0.

As shown in Table 2.2, the number of classes is larger than that of words, because the EDR Dictionary has some classes (concepts) which have no words but have definitions.

Comparison based on examples

To see the differences between those ontologies (thesauruses), we show the class(es) for the word 牛 *ushi* "beef/cow" as an example. In **Goi-Taikei**, the word 牛 *ushi* "beef/cow" is linked to two classes; ⟨537:beast⟩ and ⟨843:meat and eggs⟩ (⟨537:beast⟩ has higher priority). The hierarchic structure of ⟨537:beast⟩ is as follows: ⟨537:beast⟩ ⊆ ⟨536:animals (organism)⟩ ⊆ ⟨535:animal⟩ ⊆ ⟨534:animate⟩ ⊆ ⟨533:objects⟩ ⊆ ⟨2:concrete⟩ ⊆ ⟨1:common noun⟩. Therefore, ⟨537:beast⟩ is at Level 6. We also show the entries of **Goi-Taikei**'s Japanese-to-English transfer Dictionary, there are 4 translations *bull, cow, cattle, ox* for 牛 *ushi*. In the transfer dictionary, 牛 *ushi* is only translated as cow; 肉 *niku*, 牛肉 *gyûniku* and ビーフ *bîfu* have

[2]In **Goi-Taikei**, there are other thesauruses for verbals and proper nouns.

Table 2.2: Comparison of **Goi-Taikei** and Other Thesauruses/Ontologies: Size

Resource	Depth	Classes	Words	Words/Class	
		No.	No.	Max	Ave.
Goi-Taikei	12	2,715	300,000	93,141	110.5
Bunrui-Goihyo	5	895	96,000	1,064	112.9
EDR	around 10	410,000	194,000	10,023	37.1
WordNet	-	93,000	166,000	28	1.8

Because **WordNet** is not a tree, we can't count the depth.

Table 2.3: Comparison of **Goi-Taikei** and Other Thesauruses/Ontologies: Class and Target

Resource	Top Node(s)	Class (including largest No. of words)	Target POS
Goi-Taikei	⟨1:common noun⟩ (⟨1:proper noun⟩, ⟨1:event⟩)	⟨464:juris-diction⟩	noun (proper noun, verb,adj,adv)
Bunrui-Goihyo	⟨nominal⟩, ⟨verbal⟩, ⟨aspect⟩, ⟨other⟩	⟨person⟩	ALL
EDR	⟨concept⟩	⟨MISC⟩	ALL
WordNet	POS	⟨body parts⟩	noun,verb,adj,adv

the translation *beef*.

In the case of **Bunrui-Goihyo**, there are 2 entries which have different readings: that is 牛 *ushi* "cow" and 牛 *gyû* "beef". 牛 *ushi* "cow" is classified as ⟨mammals⟩ (⟨mammals⟩ ⊆ ⟨animals⟩ ⊆ ⟨natural objects and phenomena⟩ ⊆ ⟨nominal⟩). and 牛 *gyû* "beef" is classified as ⟨fish·meat⟩ (⟨fish·meat⟩ ⊆ ⟨food⟩ ⊆ ⟨products and tools⟩ ⊆ ⟨mammals of nominal⟩).

In the case of **EDR**, 牛 *ushi* "cow"[3] has one concept and the shortest path from the top node (⟨concept⟩) is as follows: ⟨cattle⟩ ⊆ ⟨mammals⟩ ⊆ ⟨vertebrate⟩ ⊆

[3]in **EDR**, *cattle* is used for 牛 *ushi*

⟨animals (one of a species)⟩ ⊆ ⟨animals⟩ ⊆ ⟨agent⟩ ⊆ ⟨concept⟩. Therefore, 牛 *ushi* "cow" (= ⟨cattle⟩) is at Level 6 in the shortest path. Note that there is another concept *beef*, 牛肉 *gyûniku* "beef" (= ⟨beef⟩).

Because **EDR** allows multiple inheritance, the parents of ⟨animals (one of a species)⟩ are not only ⟨animals⟩ but also ⟨species of animate life⟩, and the parents of ⟨animals⟩ are not only ⟨agent⟩ but also ⟨animate life⟩. Note that all the above concepts except ⟨cattle⟩ play the role of definition only (i.e., it corresponds to no word).

In the case of **WordNet**, *cow* has 3 senses, but the sense 3 has figurative meaning. The sense 1 of *cow* is synset ⟨cow, moo-cow⟩, the depth is 17, and it's hierarchic structure is as follows: ⟨cow, moo-cow⟩ ⊆ ⟨cattle, cows, kine, oxen, Bos taurus⟩ ⊆ ⟨bovine⟩ ⊆ ⟨bovid⟩ ⊆ ⟨ruminant⟩ ⊆ ⟨even-toed ungulate, artiodactyl,...⟩ ⊆ ⟨ungulate, hoofed mammal⟩ ⊆ ⟨placental, placental mammal,...⟩⊆ ⟨mammal, mammalian⟩ ⊆ ⟨vertebrate, craniate⟩ ⊆ ⟨chordate⟩ ⊆ ⟨animal, animate being,...⟩ ⊆ ⟨organism, being⟩ ⊆ ⟨living thing, animate thing⟩ ⊆ ⟨object, physical object⟩ ⊆ ⟨physical entity⟩ ⊆ ⟨entity⟩. Sense 2 of *cow* is synset ⟨cow⟩, the depth is 10, and it is located immediately below ⟨placental, placental mammal,...⟩.

From above examples, we can summarize as follows: hyper semantic classes like ⟨animals⟩ and concrete leaf concepts like ⟨cow⟩ are common and among different thesauruses. But they differ from each other in the hierarchical path. **Bunrui-Goihyo**'s depth is 5, the classification is the coarsest, besides that, it includes not only content words but also function words. **EDR**'s concepts have the finest granularity, and allows multiple inheritance. **WordNet**'s hierarchy is the deepest. In the case of *cow*, both biological and general classification are mixed. **Goi-Taikei** Ontology has some advantages that it has large coverage, easy-to-use tree structure, and a lot of related resources.

However, as shown above, because the thesauruses and ontologies have different features, they should be used for various roles in different systems. In this thesis, we use both **Bunrui-Goihyo** and **Goi-Taikei** for word sense disambiguation (Chapter 6), because **Bunrui-Goihyo** includes function words.Then, **WordNet** (and **EuroWordNet**) will play an important role in translation and multilingual systems.

2.2 ALT-J/E's Valency (Semantic Pattern) Dictionary

Detailed information about verb valency (subcategorization) and selectional restrictions is useful both for monolingual parsing and selection of appropriate translations in machine translation. In addition to its usage in resolving parsing ambiguities (Ikehara et al., 1991; Korhonen, 2002), verb valency information is particularly important for complicated processing such as detection and referent identification of zero pronouns (Nakaiwa and Ikehara, 1995; Yamura-Takei et al., 2002). Therefore, several dictionaries which have valency information have been constructed. In this section, we introduce the valency (pattern) dictionary from the Japanese-to-English machine translation system **ALT-J/E**. For details of other valency dictionaries, see Section 2.2.1.

The **ALT-J/E**'s valency semantic (pattern) dictionary's basic structure of a clause comes from the relationship between the main verb and nouns. The structure transfer dictionary provides this basic clause structure.

ALT-J/E provides 13,000 patterns for the valency dictionary and 3,000 patterns for the idiomatic structure dictionary. In the valency dictionary, there are, on average, 2.3 patterns for each Japanese verb. When we ignore all idiomatic and adjectival patterns there are 5,062 verbs and 11,214 valency patterns (2.2 patterns/verb).

We show some examples in Figures 2.3 and 2.4, whose head verbs are the same (指令する*shirei-suru* "order"), but the structures are different.

As shown in Figures 2.3 and 2.4, each **pattern** consists of source (Japanese) and target (English) language subcategorization information and selectional restrictions on the source side. Each argument on the Japanese side consists of head-word, a case-role, a list of case-markers and a list of selectional restrictions. There is also other information about aspectual class, verbal semantic attributes and so on, which we will not discuss here, although it is included in the patterns we create. Selectional restrictions are given as either nodes in the **Goi-Taikei** thesaurus (2,710 semantic classes; see Section 2.1) or strings. It takes an expert lexicographer an average of 30 minutes to create one pattern from scratch.

Because the valency dictionary is a transfer dictionary, the arguments associated with the predicate are linked between the two languages with indices (N1, N2, ...). Each case-slot has information such as grammatical function, case-marker, case-role (the index number gives the case-role), semantic restrictions on the filler and default order (not all the features are shown in the examples). Most arguments are NPs or PPs,

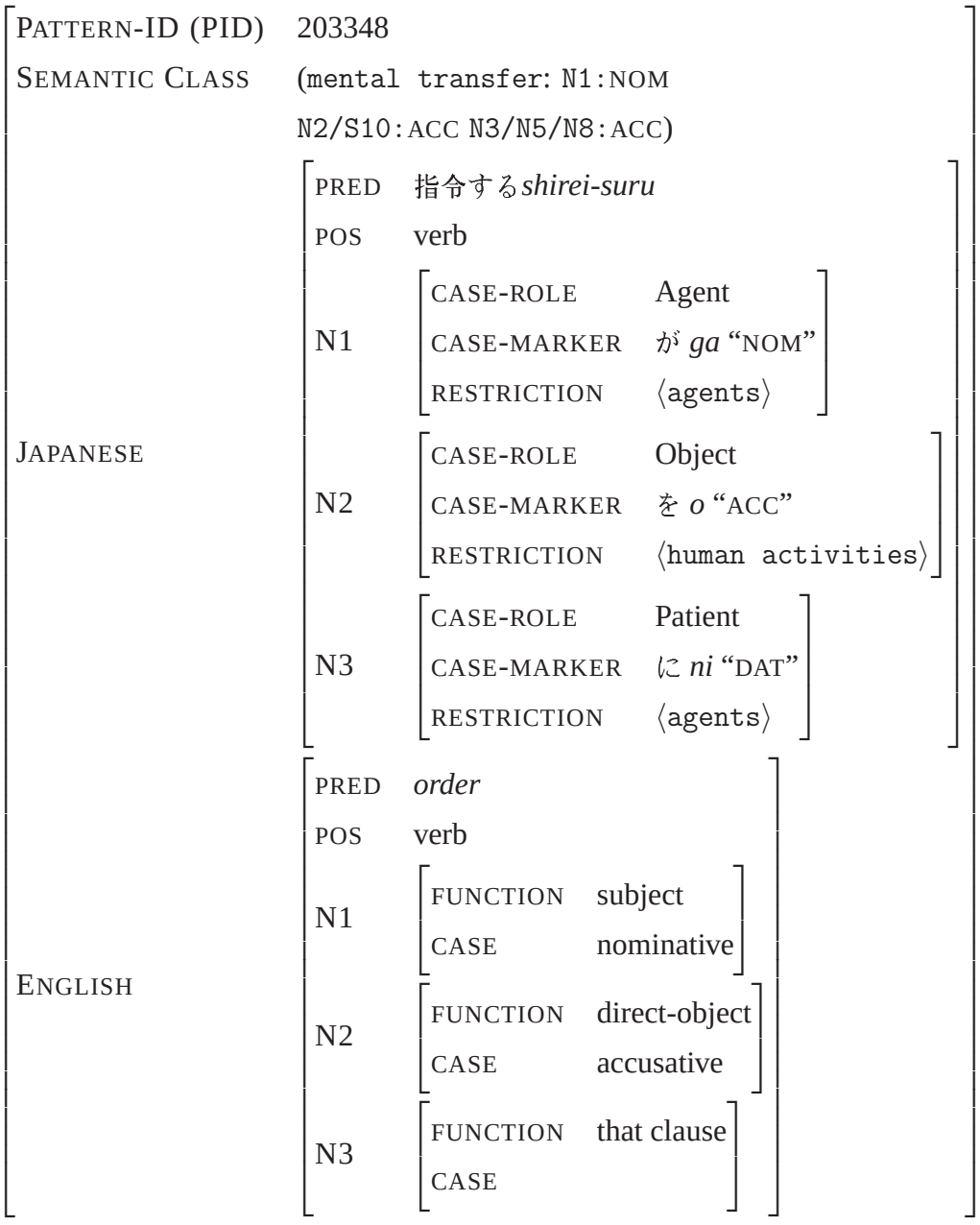


Figure 2.3: Valency (Semantic Pattern) Entry for the verb 指令する *shirei-suru* ⇔ *order* No.1 (SVOP)

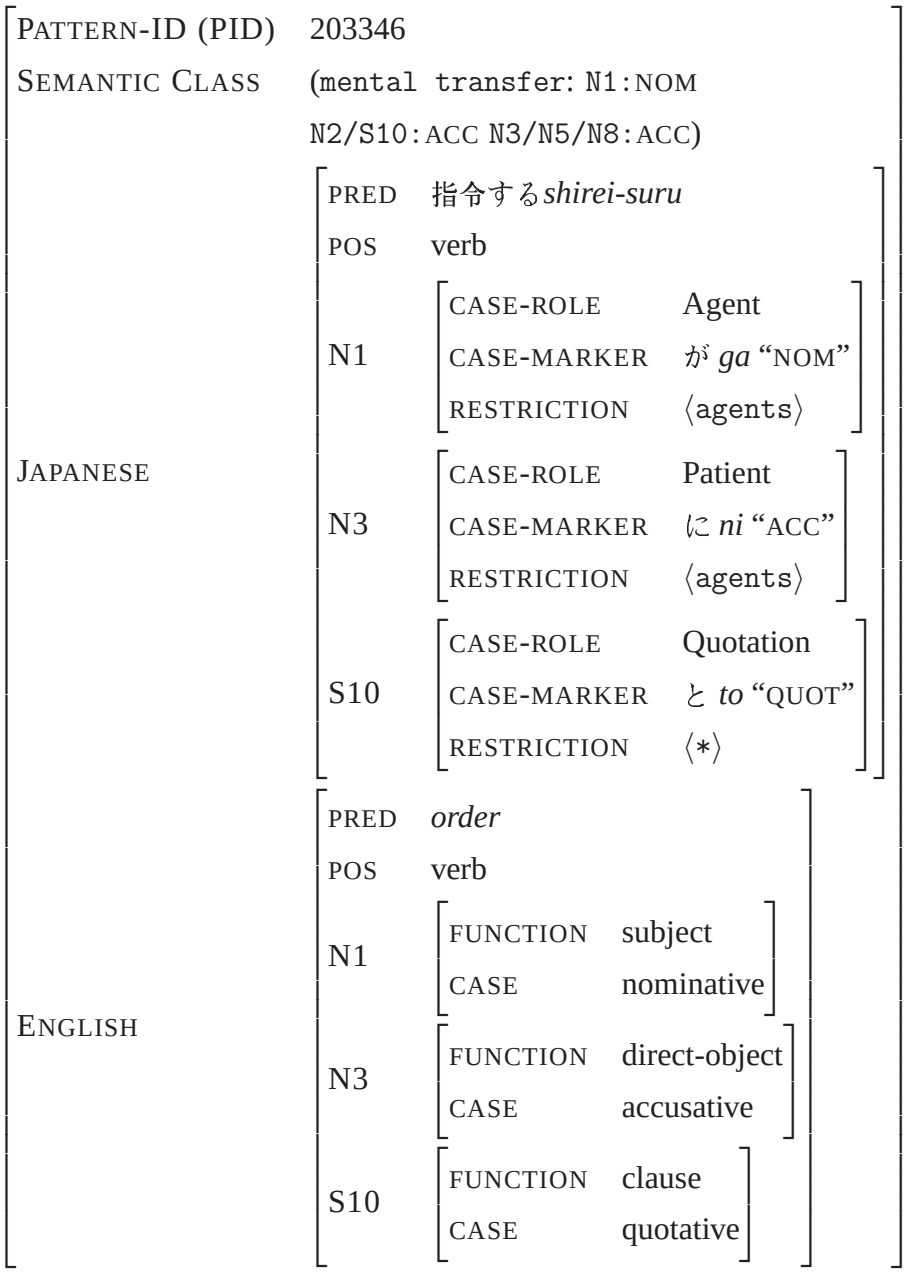


Figure 2.4: Valency (Semantic Pattern) Entry for the verb 指令する *shirei-suru* ⇔ *order* No.2 (SVPC)

but it is possible to have a sentential argument, as in Figure 2.4, where it is marked with S10. The arguments correspond between Japanese and English, thus giving the backbone of the transfer. It is possible for an argument to only appear on one side, this is useful for verbs in one language that incorporate information given as an argument in the other.

We call the combination of case-role and case-marker the **slot-type**. A verb's basic argument type is given by the combination of slot-types it allows. For example: `N1:agent`+*ga* is one slot-type, `N2:object-1`+*o* is another, and their combination is the basic transitive **frame-type**: `N1:agent`+*ga*, `N2:object-1`+*o*.

Because the Japanese slot-type combinations have not been treated as fixed case-frames, there are many minor variations, such as `N1`+*ga* with `N3`+*ni* and `N1`+*ga* with `N3`+*ni/e* which are treated as different. In most cases, these are unmotivated distinctions, and it would be advantageous to merge them, as suggested by Nomura and Muraki (1996) and Baldwin et al. (1999). This would serve to reduce the number of different frame-types.

2.2.1 Comparison with Other Resources

There are several dictionaries which have valency information, especially monolingual dictionaries. For English, there are **VerbNet** (Kipper et al., 2006) and **FrameNet** (Johnson et al., 2002) constructed based on extended verb classes (Levin, 1993). **VerbNet** groups verbs based on their semantic or syntactic features. Each verb class in **VerbNet** is described by thematic roles (29 roles:`Agent`, `Patient`, `Location`, etc.), selectional restrictions on the arguments, and frames consisting of a syntactic description and semantic predicates with a temporal function.

For example, in **VerbNet**, the following roles and restrictions are described for *order*: Agent`[+animate or +organization]` V Patient`[+animate or +organization]` Proposition,
Agent`[+animate or +organization]` V Proposition`[+oc_to_inf]`, and
Agent`[+animate or +organization]` V Proposition`[+that_comp,-tensed_that]`.
And the members of first type are followings: *allow, call, need, okay, want, permit, summon*. A few hundred of verbs are recorded in **VerbNet**.

VerbNet and **FrameNet** are part of the **SemLink** project. **SemLink** is a project whose aim is to link together different lexical resources via a set of mappings. Currently,

through the **SemLink** project, mapped resources are followings: **PropBank** (Palmer et al., 2005), **VerbNet** (Kipper et al., 2006), **FrameNet** and **WordNet** (Fellbaum, 1998).

Now, **PropBank** is a corpus which is annotated with around 4,500 argument role labels for around 3,300 verbs. For example of *order*[4]: [$_{ARG0}$ It] [$_{ARGM-DIS}$ also] [$_{rel}$ ordered] [$_{ARG1}$ P&G] [$_{ARG2-PRD}$ to produce more studies ...] [$_{ARG1}$ the plants] [$_{rel}$ ordered] * by [$_{ARG0-by}$ Florida Power]

For Japanese, **EDR** Japanese Co-occurrence Dictionary is a collection of verbs and related frames, which are extracted from corpus, and the number of records are 14,000 for 5,000 verbs.

IPAL verb/adjective dictionaries (IPA, 1987, 1994) classifies 861 Japanese basic verbs and 136 basic adjectives based on semantic and syntactic features. Each verb and adjective has case frame information, example sentences and so on. The **IPAL** basic verb list has fine information, but does not have enough coverage of valency patterns.

Both **EDR** and **IPAL** are hand-made dictionaries. There are some automatically created frame dictionaries. For Japanese, Haruno and Yamazaki (1996); Utsuro et al. (1997) extracted case frame information from syntactically annotated corpora. Kawahara and Kurohashi (2005) constructed a very large case frame dictionary from raw corpus. They parsed the raw corpus automatically, then constructed the dictionary using a highly reliable part of the parsed results. They used newspaper text (Kawahara and Kurohashi, 2005) or Web corpus (Kawahara and Kurohashi, 2006).

As above, though there are several hand-made or automatically created dictionaries, **ALT-J/E**'s valency dictionary is a one of the largest hand-made valency dictionaries. It has both fine grained and bilingual information.

2.3 Japanese Semantic Database: Lexeed

Because to build a richly informative dictionary by hand is both time consuming and costly, so we select basic words to assign hand-made rich information. We call the Japanese Semantic Database of the basic words, **Lexeed**.

In this section, we introduce **Lexeed**, see Section 2.3.1 for other dictionaries.

[4]There are 130 examples of *order* in **PropBank**.

The **Lexeed** Semantic Database of Japanese (Kasahara et al., 2004), which consists of all words with a familiarity greater than or equal to five on a scale of one to seven same as **Goi-Tokusei** (Amano and Kondo, 1998). This gives 28,000 words, divided into 46,347 different senses (the fundamental vocabulary). The examples of **Lexeed** are given in Figure 2.6 and 2.7. Each sense has a definition sentence and example sentence written using only these 28,000 familiar words (and some function words). In the case of the definition sentence of ドライバー$_1$ (Sense 1 of *doraibâ* "driver"), an originarl sentence S_1 is rewritten to S_1'. Many senses have more than one sentence in the definition: there are 81,000 defining sentences in all.

Each entry contains the word itself and its part of speech (POS) and the familiarity score along with definition and example sentences. In addition, it's also added some information by **Hinoki** project: its lexical type(s) in the grammar, links to other senses in the lexicon (such as hypernym), links to other resources (such as the **Goi-Taikei**, **WordNet** (Fellbaum, 1998), **Iwanami** (Nishio et al., 1994)). (In Figure 2.6, 2.7, all underlined features are added by the **Hinoki** project. See Section 2.4 for more details about **Hinoki**). Then we show the overview of links between the linguistic resources in Figure 2.5.

Each sense is linked with semantic classes of **Goi-Taikei** (Ikehara et al., 1997); e.g. ドライバー$_1$ (Sense 1 of *doraibâ* "driver") is linked with ⟨`942:tool`⟩, both ドライバー$_2$ and 運転手$_1$ are linked with ⟨`292:driver`⟩, ドライバー$_3$ is linked with⟨`921:plaything, sporting goods`⟩. Through the semantic classes, we can gather similer word senses; e.g. the senses which is linked with ⟨`292:driver`⟩ are followings, 運転士$_{1,2}$ *untenshi* "motorman", 運転手$_1$ *untenshu* "chauffeur", 機長$_1$ *kicho* "chief pilot", 船頭$_{1,2}$ *sendo* "boatman", テストパイロット$_1$ *tesuto pairotto* "test pilot", パイロット$_{2,3}$ *pairotto* "pilot", 飛行士$_1$ *hikoushi* "airman", ドライバー$_2$ *doraibâ* "driver", ペーパードライバー$_1$ *pêpâ doraibâ* "Sunday driver", ライダー$_1$ *raidâ* "rider", and 宇宙飛行士$_1$ *uchuuhikoushi* "astronaut". The semantic classes are principally defined for nouns (including verbal nouns), although there is some information for verbs and adjectives. All content words of **Lexeed**, including nouns, verbs and adjectives, are linked to semantic classes, as shown in Figure 2.6

In addition, **Lexeed** senses are arranged into several hierarchies (Nichols et al., 2005, 2006). These are automatically produced from the word definitions, and do not link all the senses into a single hierarchy. For example, in Figure 2.6 and 2.7, the

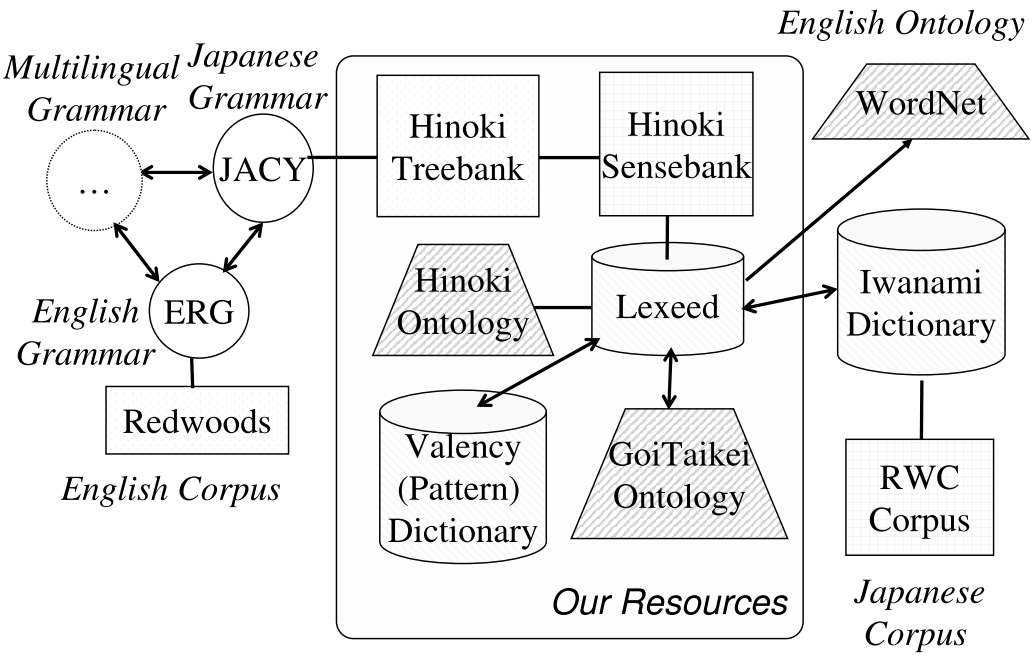


Figure 2.5: Overview of links between the Linguistic Resources

hypernym for ドライバー$_1$ is 道具$_1$ *equipment* "tool", for ドライバー$_2$ and 運転手$_1$ is 人$_1$ *hito* "person", for ドライバー$_3$ is クラブ$_2$ *kurabu* "club", and the domain for ドライバー$_3$ is ゴルフ$_1$ *gorufu* "golf".

The **Lexeed** entries can be searched via a web interface which I made[5]. Figure 2.8 shows the interface. As shown in Figure 2.8, the interface shows the several links for each word; that is hypernyms, tagged information, semantic classes (in **Goi-Taikei**), English translations, and so on.

Lexeed is used for two things. First, it defines the sense inventory used in the sensebank and ontology. Second, the definition and example sentences are used as corpora for the treebank and sensebank (See Section 2.4). Thus, as shown in Figure 2.6, 2.7, each content word in the definition and example sentences is annotated with sense tags from the same lexicon, as well as syntactic information, as part of the **Hinoki**.

[5]This interface is internal use only.

HEADWORD	ドライバー *doraibâ*	
POS	noun	Lexical-type `noun-lex`
FAMILIARITY	6.5 [1–7]	
SENSE 1	DEFINITION	S_1 ねじ まわし 。 screw turn (screwdriver) S_1' ねじ$_2$ を 差し入れ$_2$ たり 、 抜き取っ$_2$ たり する 道具$_2$ 。 A tool for inserting and removing screws .
	HYPERNYM	道具$_1$ *equipment* “tool”
	SEM. CLASS	⟨`942:tool`⟩
	IWANAMI	37515,0-0-1-0 (L ≃ R(3/3))
SENSE 2	DEFINITION	自動車$_1$ を 運転$_1$ する 人$_1$ 。 A person who drives a car
	EXAMPLE	父$_1$ は 優良$_1$ な ドライバー$_2$ として 表彰$_1$ さ$_{19}$ れ た 。 My farther was commended as a good driver
	HYPERNYM	人$_1$ *hito* “person”
	SEM. CLASS	⟨`292:driver`⟩
	IWANAMI	37515,0-0-2-0 (L ≃ R(3/3))
SENSE 3	DEFINITION	ゴルフ$_1$ で 、 遠 距離$_1$ 用 の クラブ$_3$ 。 In golf, a long-distance club. 一番$_2$ ウッド$_1$ 。 A number one wood .
	HYPERNYM	クラブ$_2$ *kurabu* “club”
	DOMAIN	ゴルフ$_1$ *gorufu* “golf”
	SEM. CLASS	⟨`921:plaything, sporting goods`⟩
	IWANAMI	37515,0-0-3-0 (L ≃ R(3/3))

Where L is an abbreviation of **Lexeed**, and R is an abbreviation of **Iwanami (RWCP)**.

Figure 2.6: Entry for the Word ドライバー *doraibâ* “driver” from **Lexeed** (with English glosses)

HEADWORD	運転手 *untenshu*	
POS	noun	Lexical-type `noun-lex`
FAMILIARITY	6.2 [1–7]	
SENSE 1	DEFINITION	電車$_1$ や 自動車$_1$ を 運転$_1$ する 人$_4$ 。 a person who drives trains and cars
	EXAMPLE	大きく$_5$ なっ たら 電車$_1$ の 運転手$_1$ に 成る$_6$ の が 夢$_3$ です 。 I dream of growing up and becoming a train driver
	HYPERNYM	人$_4$ *hito* “person”
	SEM. CLASS	⟨`292:driver`⟩
	IWANAMI	-

Figure 2.7: Entry for the Word 運転手 *untenshu* “chauffeur” from **Lexeed** (with English glosses)

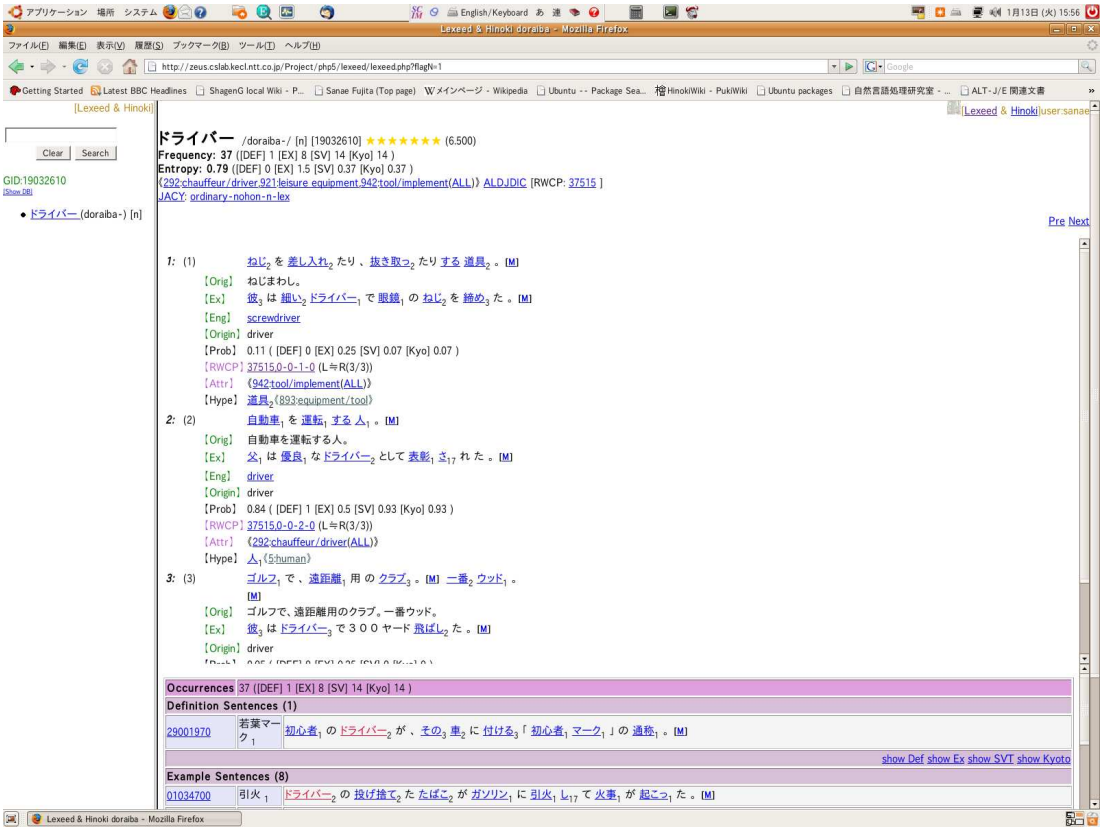


Figure 2.8: Search Interface for **Lexeed**: ドライバー *doraibā* "driver"

2.3.1 Comparison with Other Resources

Besides **Lexeed**, there are a lot of monolingual machine readable dictionaries: For Japanese, **Goi-Taikei**'s Japanese Dictionary, **JUMAN**'s Dictionary (Kyoto University, 2008), **Iwanami** (Nishio et al., 1994), and so on. **JUMAN**[6] is a Japanese morphological analysis system. Selected words **JUMAN**'s dictionary has around 30,000 selected basic words (except proper nouns) which are assigned several information by hand, but has no definitions. **Iwanami** defines the sense inventory used in the **RWCP** sensebank (originally used in the **SENSEVAL-2** competition).

To compare these dictionaries, we show their size in Table 2.4. Table 2.4 shows that the size and features of **Lexeed** is the smallest but the related sensebank is the biggest.

Table 2.4: Comparison **Lexeed** and Other Resources

Resource	Definitions	Selected	Size of Resource		Size of Sensebank
			Entries	Senses	(No. of Words)
Lexeed	Y	Y	28,000	46,000	840,000
Iwanami	Y	N	60,000	85,000	149,000
JUMAN	N	Y	30,000	-	-

To compare **Lexeed** and **Iwanami**, and also to export into **RWCP** into **Hinoki** Sensebank, we linked **Lexeed** word senses (except function words) to **Iwanami** word senses, and classified the link types by hand: that is *completely same meaning* ($=$), *almost same meaning* ($\simeq$), ***Lexeed*** *sense has wider meaning* ($\supset$), ***Lexeed*** *sense has narrower meaning* ($\subset$), *meanings of* ***Lexeed*** *and* ***Iwanami*** *overlap* (overlap), *no sense to match* ($\neg$), *others* (others)[7]. For example, each ドライバー$_{1,2,3}$ *doraibâ* "screwdriver/driver/club" (Figure 2.6) has the almost same sense ($\simeq$) in **Iwanami**. In Figure 2.6, we show this link as L $\simeq$ R (3/3): where L is an abbreviation of **Lexeed**, and R is an abbreviation of **Iwanami** (**RWCP**), and (3/3) means that the judge is done by 3 of 3 evaluators. But 運転手 *untenshu* "chauffeur" (Figure 2.7) doesn't appear in **Iwanami**.

We show the details of this classification results in Table 2.5. Though many **Lexeed** word senses are linked to many **Iwanami** senses, in Table 2.5, the link types are clas-

[6] http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[7] *others* includes *idiom*, *no entry of same index word*, etc.

sified in order of following priority: $=, \simeq, \supset, \subset, \textit{overlaps}, \neq, \textit{others}$. Table 2.5 shows that 69.4% of **Lexeed** have same or almost same **Iwanami** sense, and **Lexeed** meanings tend to be narrower than **Iwanami**.

Table 2.5: Hand-Classification of Link types of **Lexeed** and **Iwanami**

Type	No. of **Lexeed** word senses	
	No.	(%)
Lexeed $=$ **Iwanami**	318	0.8
Lexeed $\simeq$ **Iwanami**	27,620	68.6
Lexeed $\supset$ **Iwanami**	2,044	5.1
Lexeed $\subset$ **Iwanami**	6,554	16.3
Lexeed overlaps **Iwanami**	473	1.2
Lexeed $\neq$ **Iwanami**	2,895	7.2
others	376	0.9
Total	40,280	100

To compare **Lexeed** and **JUMAN**'s dictionary, we automatically linked them using lemmas and POS as pivots. Table 2.6 shows that the size of them and the entries which is existing in both dictionaries. And Table 2.7 shows that the entries which is existing in only **Lexeed** or **JUMAN**.

Table 2.6: Number of Words Existing in both **Lexeed** and **JUMAN**

POS	**JUMAN**	**Lexeed**	Same Entries	
			No.	Sample
Noun	22,419	24,634	13,228	背後 *haigo* “back”, 冷凍 *reitou* “refrigeration”
Verb	4,225	3,160	2,474	占める *shimeru* “take”, 励む *hagemu* “endeavor” 改まる *aratamaru* “be renewed”
Adj	2,350	498	344	深い *fukai* “profound”, 悔しい *kuyashii* “mortifying”
Adv	1,246	668	312	曲げて *magete* “distorted”, 文字通り *mojidori* “literally”
Others	296	583	124	だから *dakara* “so”, いろんな *ironna* “several”
Total	30,536	29,543	16,482	

Table 2.7: Number of Words Existing in Only One Dictionary (**Lexeed** or **JUMAN**)

POS	No.	Samples
JUMAN Only		
Noun	9,618	任官 *ninkan* “appointment”, 渡世 *tosei* “living”, 注視 *chushi* “gaze”
Verb	1,800	躓く *tumazuku* “stumble”, 処す *shosu* “deport”, 捻る *nejiru* “twist”
Adj	2,014	不愉快だ *fuyukai-da* “unpleasant”, 訳無い *wakenai* “easy”
Adv	935	ろくろく *rokuroku* “uselessly”, 我知らず *wareshirazu* “unconsciously”
Others	173	恐るべき *osorubeki* “fearful”, 恥ずべき *hazubeki* “shameful”
Total	14,540	
Lexeed Only		
Noun	11,406	遅番 *osoban* “late shift”, くし *kushi* “comb”, ハッカー *hakkâ* “hacker”
Verb	686	引ったくる *hittakuru* “take by force”, とろける *torokeru* “melt”
Adj	154	ふさわしい *fusawashii* “suitable”, 逞しい *takumashii* “robust”
Adv	356	誓って *chikatte* “upon my honor”, いまいち *imaichi* “lack something”
Others	459	よしよし *yoshiyoshi* “huba-huba”, 何故なら *nazenara* “because”
Total	13,061	

2.4 Japanese Treebank / Sensebank: Hinoki

In this section we describe the **Hinoki** corpus. The corpus is built from the **Lexeed** dictionary definitions, examples and newspaper text. It consists of the treebank and sensebank. The treebank uses an HPSG based Japanese grammar to encode both syntactic and semantic information. The sensebank uses **Lexeed** as a sense inventory.

The target corpus of treebank and sensebank is the same, so the **Hinoki** corpus has an important advantage over general treebank and sensebank, in that it can provide syntactic, semantic and lexical semantic information.

2.4.1 Syntactic Annotation

Syntactic annotation in **Hinoki** is *grammar based corpus annotation* done by selecting the best parse (or parses) from the all analyses derived by a broad-coverage precision grammar. The grammar is an HPSG implementation (**JACY**: Siegel and Bender, 2002), which provides a high degree of details, marking not only dependency and constituent structure but also detailed semantic relations. As the grammar is based on a monostratal theory of grammar (Head Driven Phrase Structure Grammar: HPSG, Pollard and Sag, 1994), annotation by manual disambiguation determines syntactic and semantic structure at the same time.

First, the corpus is parsed, and then the annotator selects the correct analysis (or, occasionally rejects all analyses). Selection is done through a choice of discriminants. The actual annotation process uses the same tools as the Redwoods treebank of English (Oepen et al., 2004) which was parsed by HPSG-based English grammar (ERG) (See Figure 2.1 for the relation with our resources). The system selects features that distinguish between different parses, and the annotator selects or rejects the features until only one parse is left. Using a grammar helps treebank consistency — all sentences annotated are guaranteed to have well-formed parses. The flip side to this is that any sentences which the parser cannot parse remain unannotated, at least unless we were to fall back on full manual mark-up of their analyses. The average number of decisions for each sentence is proportional to its length (around $log2$ of the number of parses). In general, even a sentence with 5,000 parses requires around 12 decisions (Tanaka et al., 2005a). Table 2.8 shows the size of target corpus of **Hinoki** project.

The **Hinoki** treebank currently consists of around 95,000 annotated definition and

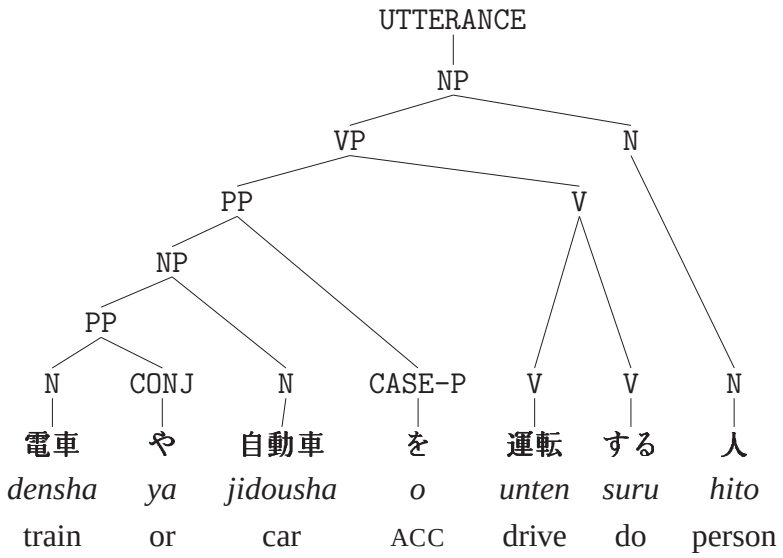
Table 2.8: Size of **Hinoki**'s Target Corpus

Corpus	Sentences	Words	Content words	Basic words
Definitions	75,000	691,072	318,181	318,181
Examples	45,000	498,977	221,224	221,224
RWCP	36,000	888,000	692,069	391,010
Kyoto	38,000	969,558	526,760	472,419

example sentences of the **Lexeed** dictionary. The definition and example sentences in the dictionary are short, around with 10 words on average, and are relatively self contained. The example sentences are relatively easy to parse. The definition sentences contain many coordinate structures and are relatively hard to parse. We are currently parsing and annotating newspaper text (Kyoto Corpus and **RWCP** Corpus) and 25% are parsed, of with around 50% are correct. We extended **JACY** by manually adding the **Lexeed** defining vocabulary, and some new rules and lexical-types, to parse dictionary sentences (Bond et al., 2004a). We still need more grammar roles and lexicon development for newspaper text. See Bond et al. (2006); Tanaka et al. (2005a) for anotation details.

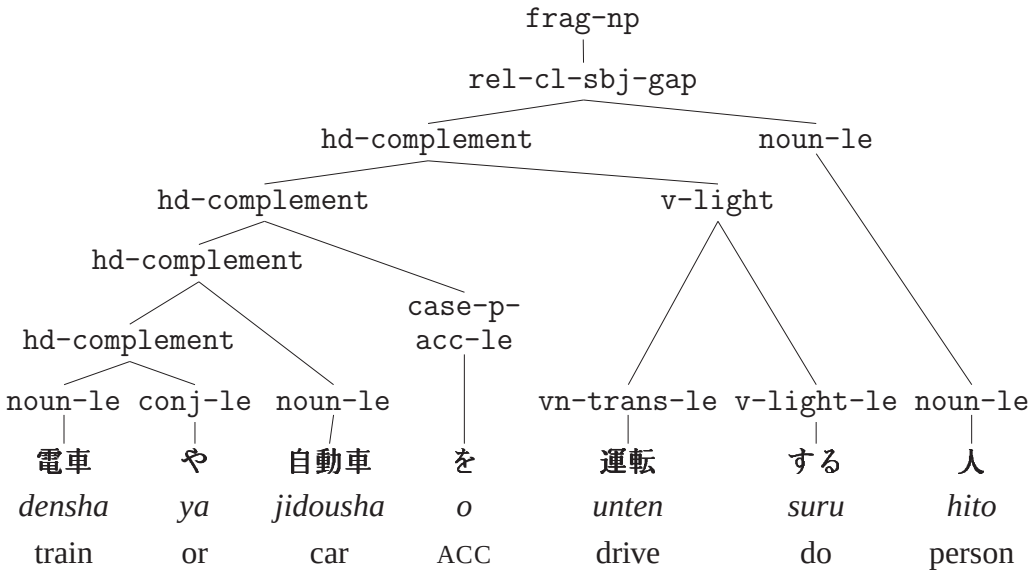
Now, we show the information included in **Hinoki** treebank. As an example, we use the definition sentence of 運転手 *untenshu* "chauffeur: somebody who drives trains and cars" (Figure 2.7).

There were 4 parses for the definition sentence. The correct parse, shown as a phrase structure tree, is shown in Figure 2.9. The two sources of ambiguity are the conjunction and the relative clause. The parser also allows the conjunction to combine 電車 *densha* "train" and 人 *hito* "person". In Japanese, relative clauses can have gapped and non-gapped readings. In the gapped reading (selected here), 人 *hito* "person" is the subject of 運転 *unten* "drive". In the non-gapped reading there is some underspecified relation between the modifee and the verb phrase. This is similar to the difference in the two readings of *the day he knew* in English: "the day that he knew about" (gapped) vs "the day on which he knew (something)" (non-gapped). Such semantic ambiguity is resolved by selecting the correct derivation tree that includes the rules applied in building the tree (Figure 2.10).



運転手$_1$ “chauffeur”: “a person who drives a train or car”

Figure 2.9: Syntactic View of the Definition of 運転手$_1$ *untenshu* “chauffeur”



運転手$_1$ “chauffeur”: “a person who drives a train or car”

Figure 2.10: Derivation Tree of the Definition of 運転手$_1$ *untenshu* “chauffeur”
Phrasal nodes are labeled with identifiers of grammar rules, and (pre-terminal) lexical nodes with class names for types of lexical entries.

The semantic representation is Minimal Recursion Semantics (MRS) (Copestake et al., 2005). We simplify this into a dependency representation, further abstracting away from quantification, as shown in Figure 2.11. We can interpret the meaning of Figure 2.11 as Figure 2.12. One of the advantages of the HPSG sign is that it contains all this information, making it possible to extract the particular view needed. In order to make linking to other resources, such as the sense annotation, easier predicates are labeled with pointers back to their position in the original surface string. For example, the predicate `densha_n_1` links to the surface characters between positions 0 and 3:電車.

2.4.2 Semantic Annotation

The lexical semantic annotation uses the sense inventory from **Lexeed** (See Section 2.3). All words in the fundamental vocabulary are tagged with their sense. For example, the word 夢 *yume* "dream, hope" (of example sentence in Figure 2.7) is tagged as sense 3 in the example sentence, with the meaning "hope, wish".

Because the **Lexeed** word senses are linked to **Goi-Taikei** semantic classes, we can get the semantic classes through the word senses. In the case of 夢$_3$ *yume* "hope, wish", we can get not ⟨`1252:dream`⟩ but ⟨`1363:hope/wish`⟩ as a semantic class.

As above, the **Hinoki** Corpus (Bond et al., 2006) consists of dictionary definition and example sentences (from **Lexeed**) and newspaper corpora (taken from the Kyoto Corpus (Kurohashi and Nagao, 2003) and **RWCP** corpus.

Lexeed definition and example sentences consist of basic words and function words only, i.e., it is self-contained. Therefore, all content words have headwords in **Lexeed**, and all word senses appear in at least in one example sentence.

Both newspaper corpora were taken from the Mainichi Daily News. **RWCP** was the text used for the Japanese dictionary task in **SENSEVAL-2** (Shirai, 2002) (which has the Senseval sense annotation). And the Kyoto Corpus is marked up with dependency analysis (Kurohashi and Nagao, 2003). We chose these corpora so that we can compare our annotation with existing annotation. Both these corpora were already word segmented and part-of-speech annotated.

Table 2.9 shows the size of the **Hinoki** sensebank. Words were segmented and mor-

```
e2:unknown<0:13>[ARG x5:_hito_n]
x7:densha_n_1<0:3>[]
x12:_jidousha_n<4:7>[]
x13:_ya_p_conj<0:4>[LIDX x7:_densha_n_1,
                    RIDX x12:_jidousha_n]
e23:_unten_s_2<8:10>[ARG1 x5:_hito_n]
e23:_unten_s_2<8:10>[ARG2 x13:_ya_p_conj]
```

Figure 2.11: Simplified Dependency View (MRS) of the Definition of 運転手$_1$ *untenshu* “chauffeur”

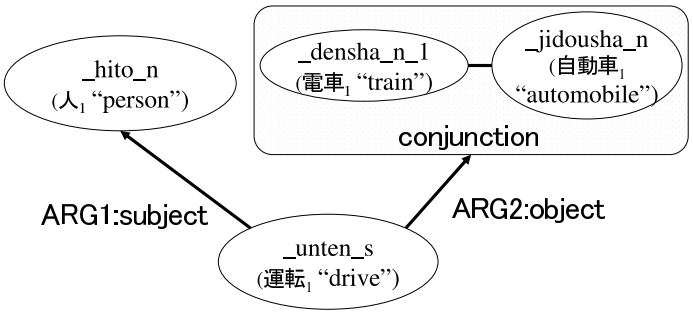


Figure 2.12: Interpretation for MRS of 運転手$_1$ *untenshu* “chauffeur” (Figure 2.11)

phologically tagged using ChaSen[8]. The Kyoto Corpus is originally morphologically analyzed using Juman[9]. We converted the tags into the IPA tagset used in ChaSen.

Table 2.9: Size of **Hinoki** Sensebank

Corpus	Sentences	Words	Content words	Basic words	Mono-semous %
Definitions	75,000	691,072	318,181	318,181	31.7
Examples	45,000	498,977	221,224	221,224	30.5
RWCP	36,000	888,000	692,069	391,010	39.3
Kyoto	38,000	969,558	526,760	472,419	36.3

Now, we are adding annotation of **Goi-Taikei**'s class over **Hinoki** corpus, we've finished the first half of the same part of Kyoto Corpus. We show the example in Table 2.10.

2.4.3 Comparison with Other Resources

Treebank

There are some morphological and syntactical annotated corpora. For English, there are **EDR** English Corpus (120,000 sentences), Penn Treebank (Marcus et al., 1994) (including Wall Street Journal, The Brown Corpus and so on), **PropBank** (Palmer et al., 2005) (85,000 sentences), which were added predicate-argument relations to the syntactic trees of the Penn Treebank. **PropBank** is also being mapped to **VerbNet** and **FrameNet** as part of **SemLink**: Mapping together **PropBank**/**VerbNet**/**FrameNet**. Here is an example of **PropBank**:
[$_{ARG1}$ Commonwealth Edison Co.] was [$_{rel}$ ordered] *-1 [$_{ARG2-PRD}$ *-2 to refund about $ 250 million *U* to its current and former ratepayers for illegal rates collected * for cost overruns on a nuclear power plant].

For Japanese, there are Kyoto Text Corpus[10], NAIST Text Corpus (Iida et al.,

[8] `http://chasen-legacy.sourceforge.jp/`

[9] `http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html`

[10] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html

2007)[11], **EDR** Japanese Corpus (200,000 sentences), and balanced corpus are now constructing by KOTONOHA project[12]. Both Kyoto Corpus and NAIST Corpus are annotated to the same 40,000 sentences of newspaper; Mainichi 1995. About 5,000 sentences of Kyoto Corpus and all of NAIST Corpus have information about predicate-argument and co-referential relations, but there target case is surface case only. Then in the case of NAIST Text Corpus, the target case markers to annotate is following: が *ga*, を *o*, に *ni*.

As above, **Hinoki** has not only predicate-argument information but also semantic information (MRS). But **Hinoki** has no information about co-referential relations. Howevere we tag the same text, we can also use lexical semantic information.

Sensebank

Several semantically tagged corpora are provided in the competitions for word sense disambiguation; Senseval-1,2,3 and SemEval. In the case of Japanese, for **SENSEVAL-2**'s Japanese dictionary task, **RWCP** corpus was provided. **RWCP** corpus are defined word senses according to **Iwanami** Japanese dictionary (See Section 2.3.1). Of 888,000 words (Table 2.9), 148,558 words are tagged with **Iwanami** senses. So **Hinoki** sensebank has much bigger coverage over **RWCP**. However, annotated resources are still lacking in several genres, therefore we need to investigate an efficient way to construct resources.

[11] http://cl.naist.jp/nldata/corpus/

[12] http://www.kokken.go.jp/en/research_projects/kotonoha/kotonoha/

Table 2.10: Simplified Example of semantic annotated **Hinoki** corpus

Word	Semantic Class	Proper Noun	
		Type	Class
村山$_p$ *Murayama* “family name”	⟨47:men and women/ gender⟩	B	⟨260:politician
富市$_p$ *Tomiichi* “first name”	⟨48:male/man⟩	I	⟨260:politician⟩
首相$_1$ *shusyo* “Prime Minister”	⟨260:politician⟩	E	⟨260:politician⟩
は *ha* “TOP”	-		
年頭 *nentou* “beginning of a year”	⟨2707:beginning⟩		
にあたり *niatari* “when”	-		
首相$_1$ *shusyo* “Prime Minister”	⟨260:politician⟩		
官邸$_1$ *kantei* “official residence”	⟨447:housing (Others)⟩		
で *de* “LOC”	-		
内閣$_1$ *naikaku* “cabinet”	⟨364:executive agency/ administrative body⟩		
記者$_1$ *kisha* “reporter”	⟨245:journalist⟩		
会$_4$ *kai* “meeting”	⟨378:society⟩		
と *to* “with”	-		
二十八 *28* “28”	⟨2586:number⟩		
日$_4$ *nichi* “day”	⟨2682:day⟩		
会見$_1$ *kaiken* “interview”	⟨1695:meeting⟩		
し *shi* “did”	⟨2050:execution⟩		
、 , “punctuation”	-		
社会党 *shakaitou* “Socialist Party”	⟨380:political party⟩	B	⟨380:political party⟩
の *no* “of”	-		
新$_5$ *shin* “new”	⟨2710:old and new OR slow and fast⟩	B	⟨382:faction/sect⟩
民主$_p$ *minshu* “democracy”	⟨1014:beliefs/ principle⟩	I	⟨382:faction/sect⟩
連合$_1$ *rengou* “union”	⟨2229:union⟩	E	⟨382:faction/sect⟩
所属$_1$ *shozoku* “belong”	⟨2475:dependence⟩		
議員$_1$ *giin* “cabinet member”	⟨260:politician⟩		
の *no* “of”	-		
...	...	...	...

The Sample Sentence is: 村山富市首相は年頭にあたり首相官邸で内閣記者会と二十八日会見し、
社会党の新民主連合所属議員の...

Chapter 3

Extending the Coverage of a Valency Dictionary

In this chapter, we present an efficient method of assigning valency information and selectional restrictions to entries in a bilingual dictionary, based on information in an existing valency dictionary. The method is based on two assumptions: words with similar meaning have similar subcategorization frames and selectional restrictions; and words with the same translations have similar meanings. Based on these assumptions, new valency entries are constructed for words in a plain bilingual dictionary, using entries with similar source-language meaning and the same target-language translations. We evaluate the effects of various measures of semantic similarity[1].

3.1 Introduction

As mentioned in Section 2.2, detailed information about verb valency (subcategorization) and selectional restrictions is important for natural language processing tasks such as monolingual parsing, accurate rule-based machine translation and automatic summarization. However, this information is not encoded in normal human-readable dictionaries, and is hard to enter manually. Therefore, for most of the language pairs, the lack of suitable language resources is a severe problem. Even when word-lists or

[1]First we reported in Fujita and Bond (2002a) about this method. Then we revised it in Fujita and Bond (2002b) and Fujita and Bond (2004c). Then put them together into a journal, Fujita and Bond (2007).

simple bilingual dictionaries exist, it is rare for them to include detailed information about the syntax and meaning of words.

Although great progress has been made in learning statistical models from annotated corpora, most commercial machine translation systems rely on detailed information compiled in lexicons. They are typically hand-built. However, adding such detailed information to dictionaries is both time consuming and costly.

In this chapter we present a method of adding new entries (we call these valency patterns, or patterns) to a bilingual valency dictionary. New patterns are based on existing patterns, so they have the same amount of detailed information. The method bootstraps from an initial hand-built lexicon, and allows new patterns to be added cheaply and effectively. Although we will use Japanese and English as examples, the algorithm is not tied to any particular language pair or dictionary. The core idea is to add new patterns to the valency dictionary by using Japanese-English pairs from a plain bilingual dictionary (without detailed information about valency or selectional restrictions), and to build new patterns for them based on existing patterns. We show the basic method in an illustration, Figure 3.1. As shown in this figure, because there are relatively large plain bilingual dictionaries, we extend hand-build bilingual valency dictionary using such bilingual dictionaries.

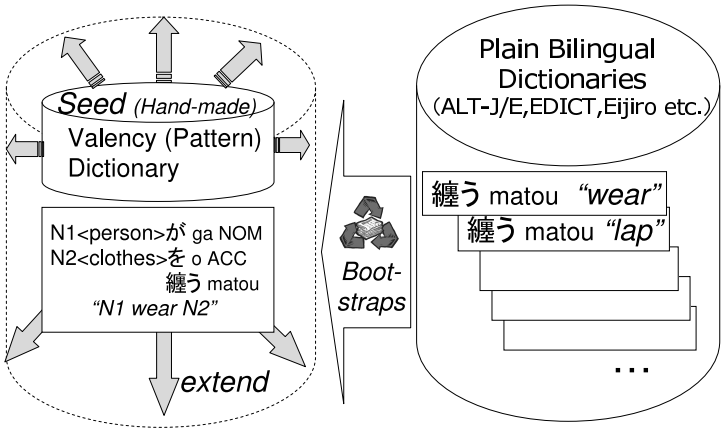


Figure 3.1: Point of the Idea for Extending the Coverage of a Valency Dictionary

3.1.1 The Number of Valency Patterns Required

Shirai (1999) estimates that at least 27,000 valency patterns are needed to cover around 80% of Japanese verbs in a typical newspaper. Various methods of creating detailed patterns have been proposed, such as the extraction of candidates from corpora (Li and Abe, 1998; Haruno and Yamazaki, 1996; Manning, 1993; Utsuro et al., 1997; Kawahara and Kurohashi, 2001), the automatic and semi-automatic induction of semantic restrictions (Akiba et al., 1995, 2000) and hand-construction (Dorr, 1997; Johnson et al., 2002; Erk et al., 2003).

However, the quality of automatically constructed monolingual patterns is still far from that of hand-constructed resources. Further, large-scale bilingual resources are still rare for most language pairs, so that it is hard to automatically build bilingual patterns.

Our work differs from corpus-based work such as Manning (1993) or Kawahara and Kurohashi (2001; 2005) in that we are using existing lexical resources rather than a corpus, and we are obtaining selectional restrictions as well as subcategorization frames. Our method is also applicable to rare words, as long as we can find them in a bilingual dictionary, and know the English translation. It does not, however, learn new frames from usage examples.

Our method adds new patterns by leveraging existing knowledge in the system dictionaries. We illustrate the method with examples of building a Japanese-English lexicon, but there is nothing in the method itself that is language specific. The basic idea is to add new patterns to the pattern dictionary by using Japanese-English pairs from a plain bilingual dictionary (without detailed information about valency or selectional restrictions), and build new patterns for them based on existing patterns.

We present both fully automatic and semi-automatic implementations in this thesis. However, even the semi-automatic implementation does not rely on detailed knowledge of the system dictionaries by the analyst. Our method is similar in principle to Ikehara et al. (1995) who add useful information to a user dictionary by comparing input word pairs to existing patterns in the system dictionary.

3.1.2 Coverage of Original Valency Patterns

To test the useful range of our algorithm, we evaluated the coverage of **ALT-J/E**'s valency dictionary (See Section 2.2) on 9 years of Japanese newspaper text (6 years of Mainichi and 3 years of Nikkei)[2] (see Table 3.1 and Figure 3.2; we graphed Table 3.1 as Figure 3.2). They consist of about 309,000,000 words. The coverage of tokens is high (92.5%), but many infrequent verb types are missing from our system (over 62% of verb types have no pattern). The verbs which have no pattern appear on average 50 times. Many of these infrequent words are still quite familiar to native speakers. For example ちょろまかす *choromakasu* "steal", 馴れ合う *nareau* "conspire" and 腫れ上がる *hareagaru* "swell up" appear only once in the 9 years of newspaper data, but they are very familiar words. The existing dictionary's coverage is good, but not complete.

We measure familiarity using **Goi-Tokusei** (Amano and Kondo, 1999) and web frequencies. **Goi-Tokusei** lists word familiarities for Japanese: The word-familiarity is a subjective rating score which represents how familiar people feel to a particular word. The rating scale is a 7-point scale (1:unfamiliar – 7:familiar): 94% of adults know the words those familiarities are higher than 5 (Amano and Kondo, 1998). The words cited above all have word familiarities greater than 5. From web pages, we can find 27,100 pages for the base form of ちょろまかす *choromakasu* "steal", 102,000 for 馴れ合う *nareau* "conspire" and 55,700 for 腫れ上がる *hareagaru* "swell up": these are widely used words.[3]

In the newspaper text, there is an average of 3.1 verbs for each sentence. Therefore, one verb in every 5 sentences has no pattern. In order to reduce the number of unknown verbs to one in 10 sentences, we need to add valency information for 2,647 verbs.

Table 3.1: Cover Ratio for Japanese Newspapers (9 years)

In lexicon	No. of Types (%)		No. of Tokens (%)	
Japanese exists	4,997	37.6	24,656,590	92.5
No pattern	8,304	62.4	2,000,710	7.5
Total	13,301	100.0	26,657,300	100.0

[2]Mainichi '91, '92, '94, '95, '99, 2000 and Nikkei '95, '96, '98.

[3]http://www.google.co.jp/, searched on 2006-03-22.

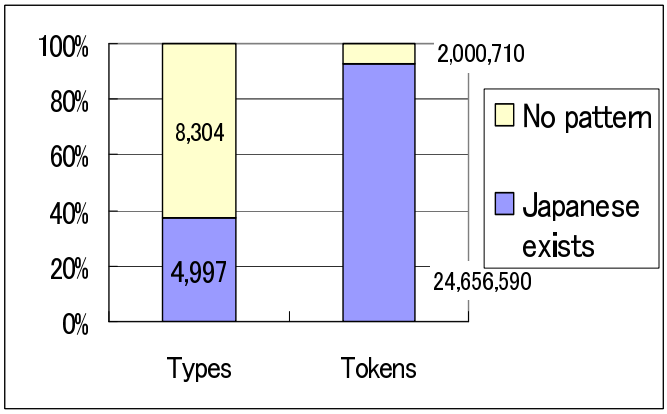


Figure 3.2: Graph of Cover Ratio for Japanese Newspapers (9 years)

3.1.3 Utility of Valency Information

In order to demonstrate the utility of the valency information, we give an example of a sentence translated with the system default information (basically a choice between transitive and intransitive), and the full valency information in (1).[4] The verb is 頼む *tanomu* "ask" [NP-*ga* NP-*ni* Cl-*to* V], which takes a clause complement. Without the valency information the translation is incomprehensible: the clause complement is misinterpreted, the zero-pronoun is not resolved and the English to-infinitive is not produced.

(1) 太郎 は 友達 に 話さ ない ように 頼んだ
Tarō wa tomodachi ni hanasa nai, yōni tanonda
Tarou TOP friend DAT talk not QUOT asked
"Tarou asked his friend not to talk."

with: `Taro asked his friend not to talk.`

without: `As Taro did not talk to his friend, * asked.`

[4]We use the following abbreviations: NOM: nominative postposition; ACC: accusative postposition; DAT: dative postposition; LOC: locative postposition; TOP: topic postposition; QUOT: quotative postposition; REC: reciprocal postposition; NP: noun phrase; Cl: clause; V: verb. The sentence is translated using **ALT-J/E** Ikehara et al. (1991).

In general, translation tends to simplify text, because the target language will not be able to represent exactly the same shades of meaning as the source text: there is some semantic loss. Therefore, in many cases, a single target language entry is the translation of multiple similar source patterns.

For example, there are 23 Japanese predicates linked to the English entry *report* in the valency dictionary used by the Japanese-to-English machine translation system **ALT-J/E** Ikehara et al. (1991). Six of these have the same frame-type as that shown in Figure 3.3. Five patterns have the frame-type shown in Figure 3.4. Three more link to a variation of that in Figure 3.3 where `N3`+*ni* is replaced by `N3`+*ni/e/made*. Collapsing such minor variations, 11 are of one type, 7 of the other, and only 2 are genuinely different. Therefore, in order to make new frames for predicates that translate into English *report*, we need to add only two patterns, one of the types in Figure 3.3 and one in Figure 3.4. Ideally, we should merge these into a single **alternation** (Levin, 1993) and link to that as suggested in Baldwin et al. (1999).

The ultimate aim of this research is to identify what kind of information is most effective in the creation of lexical patterns. In particular we wish to discover what is the minimal amount of information necessary to reliably create new patterns. Dillinger (2001) criticized previous research presented on lexical construction as paying "more attention to theoretical issues than to establishing effective processes for dictionary development". We address both theory and practice through rigorous evaluation of various methods with an emphasis on producing usable patterns as the final result.

In the following sections, first we propose the basic method of creating and refining new patterns (Section 3.2). Then we add two refinements: creating multiple patterns simultaneously using information about alternations, and merging similar patterns (Section 3.2.4). We are able to create high-quality patterns cheaply. We test the various filters to improve the quality of patterns and to make the creation of patterns more efficient. The evaluation (Section 3.3) is done with both a translation task-based evaluation and a direct evaluation by lexicographers. We then discuss the results and suggest a refined method, compare our research with other approaches and discuss further work (Section 3.4). Finally, we conclude (Section 3.5).

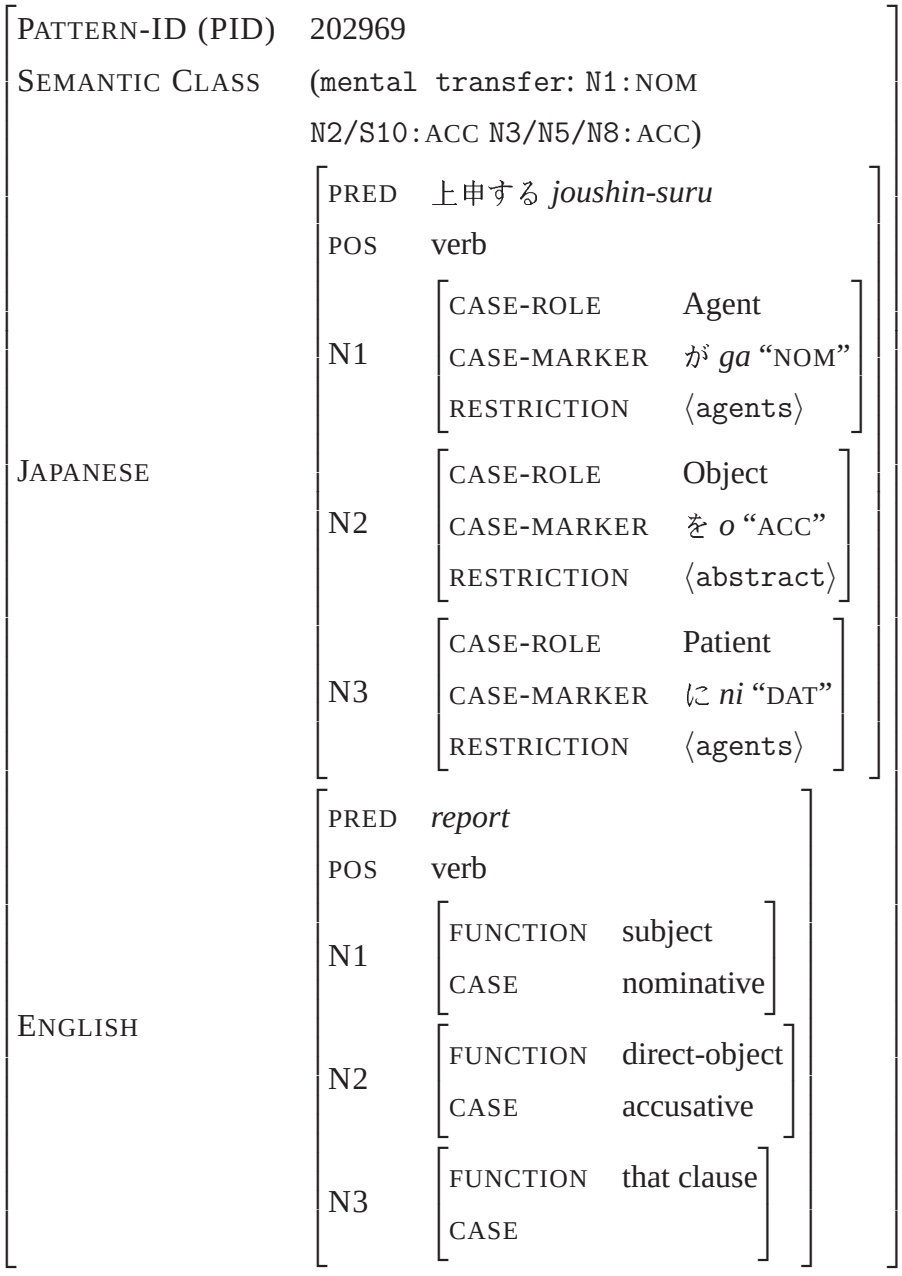


Figure 3.3: Valency (Semantic Pattern) Entry for the verb 上申する *joushin-suru* ⇔ *report* No.1 (SVOP)

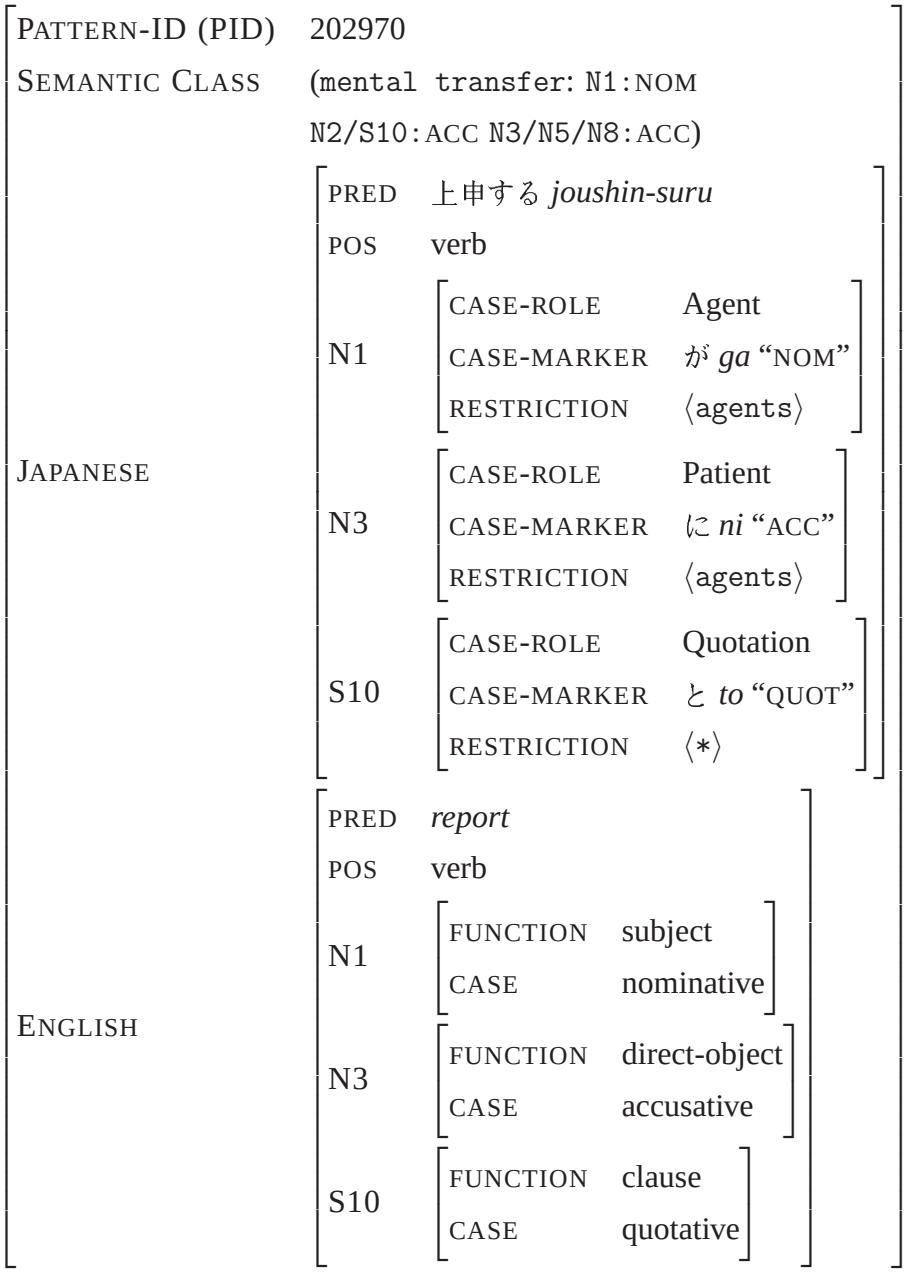


Figure 3.4: Valency (Semantic Pattern) Entry for the verb 上申する *joushin-suru* ⇔ *report* No.2 (SVPC)

3.2 Method of Creating Patterns

3.2.1 Overview of Method of Creating Patterns

The method is based on the observation that verbs with similar meanings typically have similar valency structures. That is, if there is an unknown verb (S_U) whose meaning is similar to an existing verb in the seed dictionary (the known verb S_K), we can copy the valency information of S_K for S_U. This method has some fundamental limitations. It only creates valency patterns for words for which we can find similar words in the valency dictionary. But it is simple and robust because it creates new patterns by copying from the existing patterns.

The basic method used to determine semantic similarity is translation equivalence: if two verbs have the same English translation then they have similar meanings. This massively overgenerates: one sense of a verb may overlap, but not all will. Further, the match criteria are quite loose: any pattern with the same English head.[5] Therefore *give up* and *give back* are counted as the same entry. This allows for minor inconsistencies in the target language dictionaries. In particular the valency dictionary is likely to include commonly appearing adjuncts and complements that do not normally appear in bilingual dictionaries. For example: 行く *iku* "go" is translated as *to go* in EDICT, *go* in the **ALT-J/E** word transfer dictionary and NP_1 *go from* NP_2 *to* NP_3 in **ALT-J/E**'s valency dictionary (among other translations). To match these patterns it is necessary to have some flexibility in the English matching.

A single verb may have multiple possible subcategorization (subcat) and selectional restrictions (SR). We create more patterns by using data about verbal alternations Levin (1993): if the existing verb participates in a known alternation then we create new patterns based on both alternatives (See Section 3.2.4 for more information).

To reduce the overgeneration, we investigate various methods of further constraining the creation of new patterns, such as a simple human check (pre-filter), paraphrasing, a multilingual check, semantic association scores and the strength of the link through English. Then we investigate different ways to merge similar patterns. Evaluation was done using both a task-based evaluation and checking by an expert lexicographer. These methods are described in more detail in the following sections.

[5] We exclude idiomatic patterns (those with fixed case slot fillers) and patterns headed by light verbs such as *make*, *do* and *take*.

3.2.2 Constructing Candidates

To find translation equivalences, we used a plain bilingual dictionary which contains word pairs without valency information. This was made from **ALT-J/E**'s Japanese-English word transfer dictionary and an enhanced version of EDICT (Breen, 2004) where Japanese verbal-nouns were expanded into verbs (e.g., 上申 *joushin* "report" was expanded into 上申する *joushin-suru* "report").[6]

To create a candidate S_U, an Unknown word for which we have no valency information, we find all words where E, the English translation (or translations) is linked to one or more valency patterns S_K in the valency dictionary. Figure 3.5 shows the overall flow of creating new patterns. We discuss the details about filtering methods in Section 3.2.3

For example, 纏う *matou* "wear [clothes]", matching through the translation gives 15 candidate Japanese verbs in the valency dictionary on which we can base the new entry. These include 着る *kiru* "wear", 弱る *yowaru* "wear [out]", 笑いを浮かべる *warai-o ukaberu* "wear a smile" and so on (some possible links are shown in Figure 3.6). Some of this variety of candidates comes from the polysemy of the English verb: 着る *kiru* "wear" corresponds to WordNet sense 1 "be dressed in", 弱る *yowaru* "wear out" to sense 8 "exhaust or tire though overuse or great strain or stress" and 笑いを浮かべる *warai-o ukaberu* "wear a smile" to sense 3 "wear an expression of one's attitude or personality". In fact, 纏う *matou* "wear" corresponds only to sense 1.

3.2.3 Filtering Candidates

In order to filter out inappropriate candidates, we investigate several methods of judging similarity.

Pre-filter

The simplest method is to use human judgment. This is implemented as a pre-filter (`reject`) where an analyst examines the two source language words (S_U, S_K) linked by an English translation, and rejects them if they do not have a similar meaning.

[6]EDICT typically translates Japanese verbal nouns as nouns, without giving a separate verb entry: e.g., 共同 *kyōdō* "cooperation". We used **ALT-J/E**'s English morphological dictionary and the EDICT part-of-speech codes to create 10,395 new verb entries such as: 共同する *kyōdō-suru* "cooperate".

Step 1: For each pattern (S_U-T_U) in the plain $S-T$ dictionary with no pattern in the valency dictionary

- For each valency pattern (S_K) with the same target translation (T_{UK})
  - Create a candidate pair S_U-S_K

Step 2: For each candidate pair S_U-S_K (linked by T_{UK})

- apply filtering methods Section 3.2.3
- If the candidate pair doesn't filter out (S_U-S_K)
  - Replace S_K by S_U then create a new pattern (S_U-T_{UK}) for pairs

Step 3: For each new pattern S_U-T_U (made from S_K-T_K)

1. If S_K-T_K has an alternation S_A-T_A
   also create candidate S_U-T_A Section 3.2.4
2. If there are similar new patterns
   merge them Section 3.2.4

Figure 3.5: Flow of Creating New Patterns

Many words that are obviously dissimilar are linked due to the polysemy of the English pivot. Rejecting them is a very fast process. It only becomes slow if the analyst does not recognize one of the verbs and therefore has to look it up. The strength of this method is its accuracy: the weakness is that it requires human intervention, and is thus expensive.

Consider the three pairs shown in (2). For a Japanese native speaker, (a) is clearly good, while (b) and (c) are clearly bad.

(2) Potential candidates for 纏う *matou* "wear [clothes]".

a. 纏う ⇔着る (⇔ *kiru* "wear [clothes]")

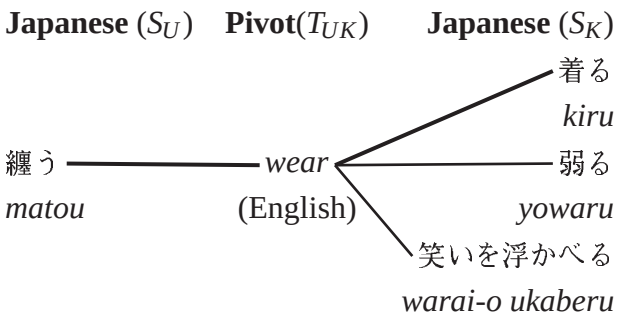


Figure 3.6: Creating Candidates through a Common Pivot Translation

b. 纏う ⇔弱る (⇔ *yowaru* “wear [out]”)

c. 纏う ⇔笑いを浮かべる (⇔ *warai-o ukaberu* “wear a smile”)

Paraphrasing

The aim of this filter is to eliminate candidate patterns with incorrect subcats, without having to use an expert bilingual lexicographer.

The filtering is done by an analyst. The analyst judges whether sentences with the candidate verb (S_U) replaced by the seed verb (S_K) (and vice-versa) are grammatical or not. Ideally, words with the same subcat will produce a grammatical paraphrase, while those with different subcats will not.

For example, both 結婚する *kekkon-suru* “marry” (S_K) and 嫁ぐ *totsugu* “marry into”(S_U) have similar meanings. But 結婚する *kekkon-suru* “marry” is a reciprocal verb: “a man and a woman marry”, 嫁ぐ *totsugu* “marry into” on the other hand is directional, “a woman marries a man/into a family” and thus the subcat is different. This can be seen in (3) and (4), where 結婚する *kekkon-suru* “marry” is replaced with 嫁ぐ*totsugu* “marry into”, but (4) is ungrammatical.

(3) 彼女 は 彼 と 結婚する。
kanojo wa kare to kekkon-suru
she TOP him REC marry
“She’ll marry (with) him.”

(4) * 彼女 は 彼 と 嫁ぐ。
kanojyo wa kare to totsugu

In order to filter out inappropriate candidates, we compare the usage of S_K with S_U using examples from a corpus. Two judgments are made for each paraphrase pair: is

the paraphrase grammatical, and if it is grammatical, are the meanings similar?

This judgment can be done by monolingual speakers of the source language. We test both directions: first we find example sentences using S_U, replace S_U with S_K and compare the paraphrased sentences. Then we find sentences for valence patterns using S_K, replace them with S_U and judge the similarity. Figure 3.7 shows the comparison using paraphrases.

For each candidate pattern S_U-T (from S_K-T)

- Extract 5 sentences using S_U from the corpus
  For each sentence
  - Replace S_U with S_K
  - Classify the paraphrased sentence into 3 grammaticality classes
    if the class is `grammatical`
    - Classify the semantic similarity into 6 classes
- Extract 5 sentences using each pattern of S_K from the corpus
  - Replace S_K with S_U
  - Test as above

Figure 3.7: Flow of Paraphrasing Check

The three grammaticality classes are: `grammatical`, `ungrammatical`, `grammatical in some context`.[7] Semantic similarity was divided into the following classes:

- `same`: S_U 脅す *odosu* “threaten” and S_K 威す *odosu* “threaten”
- `close`: S_U 具申する *gushin-suru* “report” and S_K 上申する *joushin-suru* “report”
- [S_U] `broader`: S_U 作り出す *tsukuri-dasu* “create” and S_K 発明する *hatsumei-suru* “invent”

[7]The analysts also rejected 7.9% of the example sentences as irrelevant. These were sentences where the verb did not actually appear, but that had been selected due to errors in the morphological analysis.

- [S_U] narrower: S_U 再婚する *saikon-suru* “remarry” and S_K 結婚する *kekkon-suru* “marry”
- different nuance: S_U 押収する *oushuu-suru* “expropriate” and S_K 取り上げる *toriageru* “confiscate” (S_U is more formal than S_K.)
- different: S_U 立ち向かう *tachi-mukau* “confront” and S_K 反論する *hanron-suru* “argue against” (their meanings overlap so they are classified into other classes in some context.)

Next, we give an example of the paraphrasing; for the unknown Japanese word S_U 具申する *gushin-suru* “report” we look at the existing word S_K 上申する *joushin-suru* “report” which exists in the valency dictionary, with the same English translation.

We extract 5 sentences from our corpus which use S_K, for example (5; slightly simplified here), and replace S_K with S_U (6).

(5) 経営 トップ に この こと を 上申 し、 ＯＫ が 出た。
Keiei toppu ni kono koto o joushin-shi, OK ga deta.
management top DAT this thing ACC report, ok NOM came-out
“I reported this to the top management and they OKed it.”

(6) 経営 トップ に この こと を 具申 し、 ＯＫ が 出た。
Keiei toppu ni kono koto o gushin-shi, OK ga deta.

Similarly, we extract 5 sentences from our corpus which use S_U, for example (7; slightly simplified here), and replace S_U with S_K (8).

(7) 罰則 を 重くする 必要 は ない と 具申 した。
bassoku o omoku-suru hitsuyou wa nai to gushin-shita.
penalty ACC increase need TOP nothing QUOT reported.
“I reported that there is no need to make the penal regulations more severe.”

(8) 罰則 を 重くする 必要 は ないと 上申 した。
bassoku o omoku-suru hitsuyou wa nai to joushin-shita.

Both paraphrases (6) and (8) are grammatical and both pairs (5, 6) and (7, 8) have close meanings. This is done for all five sentences containing S_U and then done in reverse for all 5 sentences matching the pattern for S_K.

The strength of this paraphrasing method is that non-experts can make the judgments and there is supporting data for them. The weaknesses are that it requires example sentences and is labor intensive.

We also investigate checking the paraphrases using web-data.[8] In this experiment we replace the target word as above and then look for the paraphrased sentence: if it exists then the paraphrase is good. However, the average length of sentence we could find to paraphrase is 19 words (38 characters). We therefore got so few hits (fewer than 1%) that the test was practically useless.

Multilingual Check

Another possible filter on overgeneration is to use multiple languages as pivots (Bond et al., 2001; Paik et al., 2001; Fujita and Bond, 2004). Because we match the entire translation in language X, there is no overgeneration due to complex verbs. When plain dictionaries are available in multiple languages, then the criterion can be varied further — for example to use all dictionaries and select these words which have at least one matching translation of X (we call this **UNION**) or to use all dictionaries and select only those words which have matching translations in all languages (we call this **INTER**).

In our experiment we used a Japanese-to-Chinese machine dictionary available in machine readable form: $J-C$ (Shogakukan and Peking Shomoinshokan, 1987); and two dictionaries available on-line: Wadoku Jiten — a Japanese-to-German dictionary $J-G$ (Apel, 2002); and Dico FJ — a Japanese-to-French dictionary $J-F$ (Desperrier, 2002). In Table 3.2, we show the number of patterns for each of the plain dictionaries used in this thesis. Three of these dictionaries (EDICT, Wadoku Jiten and Dico FJ) are available on-line, and are growing over time; the numbers given here are for the versions we used. Most bilingual entries lacked POS tags, so we matched on the surface form of all entries, even though most are not verbs or adjectives.

We use the plain dictionaries in several ways. First, we only use the pairs of S_U and S_K which have the same: (1) Chinese translation C (we call this strategy **CN**), (2) German translation G (we call this **DE**), (3) French translation F (we call this **FR**), then (4) have at least one matching translation in C, G and F (**UNION**), or finally (5) have matching translations in all of C, G and F (**INTER**).

[8] Thanks to an anonymous reviewer for the suggestion.

Table 3.2: Size of J-X Dictionaries

J-X	Japanese	X	Pairs
J-C	72,400	102,300	180,800
J-G	252,400	224,000	526,000
J-F	16,600	10,500	37,900
J-E (EDICT)	94,200	80,400	154,600
J-E (**ALT-J/E**)	323,700	276,100	415,000

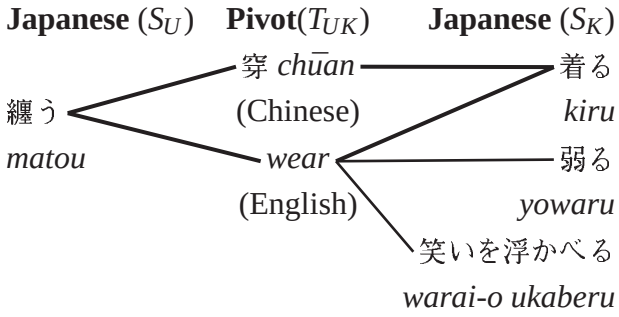


Figure 3.8: Creating Candidates through Multiple Pivots

An example of the utility of adding another language is shown in Figure 3.8. In this case, 纏う*matou*, 着る*kiru*, 弱る*yowaru* and 笑いを浮かべる*warai-o ukaberu* have the same English translation *wear*. But the word *wear* is polysemous and the Japanese pairs 纏う*matou* and 弱る *yowaru* or 笑いを浮かべる*warai-o ukaberu* don't have similar meanings. Because Chinese verbs have different patterns of polysemy to English, only the appropriate Japanese candidate pair (纏う*matou* and 着る*kiru*) is linked by both English and Chinese (Figure 3.8).

If the source word itself is polysemous (or is monosemous with multiple subcategorization frames), then there can be more than one candidate language word linked through multiple languages. In this case we will build multiple patterns. Each pattern would correspond to a different sense.

Association Scores

We also tested the use of association scores based on word-vector distances taken from word-definitions and corpora (Kasahara et al., 1997). This measure is designed to

simulate human word association.

First we used the association score to cut off subthreshold candidates. Then we used the score to rank words in order of similarity and only create patterns for words judged similar by association. We investigated creating patterns for various ranks of similarity: words more similar than a threshold, the most similar word, and words that were within the top 10, 100, 1,000 and 10,000 most similar words. The strength of this method is that it is fully automatic. The weakness is that highly associated words are not necessarily syntactically or semantically similar (for example 結婚する *kekkon-suru* "marry" and 嫁ぐ *totsugu* "marry into").

Translation Link Strength

We also evaluated the quality of the English translation link. This was measured using the dice coefficient. That is, if S_U has English translations $T(S_U)$, and they link through the valency dictionary to a Japanese word S_K with translations $T(S_K)$, then the strength of the link is:

$$\text{(3.1)}\quad \text{link strength} = \frac{2 \times (|T(S_U) \cap T(S_K)|)}{|T(S_U)| + |T(S_K)|}$$

This is similar to the one-time inverse consultation score used by Tanaka et al. (1998) to link Japanese and French through English. The strength of this method is that it is fully automatic. The weakness is that it depends entirely on the quality of the bilingual lexicon.

3.2.4 Making Candidates Robust

In order to make the system robust, we add alternative candidates and then merge similar candidates.

Adding Alternative Patterns

If the pattern in the seed valency dictionary participates in a diathesis alternation (such as *I broke the cup* ⇔ *The cup broke*), then we create candidates for both alternatives at once.

For example, the unknown verb 着火する *chakka-suru* "ignite" matches 引火する *inka-suru* "ignite" which has two alternatives in the seed dictionary linked by the

Causative/Inchoative Alternation. We make patterns for both of them, allowing us to match both (9)[9] and (10).

(9) 導火線　が　着火した。
doukasen ga chakka-shita.
fuse ACC ignited/caught fire
"The fuse ignited. / The fuse caught fire. "

(10) 彼　は　導火線　に　着火した。
kare wa doukasen ni chakka-shita.
He TOP fuse DAT ignited.
"He ignited the fuse."

This can only be done if the seed dictionary contains information about alternations. Currently, identifying alternations and adding them and to lexicons, is being done both by linguists (Furumaki and Tanaka, 2003) and computational linguists (Dorr, 1997; Bond et al., 2002; McCarthy, 2000).

Merging Patterns

Merging similar candidates is an important problem for corpus-based approaches, which normally have 10s to 1000s of candidates to merge (Li and Abe, 1998; McCarthy, 2000). In our case we have fewer candidates, and more information. Although the existence of very similar patterns does not affect the translation quality, the redundancy creates spurious ambiguity, which slows the system down and makes debugging harder.

We reduce the number of redundant patterns by merging similar patterns. First, if two patterns are identical, we merge them. We then merge candidates that only differ in their case-markers and selectional restrictions. That is, they have the same Japanese head-word, the same English head-word, the same English subcat, the same number of arguments, and the same case-roles. If the patterns have different case-markers, the merged pattern is given the union of the two sets (for example if the argument of S_{U1} has {に} *ni* "to", and the argument of S_{U2} has {に,へ} *ni,e* "to", then the merged pattern will have {に, へ} *ni,e* "to" as its case markers. However, if one of the similar

[9]Actually, (9)'s translation is *catch fire*. We use *ignite* only for explanation of the alternation in English.

patterns is from a domain-specific dictionary, it is rejected in favor of the pattern from the general dictionary, rather than being merged.

We tested two strategies for merging selectional restrictions: **parent** and **child**. All pairs of SRs from the two patterns are compared. In **parent**, if one restriction subsumes the other the least restrictive (the parent) is used. In **child**, the most restrictive (the child) is used. If neither restriction subsumes the other, then both are used. Multiple patterns can be merged, not only pairs of similar patterns.

At this step, if the original pattern is marked in the transfer lexicon as a technical term and its lemma used in other candidate patterns, we don't use the pattern. This stops us from basing patterns on very specialized usages of words if we have other alternatives.

3.3 Creation and Evaluation

In this Section, we create new patterns and apply several filtering methods (Step 1, 2 of Figure 3.5). These are then evaluated according to their effect on translation quality (Section 3.3.3) or by expert lexicographers (Section 3.3.4).

3.3.1 Target Verbs

We use the valency dictionary from the Japanese-to-English machine translation system **ALT-J/E** as a seed lexicon (See, Section 2.2).

In **ALT-J/E**'s Japanese-English word dictionary, there are 55,615 J-E pairs whose Japanese part of speech is adjective, adjectival noun or verb. There are a total of 20,925 distinct Japanese entries. However, due to the cost of making detailed entries, only those 4,937 entries have valency patterns: 15,988 entries have no pattern. Of the 55,615 J-E pairs, 35,999 have no pattern in the valency dictionary. Our method is applicable to 13,408 of these pairs: their English entry has an pattern in the valency dictionary.

In Table 3.1 and Figure 3.2, we showed that 8,304 kinds of verbs have no pattern. Of those, 4,129 (49.7 %) verbs appear in the **ALT-J/E**'s Japanese-English transfer dictionary or EDICT and have a pattern with the same translation in the valency dictionary (See Table 3.3).

Table 3.3: The Possibility of Increasing Cover Ratio for Japanese Newspapers (9 years)

In lexicon	No. of Types (%)		No. of Tokens (%)	
Japanese exists	4,997	37.5	24,656,590	92.5
English exists	4,129	31.0	1,355,552	5.1
No pattern	4,175	32.4	645,158	2.4
Total	13,301	100.0	26,657,300	100.0

3.3.2 Results of Creation using Several Filtering Methods

We targeted the 4,129 verbs to create new patterns using multilingual check, association score and the link strength. Then, because we were able to find 5 or more examples from a corpus of newspaper for 3,753 (90.9 %) of these, we targeted the 3,753 verbs to test the paraphrasing filter and pre-filter. Table 3.4 shows the number of created patterns for the target verbs through the several filters.

The original number of candidates for the 3,753 target verbs was enormous: 108,733 pairs of S_U and S_K. Most of these were removed in the pre-filtering stage, leaving 2,492 unknown verbs matching 7,902 S_Ks in the valency dictionary. After the pre-filter, there were on average 3.2 patterns/verb.

For the paraphrasing filter, analysts took about 7 minutes per verb. The data was split between three analysts, one a linguist and two people with no special training.

The other three filters (multilingual check, association scores and link check) are fully automatic.

3.3.3 Translation Task-based Evaluation of Filtering Methods

In this section we evaluated the effect on translation quality for created patterns using various filters. For each verb (S_U) we picked the two shortest sentences we could find (on average 81.8 characters/sentence: 40 words) from a corpus of 9 years of newspaper text (4 years of Mainichi and 5 years of Nikkei)[10]. This corpus had not been used in the paraphrasing filter, i.e., all the sentences were unknown. We tried to get 2 sentences for each verb, but could only find one sentence for some verbs. For the pre-filter, the

[10]Mainichi '93, '96, '97, '98 and Nikkei '90, '91, '92, '93, '94.

Table 3.4: Number of Created Valency Patterns

Filtering Methods Filter	 Condition	Patterns (P)	Verbs (V)	Average (P/V)
Pre-Filter		7,902	2,492	3.2
Para-	$S_U \Rightarrow S_K$: grammatical % $\geq$ 90	321	205	1.6
phrase	$S_K \Rightarrow S_U$: same or close % $\geq$ 90	2,716	1,428	1.9
	CN	2,077	668	3.1
Multi-	**DE**	7,826	90.7	4.6
lingual	**FR**	629	8.2	4.1
Check	**INTER**	141	51	2.8
	UNION	9,178	1,868	4.9
Association	1st ranked	2,161	1,632	1.3
Scores	score $\geq$ 0.8	89	55	1.6
	score $\geq$ 0.7	273	163	1.7
Link Strength		4,814	778	6.2

number of target sentences is too large to evaluate them all, so we did an evaluation over a sample.

We translated the test sentences both `with` the valency dictionary which has the new patterns, and `w/out` the new patterns. When there is no pattern for a verb in the valency dictionary, the system uses either the default translation in the plain dictionary or if there is no entry in the plain dictionary, the Japanese verb as is.

Translations that were identical were marked `no change`. Translations that changed were evaluated by people fluent in both languages. The evaluators were shown randomized translations to make the evaluation blind. The new translations were placed into three categories: `improved`, `equivalent` and `degraded`. All the judgments were based on the change in translation quality, not the absolute quality of the entire sentence.

For example, in (11) the change is evaluated as “B is improved compared to A”, in this case, B is `with`, i.e., `with` is `improved`.

(11) 動く ものが いると心 が なごむ もの です。
ugoku mono ga iru to kokoro ga nagomu mono desu.
move thing NOM exist if heart NOM calm down which is

A: `If there is a thing which moves, a heart is softened.`

B: `If there is a thing which moves, we calm down.`

The results of the evaluation for each filtering method are given in Table 3.5. Thresholds were chosen after examining the data over a wide range of values, although we do not show all the results here. In addition to the number of sentences which improved or degraded, Table 3.5 shows the `difference` (`improved` − `degraded`).

Table 3.5: Task-based Evaluation of New Patterns for each Filtering method

Filtering Methods	Judgment of Translation Quality								`difference`		Total
	`improved`		`no change`		`equivalent`		`degraded`				
	No.	%	No.	%	No.	%	No.	%	No.	%	No.
Pre-Filter (Estimation)		32		26		26		16		+16	
Paraphrase[11]	1,636	37.5	1,063	24.3	1,115	25.5	552	12.6	1,084	+24.9	4,366
Multi- **CN**	305	23.5	392	30.2	410	31.6	192	14.8	113	+8.7	1,299
lingual **DE**	776	24.7	991	31.6	809	25.8	561	17.9	215	+6.8	3,137
Check **FR**	39	16.6	98	41.7	70	29.8	28	11.9	11	+4.7	235
UNION	873	24.2	1,153	31.9	950	26.3	634	17.6	239	+6.6	3,610
Association Score ≥ 0.8	18	15.8	47	41.2	28	24.6	21	18.4	-3	-2.6	114
Link Strength ≥ 0.9	366	22.1	510	30.8	422	25.5	359	21.7	7	+0.4	1,657

As can be seen in Table 3.5, paraphrasing gives the best quality; that is using only the pre-filter and the grammaticality judgments, 37.5% of translations improved and only 12.6% degraded, an overall `difference` of +24.9%. Pre-filtering gives the second best quality (estimated quality). The difference using a pre-filter is +16%, which is a good result.

[11]We use the patterns that have at least one S_U to S_K paraphrase that is `grammatical`.

In the results of multilingual check of Table 3.5, the overall difference ranges from +4.7% to +8.7%. The biggest improvement was for **CN**, which comes from a totally different language family to English. In all cases, the number of `improved` sentences is greater than those `degraded`, but **UNION** creates the most patterns, and has an overall difference of +6.6%.

Table 3.5 shows that the scores from association and link strength are not high. Therefore, we conclude that association scores and link strength are not suitable filters for calculating syntactic or semantic similarity in this task.

In summary, paraphrasing gives the best quality of translation, but pre-filtering is cheaper and satisfies both quantity and quality. Of the fully automatic methods, the multilingual check using **UNION** gives the best results. The translation results are analyzed in more detail in Section 3.4.1.

3.3.4 Lexicographers' Evaluation of Filtering Methods

In this section we evaluate the effectiveness of the filters, using direct analysis by expert lexicographers as our gold standard. The results of several methods are given in Table 3.6.

In Table 3.6, precision is the percentage of acceptable patterns that passed the filter over all patterns that passed the filter. Recall is the percentage of acceptable patterns that passed the filter over all acceptable patterns[12]. The baseline is to use all patterns that passed the pre-filter: this gives a precision of 53.4% and 100% recall.

The highest precision (72.3%) came from only using patterns where the unknown verb (S_U) was the most similar to the known verb (S_K). The recall, however is a disappointing 3%. Using the paraphrase tests based on sentences where the unknown verb replaced the known verb, gave almost as high a precision and a higher recall (71.8% and 23.7% respectively).

Next we considered the multilingual filter. Using one dictionary (the strategies; **CN**, **DE** and **FR**), **DE** gives the highest recall, but precision is not so high. **CN** gives 11.4% recall, but its precision is higher than **DE**. This can be explaned by the following: (1) because the Japanese-to-German dictionary is larger than the Japanese-to-Chinese dictionary, the Japanese-to-German dictionary has more polysemy and the polysemy

[12]The number of patterns that passed pre-filter is enormous, so we evaluated a sample set. In this sample set, the total number of acceptable patterns is 4,272.

Table 3.6: Lexicographers' Evaluation of New Patterns for each Filtering method

Filtering Methods		Precision	Recall	F-score
Filter	Condition	(%)	(%)	(%)
Pre-filter		53.4	100.0	53.5
Para-phrage	$S_U \Rightarrow S_K$: grammatical % ≥ 90	57.1	61.0	59.0
	$S_U \Rightarrow S_K$: ungrammatical % $= 0$	57.1	61.3	59.1
	$S_U \Rightarrow S_K$: same or close % ≥ 90	70.2	22.0	33.5
	$S_K \Rightarrow S_U$: grammatical % ≥ 90	61.7	55.1	58.2
	$S_K \Rightarrow S_U$: ungrammatical % $= 0$	61.7	55.7	58.5
	$S_K \Rightarrow S_U$: same or close % ≥ 90	71.8	23.7	35.6
Multi-lingual Check	**CN**	66.9	3.2	6.1
	DE	63.2	11.4	19.3
	FR	64.1	0.9	1.8
	INTER	73.8	0.2	0.4
	UNION	62.3	13.2	21.8
Association Score	1st ranked	72.3	3.0	5.8
Link Strength	score ≥ 0.9	59.6	7.1	12.7

makes the accuracy low. (2) German and English are in the same Germanic group of Indo-European language family, but Chinese and English are in different language families. So, Chinese is more effective for filtering-out the wrong pairs caused by English polysemy.

Even so, the most forgiving method, **UNION** gives good precision, and its recall is higher than the remainder. So **UNION** is the most useful of the multilingual filtering strategies.

These results show the same trends as the task-based evaluation: Paraphrasing gives the highest score, the pre-filter is next, and the multilingual pivot using **UNION** is the best of the fully automatic filters. We discuss the results of the lexicographers' evaluation in more detail in Section 3.4.2.

3.3.5 Evaluation of Alternations and Merging

In this section, we evaluate the methods used to make the new patterns robust (Step 3 of Figure 3.5). For this evaluation, we take the patterns which passed the pre-filter (those which satisfied both quantity and quality, and including the patterns made through paraphrasing) and we use them as a basic set of new patterns. We add some alternative patterns, and then merge any similar patterns. Then, analysts evaluate the created, added and merged new patterns.

An additional 178 patterns were made using alternations. At the next step, we were able to merge 2,891 similar patterns into 1,183, leaving 6,327 candidate patterns for 2,492 verbs. The maximum number of patterns merged into one was nine (勘違いする *kanchigai-suru* "mistake"). Half the mergers used the **parent** method and half used the **child** method Section 3.2.4. Table 3.7 shows the number of patterns which had case-markers (CM) or SR merged. In Table 3.7, 50% of the patterns had CMs merged and over 97% had SRs merged.

Table 3.7: Number of Merged Patterns

	`parent`		`child`		Total	
	No.	%	No.	%	No.	%
Both Merged	324	54.8	311	52.5	635	53.7
only SR Merged	254	43.0	268	45.3	522	44.1
Same	13	2.2	13	2.2	26	2.2
Total	591	100	592	100	1,183	100

After merging, there were 2.5 patterns/verb, a much closer ratio to that of the seed lexicon.

The results of the analysis are given in Table 3.8. Separate columns are shown for patterns made using alternations, patterns that were merged using the **parent** method, patterns that were merged using the **child** method, the remainder of the patterns and all the patterns. These results are used to evaluate the similarity filters.

The evaluation took around 5 minutes per verb. Each pattern was marked as: `acceptable`, `fixable` or `useless`: `acceptable` patterns could be used as they were; `fixable` patterns could be used with minor revisions; `useless` patterns were so poor

Table 3.8: Lexicographers' Evaluations for New Patterns

	Alter-nation		Merge							
			Parent		Child		Remainder		Total	
Result	No.	%	No.	%	No.	%	No.	%	No.	%
Acceptable	53	30.8	366	61.9	333	56.3	2,505	50.4	3,257	51.5
Fixable	63	36.6	195	33.0	231	39.0	1,803	36.3	2,292	36.2
Useless	56	32.6	28	4.7	26	4.4	640	12.9	750	11.9
-	0	0	2	0.3	2	0.3	24	0.5	28	0.4
Total	172	100	591	100	592	100	4,972	100	6,327	100

that it would be easier to create a pattern from scratch.

The majority of patterns that passed the pre-filter were usable as they were (51.5%). A further 36.2% were usable with minor revisions, giving 87.7% potentially useful patterns. These are encouraging results.

Patterns made using the alternations were worse overall, while those made by merging were substantially better. One of the reasons for the poor quality of the alternations is that they added another transformation to the original. If we consider only alternations of acceptable patterns, then they are acceptable 30.8% of the time. Therefore, it is better to make patterns using alternations after all other filters have been applied.

Fewer fixes were necessary for the patterns merged with more general restrictions (**parent:child** — 61.9%:56.3%) than with the more restricted patterns, although both were better than the remainder.

Examining the kinds of changes needed by the merged patterns showed the child set needed their SRs corrected more often. This shows clearly that merging to the least restrictive values (the parent strategy) is the best.

3.4 Discussion

In this section, we analyze the results of translation evaluation (Section 3.3.3) and direct evaluation (Section 3.3.4) in more detail. Then, based on the results of the analyses, we refine our proposed method.

3.4.1 Analysis of the Translation Results

First, we analyze the reasons for the `improved` and `degraded` translations.

Reasons for `improved` results: (1) The system was able to translate previously unknown words. The translation may not be the best but it is better than an unknown word. (2) A new pattern with a better translation was selected. (3) The sentence was translated using the correct subcategorization, which allowed a zero pronoun to be supplemented or some other improvement.

We show some examples of the changed translations, using simplified example sentences.

In (12) the English valency information supplies the subcategorized preposition *for* in *wish for*. The default translation makes the argument a plain direct object, which is ungrammatical for *wish*.

(12) 国民　の 大半　が 平和　を 欲し、 その ため の 危険 を 負う 覚悟　が
kokumin no taihan ga heiwa o hosshi, sono tame no kiken o ou kakugo ga
ある と 信じて きた。
aru to shinjite kita.

`w/out: It was believed that national most wished peace and that there was the preparedness that we owe danger for that purpose to.`

`with: It was believed that national most wished for peace and that there was the preparedness that we owe danger for that purpose to.`

In (13), the translation with is an improvement.

(13) ＮＡＴＯは セルビア人 に 宣戦　を 布告した。
NATO ha Serbia-jin ni sensen o fukoku-shita.

`w/out: NATO decreed a declaration of war to Serbia person.`

`with: NATO announced a declaration of war to Serbia person.`

Reasons for `degraded` results. (1) A new pattern was selected whose translation was less appropriate. (2) the detailed nuance was lost. For example, 撫で上げる *nade-ageru* “brush up” became simply *brush*.

The main reason for these degradations was a change in the default translation. When there is no pattern available, **ALT-J/E** uses a translation from its word dictionary.

As there is little information available to choose between alternatives, the first listed translation is used (the first listed translation is meant to be the most general translation). However, when we made patterns, we looked at all listed translation equivalents. 51% of the time we were able to make a pattern with the first listed translation, 27% with the second, 11% with the third, and 11% with the fourth, fifth or sixth. However, when a pattern exists **ALT-J/E** uses it in preference to an entry in the word dictionary. Therefore, the translation was changed for many patterns. Sometimes the new translation was an improvement, but sometimes it was not. For example, 口答えする *kuchi-gotae-suru* "answer back" had two translations in the word dictionary: (1) *answer back* and (2) *retort*. We could only find a pattern for *retort* and so this became the system's choice. However, *answer back* was in fact a better translation in the examples we saw.

3.4.2 Analysis of the Lexicographers Evaluation

Direct evaluation shows two things that the task-based evaluation did not make clear. The first is the utility of merging similar patterns: the resulting patterns are of high quality, and the dictionary becomes more compact. When merging, the best strategy is to create new patterns with less restrictive selectional restrictions. The second is that evaluation by paraphrasing is no better than using expert lexicographers. Although using paraphrasing does improve the quality of the dictionary, it is quicker and more accurate to use lexicographers directly (5 minutes vs 7 minutes). Further, paraphrase judgments are hard to make for untrained analysts: linguists made paraphrase judgments with higher accuracy.

For example in (14), (15), the meaning is changed in (15) but if we assume the special state, (15) will be acceptable. This falsifies the claim that paraphrase judgments can be done cheaply with untrained analysts, and makes it less effective to use paraphrasing as a filter.

(14) ヴァカーリ夫妻　　が　著した　　和英辞典。
Vaccari-fusai　　*ga*　*arawa-shita*　*waei-jiten.*
Mr. and Mrs. Vaccari ACC wrote　　Japanese-English dictionary.
"The Japanese-English dictionary is written by Mr. and Mrs. Vaccari."

(15) ？ヴァカーリ夫妻 が 表記した 和英辞典。
Vaccari-fusai ga hyouki-shita waei-jiten.
"The Japanese-English dictionary is signed on the front by Mr. and Mrs. Vaccari."

From a practical point of view the results are encouraging: we can produce useful new patterns with only a simple monolingual judgment as pre-filter: "are these verbs similar in meaning?", and it has been shown that these patterns improve the quality of translation in 32% of sentences versus degradations in only 16%.

The quality can further be improved by the candidates being checked by lexicographers. This is relatively expensive, at an additional 5 minutes per verb, but is still cheaper than creating patterns from scratch. Preliminary investigation shows that even correcting the fixable patterns takes less than 10 additional minutes per pattern on average, for a total of 15 minutes per pattern.

At the end of these experiments, we increased the valency type coverage about 1.5 times (from 4,997 to 7,427) and cover almost half of the missing tokens. This means in practice that the number of sentences which have unknown verbs decrease from one in 5 to one in 9 using the data from 9 years' newspapers (see Table 3.9 and Figure 3.9: we graphed Table 3.9 as Figure 3.9).

Overall, our results show that hand-compilation is still necessary for building high quality lexicons. However, semi-automatic acquisition of candidates, and merging the acquired candidates can increase efficiency considerably.

3.4.3 Refining the Method

As we showed in Section 3.4.2, the method for creating new patterns using only a pre-filter is very effective. The pre-filtering is a very simple judgement, done by people. So, to reduce the cost, we examine whether we can use automatic filters instead of human pre-filters, even if only for some of the target words.

Table 3.6 showed that the most efffective automatic filter was the Multilingual check, used after a pre-filter. So, we ran the multilingual check without applying a pre-filter. The results for Table 3.6 are shown after applying the pre-filter. Here, we consider doing the multilingual check without applying the pre-filter.

Table 3.9: Cover Ratio of Created Patterns for Japanese Newspapers (9 years)

In lexicon	No. of Types (%)		No. of Tokens (%)	
Japanese exists	4,997	37.5	24,656,590	92.5
Created Japanese exists	2,430	18.3	886,126	3.3
No pattern	5,874	44.2	1,114,584	4.2
Total	13,301	100.0	26,657,300	100.0

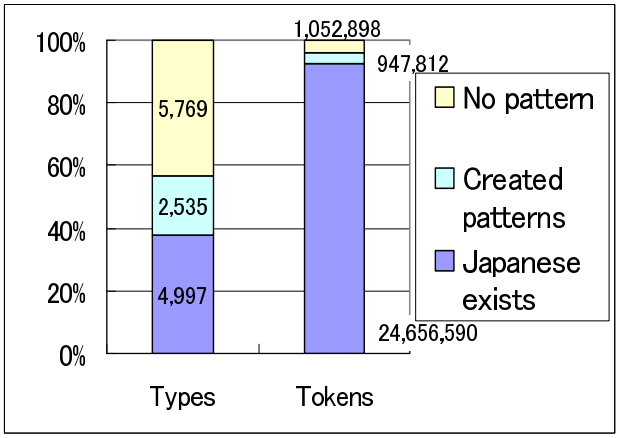


Figure 3.9: Graph of Cover Ratio of Created Patterns for Japanese Newspapers (9 years)

Table 3.10 shows the results using the multilingual check without applying a pre-filter. The Japanese-to-French dictionary is relatively small、so in Table 3.10, we use Japanese-to-Chinese ($J-C$) and Japanese-to-German ($J-G$) dictionaries. It means in Table 3.10, **UNION** has at least one matching translation in C and G, and **INTER** has matching translations in both of C, and G. Table 3.10 shows the number of target verb pairs whose S_U and S_K both exist in the dictionary, and the number after applying pre-filter, too.

Table 3.10: Number of Creatable Valency Patterns Using Multilingual Check

Filtering Methods		Without Pre-filter			Through Pre-filter		
Filter	Condition	(T[13])	(P[14])	(V[15])	(T)	(P)	(V)
No Filter		108,733				7,902	2,492
CN	S_U, S_K exist in $J-C$	26,715			9,540		
	C_U=C_K	1,474	2,077	890	938	1,389	545
DE	S_U, S_K exist in $J-G$	91,357			31,389		
	G_U=G_K	6,178	7,826	2,892	3,803	4,944	1,631
UNION	S_U, S_K exist	92,628			9,540		
	C_U=C_K or G_U=G_K	6,981	8,933	2,729	4,245	5,594	1,592
INTER	S_U, S_K exist	25,444			9,053		
	C_U=C_K, G_U=G_K	671	970	470	496	739	323

From Table 3.10, 470 patterns for 671 pairs are creatable if we use **INTER**, but of those, only 496 pairs (73.9%) went through the pre-filter. We checked 34 pairs (19.4%) of the remaining 175 pairs which are not through the pre-filter. The 34 pairs can expand to 45 patterns. Of those, 37 patterns (82.2%) should be acceptable.

For example, the pair of S_U 娶る *metoru* "marry" and S_K 貰う *morau* "marry" was rejected in the pre-filter, but it has same Chinese and German translation. From this pattern of S_K 貰う *morau* "marry" we can create an acceptable new pattern for S_U 娶る *metoru* "marry". Now, 貰う *morau* has polysemy. From the Gakken Japanese Dictionary (Kindaichi and Ikeda, 1988), 貰う *morau* has 6 senses: That is (1) *given*,

(2) *get*, (3) *marry*, (4) *have a break*, (5) *take on*, and (6) *win*.

At the pre-filter stage, it is likely that the less familiar meanings of the word[16] were not considered by the analyst. The multilingual check can not only reduce the cost, but also cover the mistakes due to human error. Of course, not all the patterns produced through the multilingual check are correct, but doing the check before the pre-filter is cost effective.

The multilingual check is useful, but only 23.2% of the target pairs are seen with both S_U and S_K in the Japanese-to-Chinese dictionary. Even in the Japanese-to-German dictionary, only 79.3% of the target pairs are seen. So, we should use the pre-filter for the remaining patterns.

We therefore propose a method of building information-rich lexicons that proceeds as follows: (1) build a seed lexicon by hand; (2) extend it automatically using more than one bilingual lexicon; (3) extend it semi-automatically using bilingual lexicons and a simple pre-filter check; (4) merge any similar patterns, making the selectional restrictions broader rather than narrower; (5) revise the new patterns as far as possible.

This method is also applicable to work in new language pairs. It will always be the case that simple bilingual lexicons are larger than information-rich lexicons — therefore it will be worthwhile using the former to extend the latter.

Our work is similar in spirit to that of Dorr et al. (2002), who link two information-rich resources (one English and one Chinese) using a bilingual dictionary. They then use the bilingual dictionary to fill in gaps, effectively using a simpler resource to increase the size of the information-rich lexicons.

Kanamaru et al. (2005) examined a method to get Japanese frames using the English FrameNet (Johnson et al., 2002) and an English-Japanese bilingual corpus. They found candidate Lexical Units via the manually translated words. The method has only been evaluated for a single verb: 襲う *osou* “attack”. This method can also be used to provide a bilingual valency dictionary, using a well aligned bilingual corpus instead of a plain dictionary.

An earlier version of this work (Fujita and Bond, 2002a) inspired Hong et al. (2004)

[10]T is the No. of Target Verb Pairs.

[11]P is the No. of Patterns.

[12]V is the No. of Verb Types.

[16]The pre-filter rejected patterns based on senses 3 and 6: that is, pairs with S_K 娶る *metoru* “marry” and S_U 博する *hakusuru* “win”.

to use the same method to create Korean-Chinese patterns, extending the number of patterns in their pattern-based machine translation system from around 110,000 to 350,000. They used three automatic checks: (1) the verbs must have the same voice; (2) neither verb must be an idiom and (3) the target language verb cannot be a light verb (support verb). The automatically created verbs were then checked by a lexicographer and non-synonyms rejected (pre-filter). The newly created verbs raised the percentage of perfectly matched patterns from 59.2% to 64.4% a gain of 5.2%. This shows that the general approach is fully extensible: it works for different systems and for different language pairs.

Future Work

We would like to experiment with more aggressive merging. In this thesis, we only merged patterns with the same case-roles and same English subcat Section 3.2.4. But when the case-roles differ only with adjunct case-slots, we could potentially merge them.

For example, すっぱ抜く *suppanuku* "expose" had the following two candidate patterns as shown in Figure 3.10 and 3.11.

┌ **N1** ⟨3:agent⟩	が *ga*
├ **N2** ⟨1236:human-activities, ...⟩	を *o*
├ **N3** ⟨3:agent⟩	に *ni*
└ すっぱ抜く *suppanuku* "expose"	

Figure 3.10: Candidate Pattern for すっぱ抜く *suppanuku* "expose" (1)

┌ **N1** ⟨3:agent⟩	が *ga*
├ **N2** ⟨1236:human-activities, ...⟩	を *o*
└ すっぱ抜く *suppanuku* "expose"	

Figure 3.11: Candidate Pattern for すっぱ抜く *suppanuku* "expose" (2)

They have different case-roles, but the N3+*ni* of Figure 3.10 is an adjunct case-marker, and the two patterns should be merged.

3.5 Conclusion

In this thesis we present a method of assigning valency information and selectional restrictions to entries in a bilingual dictionary. The method exploits existing dictionaries and is based on two basic assumptions: words with similar meaning have similar subcategorization frames and selectional restrictions; and words with the same translations have similar meanings.

A prototype system allowed 6,327 new patterns to be built, using only simple human judgement (pre-filter). Of those more than 51% were usable as is, and more than 36% were usable with minor revisions, giving 87.7% potentially useful patterns. The cost, including human revisions, is less than 6 minutes per pattern. Furthermore, even before applying human revisions, adding the created patterns to a Japanese-to-English machine translation system improved the translation for 32% of sentences using these verbs, and degraded it for only 16%, a substantial improvement in quality.

Chapter 4

Acquisition of Valency Entries using Alternation Data

In this chapter, we present a method that uses alternation data to add new entries to an existing lexicon. If the existing lexicon has only one half of the alternation, then our method constructs the other half. The new entries have detailed information about argument structure and selectional restrictions. We also show that it is possible to simultaneously add entries in two languages if your exisiting lexicon has such information. In this section we focus on one class of alternations, but our method is applicable to any alternation[1].

4.1 Introduction

In this chapter we propose a method of acquiring detailed information about predicates, including argument structure, semantic restrictions on the arguments and translation equivalents. It combines two heterogeneous knowledge sources: a seed lexicon, and information about verbal alternations. Ultimately, we will use the method with a range of alternations, however, as a proof-of-concept in this section, we consider transitive alternations where the object of the transitive is the same as the subject of the intransitive (e.g. *the acid dissolved the metal* $\Leftrightarrow$ *the metal dissolved (in the acid)*) (Levin, 1993, 26–33). The algorithm can, however, be extended to other alternations.

[1]First we reported in Fujita and Bond (2004a), then revised in a journal, Fujita and Bond (2005).

We focus on acquiring Japanese verbs, using the valency (pattern) dictionary from the Japanese-to-English Machine Translation System **ALT-J/E** as our seed lexicon (See Section 2.1, 2.2). Using this, we actually create Japanese and English entries at the same time.

4.2 Alternations

Most verbs have more then one possible argument structure (subcat). These can be regularized into pairs of alternations, where two argument structures link similar semantic roles into different subcats. Over 80 alternation types have been identified for English (Levin, 1993). However, in this section we will only be considering those between transitive and intransitive uses of verbs, where the subject of the intransitive verb (**S**) is the same as the object of the intransitive verb (**O**). We will call the subject of the transitive verb **A** (absolutive).

In order to compare English and Japanese alternations, we compiled a list of 449 Japanese verbs that took transitive/intransitive alternations, based on data from Jacobsen (1981), Bullock (1999) and the Japanese/English dictionary EDICT (Breen, 1995). Japanese, unlike English, typically morphologicaly marks the transitivity alternation. A typical pair is given in (16) (See Figure 4.2 for more detail including both the subcategorization frame and the selectional restrictions).

(16)

	Vi		Vt
溶ける	***S*** *tokeru*	⇔	溶く ***A O*** *toku*
S *dissolve*		⇔	***A*** *dissolve* ***O***

To contrast the Japanese with English, we also investigated the English translations of the Japanese **S** = **O** transitive pairs. Many verbs had multiple translation equivalents, there were 839 Japanese-English pairs in all. The classification of the English types is given in Table 4.1.

We divide the entries into five classes. The first three are those where the main English verb is the same. The most common class (30%) is those where the English verb also allows the **S** = **O** transitive alternation. The next most common (20%) is entries where the Japanese intransitive verb can be translated by making the transitive verb's translation passive: ***A*** *omit* ***O/S*** *be omitted*. In the third class (6%) the English

Table 4.1: Classification of English Alternation

Japanese		English Translation (Structure)		Type	No.	(%)
Vi	Vt	Vi	Vt			
弱まる	弱める	**S** weaken	**A** weaken **O**	**S** = **O**	138	30.0
		(***S*** *Vi*	***A*** *Vt* ***O***)			
漏れる	漏らす	**S** be omitted	**A** omit **O**	passive	91	19.8
		(***S*** *be Vt-ed*	***A*** *Vt* ***O***)			
泣く	泣かす	**S** cry	**A** make **O** cry	synthetic	30	6.5
		S *Vi/be Adj*	***A*** *Vc* ***O*** *Vi/Adj*			
亡くなる	亡くす	**S** pass away	**A** lose **O**	—	197	42.8
		(***S*** *Vi*	***A*** *Vt* ***O***)			
じゃれる	じゃらす	**S** play	**A** play with **O**	—	4	0.9
		(***S*** *Vi*	***A*** *Vt prep* ***O***)			

Vc is control verb such as *make,get,let,become*.
Many entries also include information about non-core arguments/adjuncts.

is made transitive synthetically: a control verb (normally *make*) takes an intransitive verb or adjective as complement: ***S*** *cry/****A*** *make* ***O*** *cry*. The last two are those where either different translations are needed (44%), or the same English verb is used but the valency change is not one of those described above: ***S*** *play/****A*** *play with* ***O***. We show the details of this classification results in Appendix B.

From the point of view of constructing lexical entries, if the English main verb stays the same, then we can automatically construct a usable English translation equivalent along with the Japanese alternation. This should be possible 56.3% of the time. There are two caveats. The first is that the translation may not be the best one, most verbs can have multiple translations, and we are only creating one. The second is that this upper limit is almost certainly too low. For many of the alternations, although our table contained different verbs, translations using identical ones could also be constructed.

4.3 Comparing Selectional Restrictions of A, O and S

In alternations, a given semantic role can appear in two different syntactic positions: for example, the DISSOLVED role is the subject of intransitive *dissolve* and the object of the transitive. Baldwin et al. (1999) hypothesized that selectional restrictions (SRs) stay constant in the different syntactic positions. Dorr (1997), who generates both alternations from a single underlying representation also seems to make this assumption. Kilgarriff (1993), on the other hand, specifically makes the subject ⟨`+sentient, +volition`⟩, while the object is ⟨`+changes-state, +causally affected`⟩. In this section we attempt an empiric approach and measure the differences by examining the semantic classes used as SR of **A**, **O** and **S** of verbs in the **S** = **O** alternation.

Our source of data for the selectional restrictions is **ALT-J/E**'s valency (pattern) dictionary (Section 2.2). It consists of linked pairs of Japanese and English verbs. Both verbs have information about the argument structure (subcat) of the verb. In addition to the core arguments, adjunct cases are added to many patterns, to help in disambiguation (for more details, see Section 2.2). This is common in large NLP lexicons, such as COMLEX (Grishman et al., 1998), but rarely considered by linguists.[2] The Japanese side has selectional restrictions (SR) on the arguments. The arguments are linked between the two languages using case-roles.

Each English entry is separated into the **skeleton**, which gives the argument structure, and the **flesh**, which adds the predicate and any fixed arguments (such as prepositions, particles, and nouns in multiword expressions like *kick the bucket* (Yokoo et al., 1994)). There are 616 different **skeleton**s, with the most common ten (***A*** *Vt* ***O*** *,* ***S*** *be Adj*, ...) covering 72% of the entries. We show simplified examples of entries in Figures 4.2 and 4.3.[3] The flesh is shown underlined.

Bond et al. (2002) have previously identified alternation pairs between entries in the dictionary. Of those links, there were 449 pairs where **S** = **O**. The SRs take the form of a list of semantic classes, strings or ∗, which matches anything. The semantic classes are from the **Goi-Taikei** ontology of 2,710 categories (Ikehara et al., 1997). It is an unbalanced hierarchy with a maximum depth of 12 (Level 11). The top node (Depth 1, Level 0) is ⟨`1:noun`⟩. Depth 12 (Level 11) includes ⟨`1960:cultivation`⟩,

[2] For example, the COMLEX 3.0 entry for *surprised* notes that it coocurs with *about* and *at*.

[3] Actually, each entry has the same information with Figures 3.3 and 3.4.

⟨1993:appearance⟩ and so on. The lower the level, the more specialized the meaning, and thus the more restrictive the SR. Because ∗ matches anything, even non-nouns, it's the loosest restriction. Strings, which only match specific words, are the strictest restrictions.

Figure 4.1 shows the distribution of the depth of the semantic restrictions for the **A**, **O** and **S** arguments.

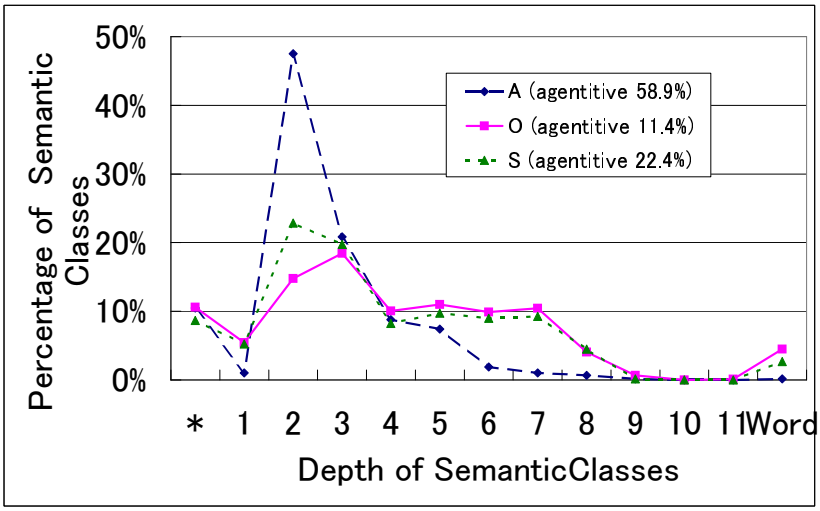


Figure 4.1: The Level of Semantic Classes

The absolute arguments (**A**) have the loosest restrictions. The most common depth is level 2, which includes ⟨3:agent⟩ and ⟨2:concrete⟩. The subject (**S**) and object (**O**) arguments show similar distributions, although **O** tends to be slightly more restrivtive.

This difference between **A** and the other two arguments was expected. SRs are used to distinguish senses.[4] In an intransitive verb, with only one argument, its SRs must have all the discrimination, and so should be relatively deep. In the transitive verb, when the object has deep semantic restrictions, the subject is not so important as a discriminator.

[4]In the **Goi-Taikei**, not just to disambiguate the Japanese sense, but also to choose the English translation.

In the **Goi-Taikei** hierarchy, semantic classes subsumed by ⟨3:agent⟩ are ⟨+sentient, +volition⟩. **A** was very agentitive, with 58.9% of the SRs being subsumed by ⟨3:agent⟩. **S** is slightly agentitive (22.4%) and **O** is the least agentitive.

In summary, the distribution of SRs is similar for the same semantic roles, even in the different grammatical positions of **S** and **O**. They are not, however, identical. In particular, **S** is more agentitive than **O**.

4.4 Method of Creating Valency Entries

In this section we describe how we create new entries. Our resources are (1) a seed lexicon of high quality hand-made entries; and (2) a list of verbal alternations. Our strategy is to look for verbs which participate in an alternation, but for which an entry exists for only one alternative. We then build the other alternate by a process of analogy with the known entries which participate in this alternation.

4.4.1 Target

In this experiment, we only look at one family of alternations, the **S** = **O** alternation. The candidate words are thus intransitive verbs with no transitive alternate, or transitive entries with no intransitive alternate. Alternations should be between entries, but the alternation list is of words. Many of the candidate words (those that have a entry for only one alternate) have several entries. Only some are suitable as seeds. We don't use entries which are intransitive lemmas but have an accusitive argument or which have both topic and nominative, such as (17).

(17) N1:⟨4:people⟩ は N3:⟨"力"⟩ が 抜ける
N1 *ha* *N3:"chikara"* *ga* *nukeru*
N1 TOP N3:power NOM lose
"N1 lose N1's energy"

There are 129 entries (25 lemmas) which have only intransitive entries, and 84 entries (40 lemmas) which have only transitive entries. We create intransitive entries using the existing transitive entries, and transitive entries using the existing intransitive entries.

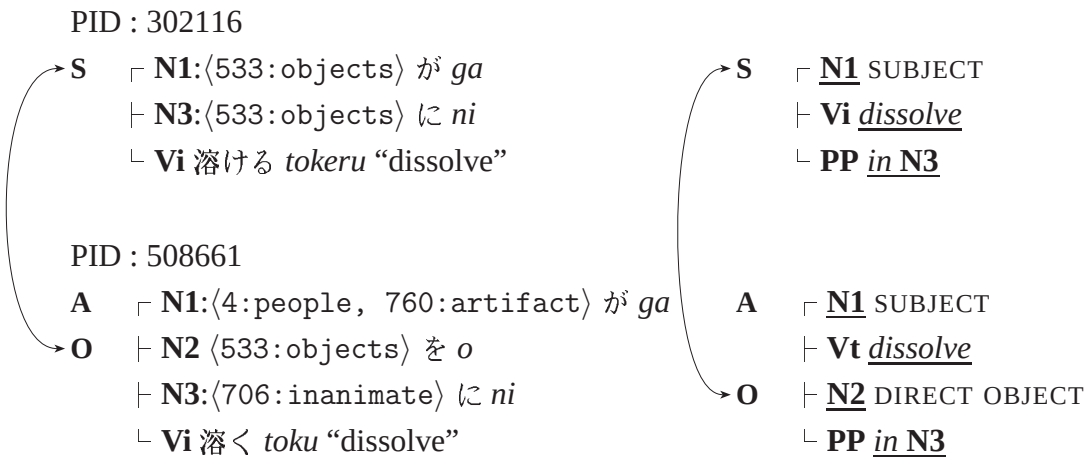


Figure 4.2: Existing Entries (which undergo the **S** = **O** alternation): 溶く *toku* “dissolve” ⇔溶ける *tokeru* “dissolve”

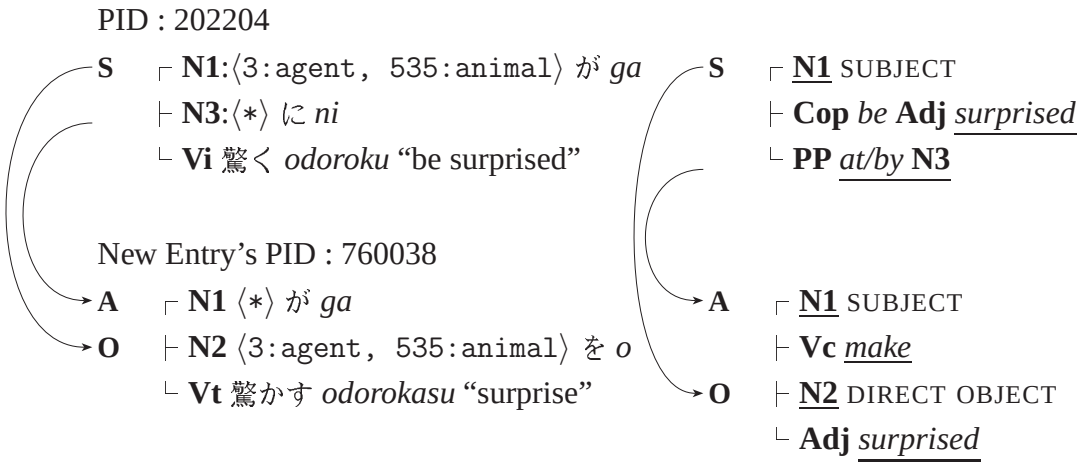


Figure 4.3: Seed: 驚く *odoroku* “be surprised” ⇒ New entry 驚かす *odorokasu* “surprise”

4.4.2 Creating the Japanese subcat and SRs

In creating the intransitive entries from the transitive entries, we map the **O**'s SRs onto the **S**'s SRs, and change the case marker from accusitive to nominative. We delete the **A** argument, and transfer any other arguments as they are.

To create the transitive entries, we map the intransitive **S**'s SRs onto the new **O**'s SRs, and give it an accusitive case-marker. Then we add a causitive argument as absolutive subject (**A**) with a default SR of ⟨3:agent⟩ and a nominative case-marker (⟨3:agent⟩ is the most frequent SR for transitive verbs undergoing this alternation).

If the intransitive entry has a demoted subject argument (where the Japanese case-marker is *ni* and the English preposition is *by*), We promote it to subject and use its SR instead of the default. We show this entry in Figure 4.3. The we use the same other case-frames as in the transitive entries.

4.4.3 Creating English Side

There are basically four choices for the English side: For the translation of a Japanese transitive, the English can be transitive (Vt), or an adjective/intransitive verb embedded in a control verb (Vc = synthetic): ***A*** *make* ***O*** *cry*. The intransitive side can be an intransitive verb (or adjective) (Vi), or a passive transitive (Vp = passive).

To creat an intransitive entry from a transitive, we see if the original translation was of type Vc or Vt. If it was Vc, then the complement of the control verb becomes the head of the new entry.[5]

If the transitive entry was Vt, then, if the English verb undergoes the **S** = **O** alternation, create an entry headed by an intransitive verb, otherwise passivize the verb. These operations are summarized in Figure 4.4. To judge whether an English verb could undergo the **S** = **O** alternation, we used the over-simple test that both transitive and intransitive entries appeared in our seed lexicon. This is reversed to make the transitive entries.

In the implementation, this process was made complicated by the presence of various extra arguments. Many adjuncts such as source or goal, were included in the seed lexicon to aid in selecting translations. In addition, many Japanese verbs were trans-

[5]We made a special rule for the English Vt *have*. In this case the intransitive alternation will be *There is*: for example, 「及ぼす」 ***A*** *have* ***O*** *on X* ⇒「及ぶ」 *There be* ***S*** *on X*

Creating Intransitive entries:

- if the original subcat has a control verb (*make,have,get,cause*)
  - **A** Vc **O** Vi/Adj $\Rightarrow$ **S** Vi/be Adj
- else (original head is Vt)
  - if the transitive head undergoes the **S** = **O** alternation
    - **A** Vt **O** $\Rightarrow$ **S** Vi
  - else
    - **A** Vt **O** $\Rightarrow$ **S** be Vt-ed

Creating Transitive Entries :
If the original subcat is:

- **S** Vi
  - if the intransitive head undergoes the **S** = **O** alternation
    S Vi $\Rightarrow$ **A** Vt **O**
  - else $\Rightarrow$ **A** Vc **O** Vi
- **S** be Adj $\Rightarrow$ **A** Vc **O** Adj
- **S** be Vt-ed $\Rightarrow$ **A** Vt **O**
- **S** Vt $\Rightarrow$ **A** Vt **O**

In this case, we use *make* as a control verb, Vc

Figure 4.4: Method of Creating the English Side

lated as verb-particle constructions or other multiword expressions. In cases were there was more than one candidate **skeleton** that could be used in the new entries, the most frequently used one in the known alternation examples was used. If there was a choice betwen unseen **skeleton**s, the most frquent overall was chosen. Finally, if an **skeleton** could not be found automatically, a default entry was used, with the expectation that it would need to be hand corrected. The defaults were: ***S*** *Vi* or ***S*** *be Vt-ed/Adj* (for the intransitive side); ***A*** *Vt* ***O*** (for the transitive side).

4.4.4 Evaluation

A total of 213 new entries were created for 65 verbs using the method outlined in Section 4.4. The quality was evaluated by expert lexicographers familiar with the seed lexicon. The evaluation was divided into two steps: (1) a decision as to whether or not the Japanese subcat and SRs gave a possible entry or not; (2) for the possible entries, how much hand correction was needed to make a correct entry.

4.4.5 Evaluation: Entry Possible/Impossible

The results of the judgement as to whether the Japanese subcat and SRs gave a possible entry or not are given in Table 4.2. A majority, (65%) were possible. Looking at the results per verb (recalling that one verb can have multiple entries), there was at least one possible entry for every verb. 69% of the verbs had at least one entry that was usable as is.

Table 4.2: Is the Japanese Expression possible?

	Possible		Impossible		Total	
Created	No.	%	No.	%	Entries	(Verbs)
`Vi`	45	53.6	39	46.4	84	(25)
`Vt`	93	72.1	36	27.9	129	(40)
Total	138	64.8	75	35.2	213	(65)

An example of an impossible entry is (18). The expression 捕らわれる*torawareru* “be caught” is possible, but not with the semantic restriction

⟨2:concrete, 2306:material-phenomenon⟩ on the subject and the adjunct case shown. Another entry created for *torawareru* (19) was judged to be possible. We discuss in more detail in Section 4.5.1.

(18) * N1:⟨2:concrete, 2306:material-phenomenon⟩ が 捕らわれる
N1 *ga* *torawareru*
N1 NOM be picked up
" N1 be caught"

(19) N1:⟨4:people, 535:animal, 760:artifact⟩ が 捕らわれる
N1 *ga* *torawareru*
N1 NOM be caught
"N1 be caught"

4.4.6 Evaluation: Fine Tuning

For entries where the basic Japanese structure was close to being correct, the lexicographers hand-corrected them to be good entries. In the next two sections we look at how much correction was needed, for first the Japanese, and then the English halves. All results are given looking only at those entries judged as possible: 45 intransitive and 138 transitive entries.

4.4.7 Japanese Side

The Japanese results are summarized in Table 4.3 (note: a single entry may have more than one part corrected). 82% of the entries needed no correction. In particular, the Vi entries were good 93% of the time. The most common change was to tweak the semantic restrictions.

The changes in SRs of the transitive verbs are shown in Table 4.4. Most of the time the change was in the **A** argument (which was given the restriction ⟨3:agent⟩ by default). The majority of the corrections were making the **A**'s SR more restrictive: changing the semantic class to its descendant. The corrections to the **O**'s SR were various. They were mainly made to reflect the fact that not all uses of a verb can alternate, the **O**'s SR is not always the same as the **S**'s, as was showed in Section 4.3.

Table 4.3: Japanese Evaluation (Fine Tuning)

Part Corrected	Vi Created		Vt Created		Total	
	No.	%	No.	%	No.	%
SRs	2	4.4	19	20.4	21	15.2
Case-role and Case-marker	2	4.4	1	1.1	3	2.2
Case-marker (only)	0	0	2	2.2	2	1.4
Japanese O.K.	42	**93.3**	71	**76.3**	113	**81.9**
	45 entries		93 entries		138 entries	

Table 4.4: Analysis of Corrected SR Vt

How to Correct SR	Case-role (No.)			
	A	**O**	**S**	**XI**
Add	2	4	0	0
Delete	0	1	0	0
Subsumed or Lower Level SR	9	0	1	0
Lower Level SR (not subsumed)	2	1	0	0
Same Level SR	0	1	0	0
Higher Level SR	0	0	0	1
Total	13	7	1	1

Table 4.5: English Evaluation (Fine Tuning)

Part Corrected	Vi Created		Vt Created		Total	
	No.	%	No.	%	No.	%
English Verb	14	31.1	34	36.6	48	34.8
Subcat	14	31.1	34	36.6	48	34.8
Other element	16	35.6	42	45.2	58	42.0
English O.K.	29	64.4	49	52.7	78	56.5
Both Japanese and English O.K.	29	**64.4**	48	**51.6**	77	**55.8**
	45 entries		93 entries		138 entries	

4.4.8 English Side

As we predicted in Section 4.2, good English translations with the same main verb could only be made around 56% of the time. Although many entries had to be corrected, the lexicographers found it to be faster than creating the new entries from scratch.

The results are given in Table 4.5 (note: a single entry may have more than one part corrected). We managed to make almost exactly as many good English entries as we predicted. In general the intransitive entries are better than the transitive ones. This is because we was adding information to make the transitives, and deleting it to make the intransitives.

4.5 Discussion

The above results show that alternations can be used to create rich monolingual entries, and to some extent bilingual entries. In this section we discuss some of the reasons for errors, and suggest ways to improve the method.

4.5.1 Rejecting Impossible Candidates

To make the construction fully automatic, a test for whether a Japanese entry is possible or not is required.

One possibility would be to add a corpus based filter: if no entries can be found that fit the entry, then it should be rejected. The problem with this approach is that many of the entries we created were for infrequent verbs. The average frequency in 16 years of Japanese newspaper text was only 173, and 22 verbs never appeared, although all were familiar to native speakers.

Another, more hopeful, possibility is to learn a classifier based on features in the entries themselves. The agentitivity of the SR of the created intransitive entries (the most problematic group, see Table 4.2), seems a good cue, only 5.6% of the SRs of the impossible entries were subsumed by ⟨`3:agent`⟩, compared to 14.9% of the possible entries. A classifier could also be trained on the known alternation pairs, although they provide only positive evidence.

4.5.2 Improving the English Translations

The numbers of the different types of translations are compared for the reference data (Section 4.2), the entries created by our method (Section 4.4) and the entries corrected by expert lexicographers (Section 4.4.6) in Table 4.6. The first three rows show entries with the same English main verb.

Table 4.6: A Comparison of Reference Data with Created Alternations

English Structure		Reference		Vi Created		Vi Corrected		Vt Created		Vt Corrected	
Vi	Vt	No.	(%)	No.	(%)	No.	(%)	No.	(%)	No.	(%)
S *Vi*	***A*** *Vt* ***O***	138	30.0	16	35.6	5	12.8	31	33.3	17	22.7
S *be Vt-ed*	***A*** *Vt* ***O***	91	19.8	27	60.0	34	87.2	8	8.6	8	10.7
S *Vi/be Adj*	***A*** *Vc* ***O*** *Vi/Adj*	30	6.5	0	0	0	0	44	47.3	32	42.7
Different Head		10	10.8	2	4.4	0	0	10	10.8	18	24.0
Total		259	56.3	45	100	39	100	93	100	75	100

We focus on three discrepancies:

1. In the class "***S*** *Vi* $\Leftrightarrow$ ***A*** *Vt* ***O***", There are fewer "Corrected" entries than "Created" for both Vi and Vt.

2. In the class "***S*** *Vi/be Adj* $\Leftrightarrow$ ***A*** *Vc* ***O*** *Vi/Adj*", there are no entries for Vi but many for Vt.

3. The total No. of "Corrected" is less than "Created". For 24 entries (17.4 %) the lexicographers chose a different English translation.

The first discrepancy is caused by our implementation over estimating the number of English verbs that undergo the **S** = **O** alternation. We used the very simple approximation that any English verb that had both transitive and intransitive entries in our lexicon could undergo the alternation. This overestimates for two reasons (i) the verbs may have different meanings, and thus not be alternations at all; (ii) the verbs may undergo other alternations, such as **A** = **S**. Looking at the corrected data, in Vi, 7 entries are corrected to "***S*** *be Vt-ed*" from ***S*** *Vi*. In Vt, 3 entries are corrected to "***A*** *Vc*

O** Vi/Adj*" and 6 to different English verbs and the other 6 are corrected to intransitive verbs in different constructions, such as (20) (created from 輝やく ***S:⟨555:face⟩*-ga [**X:**⟨1000:abstract⟩-de] kagayaku* "**S** shine with **X**").

(20) N1:⟨3:agent⟩ が N2:⟨555:face⟩ を N12:⟨1000:abstract⟩ に/で 輝かす
N1 *ga* *N2* *o* *N12* *ni/de kagayakasu*
N1 NOM N2 ACC N12 by shine
"N1's N2 shine with N12"

The second discrepency is in the frequency of the control verb construction. In Vi, no original transitive entry used control verbs. In general, when the lexicographers create an entry, they prefer a simple entry to a synthetic one. Looking at the linguists' reference data, about 6.5% of the examples used control verbs. In the constructed data, 47.3 % (44 entries) use the control verb *make*, more than any other category. Of those 44 entries, 17 entries are corrected by the lexicographers. 8 entries are corrected to intransitive verbs; 5 entries are corrected to different English head and 2 were corrected to different control verbs (*let* or *have*), the remain 2 were corrected to ***A** Vt **O***. For example, when the original intransitive entry is *N1 be exhausted*, *exhausted* is defined as adjective in the existing dictionary. So we create a new entry *N1 make N2 exhausted*$_{\text{adj}}$. However, because *exhaust* is a transitive verb, it was corrected to *N1 exhaust N2*. The algorithm needs to optionally convert adjectives to verbs in cases where there is overlap between the adjective and past particle.

Finally, we consider those Japanese alternations where the transitive and intransitive alternatives need translations with diiferent English main verbs. A good example of this is Vi 亡くなる *nakunaru* "**S** pass away" and Vt 亡くす *nakusu* "**A** lose **O**".[6] These are impossible to generate using our method. Even with reliable English syntactic data, it would be hard to rule out *pass away* as a possible transitive verb or *lose* as an intransitive. They can only be ruled out by using data linking the subcat with the meaning, and this would need to be linked to the Japanese verbs' meanings. This may become possible with larger linked multi-lingual dictionaries, such as those under construction in the Papillon project[7], but is not now within our reach.

In summary, we could improve the construction of the English translations by using richer English information, especially about alternations.

[6] *My friend passed away* ⇔ *I lost my friend*.

[7] http://www.papillon-dictionary.org/

4.6 Future Work

This research can be extended in four ways. The first is to work on automatically rejecting impossible Japanese entries. The second is to use richer English information about alternations to improve the quality of the English entries. Both of these will will improve the quality of the created entries.

The third is to apply the method to other alternations, using either linguists' data or automatically acquired alternations (Oishi and Matsumoto, 1997; Bond et al., 2002). The last is to carry out a task based evaluation, using the extended dictionary in a machine translation system.

In addition, new entries are being added to the seed lexicon from a variety of sources. When these new entries are one half of a known alternation, we apply this method to create the other half. Even with just this one alternation, we have already added a hundred new entries in this manner (although not all were correct).

We hope that our extended lexicon will be useful not only for NLP applications, but for research into the nature of alternations themselves.

4.7 Conclusion

We presented a method that uses alternation data to add new entries to an existing lexicon. The new entries have detailed information about argument structure and selectional restrictions. If the existing lexicon has only one half of the alternation, then our method constructs new Japanese entries with 69% of the verbs having one or more correct entries. We also showed that it is possible to simultaneously add entries to a second language with a reduced accuracy of 56% if your exisiting lexicon has such information. In this section we focused on one class of alternations, but it is applicable to any alternation.

Chapter 5

Exploiting Semantic Information for HPSG Parse Selection

In this chapter, we investigate the use of semantic information in parse selection.We present that sense-based semantic features combined with ontological information are effective for parse selection. Training and testing on the definition and example subset of the **Hinoki** corpus (See Section 2.4), a combined model give a improvement in parse selection accuracy over a model using only syntactic features[1].

5.1 Introduction

Recently, significant improvements have been made in combining symbolic and statistical approaches to various natural language processing tasks. In parsing, for example, symbolic grammars are combined with stochastic models (Oepen et al., 2004; Malouf and van Noord, 2004). Much of the gain in statistical parsing using lexicalized models comes from the use of a small set of function words (Klein and Manning, 2003). Features based on general relations provide little improvement, presumably because the data is too sparse: in the Penn treebank standardly used to train and test statistical parsers *stocks* and *skyrocket* never appear together. However, the superordinate concepts *capital* ($\supset$ *stocks*) and *move upward* ($\supset$ *sky rocket*) frequently appear together, which suggests that using word senses and their hypernyms as features may be useful

[1]We reported about this experiment in Fujita et al. (2007).

However, to date, there have been few combinations of sense information together with symbolic grammars and statistical models. We hypothesize that one of the reasons for the lack of success is that there has been no resource annotated with both syntactic and semantic information. In this chapter, we use a **Hinoki** corpus (See Section 2.4) , with both syntactic information (HPSG parses) and semantic information (sense tags from a **Lexeed** lexicon (See Section 2.3)). We use this to train parse selection models using both syntactic and semantic features. A model trained using syntactic features combined with semantic information outperforms a model using purely syntactic information by a wide margin (69.4% sentence parse accuracy vs. 63.8% on definition sentences).

5.2 Parse Selection

Combining the broad-coverage **JACY** grammar and the **Hinoki** corpus, we build a parse selection model on top of the symbolic grammar. Given a set of candidate analyses (for some Japanese string) according to **JACY**, the goal is to rank parse trees by their probability: training a stochastic parse selection model on the available treebank, we estimate statistics of various features of candidate analyses from the treebank. The definition and selection of features, thus, is a central parameter in the design of an effective parse selection model.

5.2.1 Syntactic Features

The first model that we trained uses syntactic features defined over HPSG derivation trees as summarized in Table 5.1. For the closely related purpose of parse selection over the English Redwoods treebank, Toutanova et al. (2005) train a discriminative log-linear model, using features defined over *derivation trees* with non-terminals representing the *construction types* and *lexical types* of the HPSG grammar. The basic feature set of our parse selection model for Japanese is defined in the same way (corresponding to the PCFG-S model of Toutanova et al. (2005)): each feature capturing a sub-tree from the derivation limited to depth one. Table 5.1 shows example features extracted from our running example (Figure 2.10 in Section 2.4) in our MaxEnt models. In Table 5.1, the feature template #1 corresponds to local derivation sub-trees. We

Table 5.1: Example structural features (SYN-1 and SYN-GP) extracted from the derivation tree in Figure 2.10

#	sample features
1	⟨0 rel-cl-sbj-gap hd-complement noun-le⟩
1	⟨1 frag-np rel-cl-sbj-gap hd-complement noun-le⟩
1	⟨2 △ frag-np rel-cl-sbj-gap hd-complement noun-le⟩
2	⟨0 rel-cl-sbj-gap hd-complement⟩
2	⟨0 rel-cl-sbj-gap noun-le⟩
2	⟨1 frag-np rel-cl-sbj-gap hd-complement⟩
2	⟨1 frag-np rel-cl-sbj-gap noun-le⟩
3	⟨1 conj-le ya⟩
3	⟨2 noun-le conj-le ya⟩
3	⟨3 ⊢ noun-le conj-le ya⟩
4	⟨1 conj-le⟩
4	⟨2 noun-le conj-le⟩
4	⟨3 ⊢ noun-le conj-le⟩

The first column numbers the feature template corresponding to each example; in the examples, the first integer value is a parameter to feature templates, i.e. the depth of grandparenting (types #1 and#2) or *n*-gram size (types #3 and #4). The special symbols △ and ⊢ denote the root of the tree and left periphery of the yield, respectively.

will refer to the parse selection model using only local structural features as SYN-1.

Dominance Features

To reduce the effects of data sparseness, feature type #2 in Table 5.1 provides a back-off to derivation sub-trees, where the sequence of daughters is reduced to just the head daughter. Conversely, to facilitate sampling of larger contexts than just sub-trees of depth one, feature template #1 allows optional grandparenting, including the upwards chain of dominating nodes in some features. In our experiments, we found that grandparenting of up to three dominating nodes gave the best balance of enlarged context *vs.* data sparseness. Enriching our basic model SYN-1 with these features we will hence-

forth call SYN-GP.

N-Gram Features

In addition to these dominance-oriented features taken from the derivation trees of each parse tree, our models also include more surface-oriented features, viz. *n*-grams of lexical types with or without lexicalization. Feature type #3 in Table 5.1 defines *n*-grams of variable size, where (in a loose analogy to part-of-speech tagging) sequences of lexical types capture syntactic category assignments. Feature templates #3 and #4 only differ with regard to lexicalization, as the former includes the surface token associated with the rightmost element of each *n*-gram (loosely corresponding to the emission probabilities in an HMM tagger). We used a maximum *n*-gram size of two in the experiments reported here, again due to its empirically determined best overall performance.

5.2.2 Semantic Features

In order to define semantic parse selection features, we use a reduction of the full semantic representation (MRS) into 'variable-free' *elementary dependencies*. The conversion centrally rests on a notion of one *distinguished* variable in each semantic relation. For most types of relations, the distinguished variable corresponds to the main index (ARG0 in the examples above), e.g. an event variable for verbal relations and a referential index for nominals. Assuming further that, by and large, there is a unique relation for each semantic variable for which it serves as the main index (thus assuming, for example, that adjectives and adverbs have event variables of their own, which can be motivated in predicative usages at least), an MRS can be broken down into a set of basic dependency tuples of the form shown in Figure 2.9 (Oepen and Lønning, 2006).

All predicates are indexed to the position of the word or words that introduced them in the input sentence (`<start:end>`). This allows us to link them to the sense annotations in the corpus.

Basic Semantic Dependencies

The basic semantic model, SEM-Dep, consists of features based on a predicate and its arguments taken from the elementary dependencies. For example, consider the

Table 5.2: Example semantic features (SEM-Dep) extracted from the dependency tree in Figure 2.9.

#	**sample features**
20	⟨0 _unten_s ARG1 _hito_n_1 ARG2 _ya_p_conj⟩
20	⟨0 _ya_p_conj LIDX _densha_n_1 RIDX _jidousha_n_1⟩
21	⟨1 _unten_s ARG1 _hito_n_1⟩
21	⟨1 _unten_s ARG2 _jidousha_n_1⟩
21	⟨1 _ya_p_conj LIDX _densha_n_1⟩
21	⟨1 _ya_p_conj RIDX _jidousha_n_1⟩
22	⟨2 _unten_s _hito_n_1 _jidousha_n_1⟩
23	⟨3 _unten_s _hito_n_1⟩
23	⟨3 _unten_s _jidousha_n_1⟩
...	

dependencies for *densha ya jidousha-wo unten suru hito* "a person who drives a train or car" given in Figure 2.9. The predicate *unten* "drive" has two arguments: ARG1 *hito* "person" and ARG2 *jidousha* "car".

From these, we produce several features (See Table 5.2). One has all arguments and their labels (#20). We also produce various back offs: #21 introduces only one argument at a time, #22 provides unlabeled relations, #23 provides one unlabeled relation at a time and so on.

Each combination of a predicate and its related argument(s) becomes a feature. These resemble the basic semantic features used by Toutanova et al. (2005). We further simplify these by collapsing some non-informative predicates, e.g. the `unknown` predicate used in fragments.

Word Sense and Semantic Class Dependencies

We created two sets of features based only on the word senses. For SEM-WS we used the sense annotation to replace each underspecified MRS predicate by a predicate indicating the word sense. This used the gold standard sense tags. For SEM-Class, we used the sense annotation to replace each predicate by its **Goi-Taikei** semantic class.

Table 5.3: Example semantic class features (SEM-Class).

#	sample features
40	⟨0 _unten_s ARG1 C4 ARG2 C988⟩
40	⟨1 C2003 ARG1 C4 ARG2 C988⟩
40	⟨1 C2003 ARG1 C4 ARG2 C988⟩
40	⟨0 _ya_p_conj LIDX C988 RIDX C988⟩
41	⟨2 _unten_s ARG1 C4⟩
41	⟨2 _unten_s ARG2 C988⟩
...	

In addition, to capture more useful relationships, conjunctions were followed down into the left and right daughters, and added as separate features. The semantic classes for 電車$_1$*densha* "train" and 自 動車$_1$*jidousha* "car" are both ⟨988:land vehicle⟩, while 運転$_1$ *unten* "drive" is ⟨2003:motion⟩ and 人$_4$ *hito* "person"is ⟨4:human⟩. The sample features of SEM-Class are shown in Table 5.3.

These features provide more specific information, in the case of the word sense, and semantic smoothing in the case of the semantic classes, as words are binned into only 2,700 classes.

Superordinate Semantic Classes

We further smooth these features by replacing the semantic classes with their hypernyms at a given level (SEM-L). We investigated levels 2 to 5. Predicates are binned into only 9 classes at level 2, 30 classes at level 3, 136 classes at level 4, and 392 classes at level 5.

For example, at level 3, the hypernym class for ⟨988:land vehicle⟩ is ⟨706:inanimate⟩, ⟨2003:motion⟩ is ⟨1236:human activity⟩ and ⟨4:human⟩ is unchanged. So we used ⟨706:inanimate⟩ and ⟨1236:human activity⟩ to make features in the same way as Table 5.3.

An advantage of these underspecified semantic classes is that they are more robust to errors in word sense disambiguation — fine grained sense distinctions can be ignored.

Valency Dictionary Compatability

The last kind of semantic information we use is valency information, taken from the Japanese side of the **Goi-Taikei** Japanese-English valency dictionary as extended by Chapter 3. This valency dictionary has detailed information about the argument properties of verbs and adjectives, including subcategorization and selectional restrictions (For more details, see Section 2.2). A simplified entry of the Japanese side for 運転する*unten-suru* "drive" is shown in Figure 5.1.

Each entry has a predicate and several case-slots. Each case-slot has information such as grammatical function, case-marker, case-role (N1, N2, ...) and semantic restrictions. The semantic restrictions are defined by the **Goi-Taikei**'s semantic classes.

On the Japanese side of **Goi-Taikei**'s valency dictionary, there are 10,146 types of verbs giving 18,512 entries and 1,723 types of adjectives giving 2,618 entries.

PID : 300513 (PID is the verb's Pattern ID.)
┌ **N1** ⟨4:people⟩ が *ga*
├ **N2** ⟨986:vehicles⟩ を *o*
└ 運転する *unten-suru* "drive"

Figure 5.1: 運転する*unten-suru* "N1 drive N2".

The valency based features were constructed by first finding the most appropriate pattern, and then recording how well it matched.

To find the most appropriate pattern, we extracted candidate dictionary entries whose lemma is the same as the predicate in the sentence: for example we look up all entries for 運転する *unten-suru* "drive". Then, for each candidate pattern, we mapped its arguments to the target predicate's arguments via case-markers. If the target predicate has no suitable argument, we mapped to comitative phrase. Finally, for each candidate patterns, we calculate a matching score[2] and select the pattern which has the best score.

Once we have the most appropriate pattern, we then construct features that record how good the match is (Table 5.4). These include: the total score, with or without the verb's Pattern ID (High/Med/Low/Zero: #31 0,1), the number of filled arguments (#31

[2]The scoring method follows Bond and Shirai (1997), and depends on the goodness of the matches of the arguments.

Table 5.4: Example semantic features (SP)

#	sample features
31	⟨0 High⟩
31	⟨1 300513 High⟩
31	⟨2 2⟩
31	⟨3 R:High⟩
31	⟨4 300513 R:High⟩
32	⟨1 _unten_s High⟩
32	⟨4 _unten_s R:High⟩
33	⟨5 N1 C High⟩
33	⟨7 C⟩
...	

2), the fraction of filled arguments vs all arguments (High/Med/Low/Zero: #31 3,4), the score for each argument of the pattern (#32 5) and the types of matches (#32 5,7).

These scores allow us to take advantage of information about word usage in an exisiting dictionary.

5.3 Evaluation and Results

We trained and tested on a subset of the dictionary definition and example sentences in the **Hinoki** corpus. This consists of those sentences with ambiguous parses which have been annotated so that the number of parses has been reduced (Table 5.5). That is, we excluded unambiguous sentences (with a single parse), and those where the annotators judged that no parse gave the correct semantics. This does not necessarily mean that there is a single correct parse, we allow the annotator to claim that two or more parses are equally appropriate.

Dictionary definition sentences are a different genre to other commonly used test sets (e.g. newspaper text in the Penn Treebank or travel dialogues in Redwoods). However, they are valid examples of naturally occurring texts and a native speaker can read and understand them without special training. The main differences with

Table 5.5: Data of Sets for Evaluation

Corpus		# Sents	Length (Ave)	Parses/Sent (Ave)
Definitions	Train	30,345	9.3	190.1
	Test	2,790	10.1	177.0
Examples	Train	27,081	10.9	74.1
	Test	2,587	10.4	47.3

newspaper text is that the definition sentences are shorter, contain more fragments (especially NPs as single utterances) and fewer quoting and proper names. The main differences with travel dialogues is the lack of questions.

5.3.1 A Maximum Entropy Ranker

Log-linear models provide a very flexible framework that has been widely used for a range of tasks in NLP, including parse selection and reranking for machine translation. We use a *maximum entropy / minimum divergence* (MEMD) modeler to train the parse selection model. Specifically, we use the open-source **Toolkit for Advanced Discriminative Modeling** (TADM:[3] Malouf, 2002) for training, using its *limited-memory variable metric* as the optimization method and determining best-performing convergence thresholds and prior sizes experimentally. A comparison of this learner with the use of support vector machines over similar data found that the SVMs gave comparable results but were far slower (Baldridge and Osborne., 2007). Because we are investigating the effects of various different features, we chose the faster learner.

5.3.2 Results

The results for most of the models discussed in the previous section are shown in Table 5.6. The accuracy is exact match for the entire sentence: a model gets a point only if its top ranked analysis is the same as an analysis selected as correct in **Hinoki**. This is a stricter metric than component based measures (e.g., labelled precision) which

[3] `http://tadm.sourceforge.net`

Table 5.6: Parse Selection Results

Method	Definitions		Examples	
	Accuracy (%)	Features (×1000)	Accuracy (%)	Features (×1000)
SYN-1	52.8	7	67.6	8
SYN-GP	62.7	266	76.0	196
SYN-ALL	**63.8**	316	**76.2**	245
SYN baseline	16.4	random	22.3	random
SEM-Dep	57.3	1,189	58.7	675
+SEM-WS	56.2	1,904	59.0	1,486
+SEM-Class	57.5	2,018	59.7	1,669
+SEM-L2	60.3	808	62.9	823
+SEM-L3	59.8	876	62.8	879
+SEM-L4	59.9	1,000	62.3	973
+SEM-L5	60.4	1,240	61.3	1,202
+SP	59.1	1,218	68.2	819
+SEM-ALL	**62.7**	3,384	**69.1**	2,693
SYN-SEM	**69.5**	2,476	**79.2**	2,126
SEM baseline	20.3	random	22.8	random

award partial credit for incorrect parses. For the syntactic models, the baseline (random choice) is 16.4% for the definitions and 22.3% for the examples. Definition sentences are harder to parse than the example sentences. This is mainly because they have more relative clauses and coordinate NPs, both large sources of ambiguity. For the semantic and combined models, multiple sentences can have different parses but the same semantics. In this case all sentences with the correct semantics are scored as good. This raises the baselines to 20.3 and 22.8% respectively.

Even the simplest models (SYN-1 and SEM-Dep) give a large improvement over the baseline. Adding grandparenting to the syntactic model has a large improvement (SYN-GP), but adding lexical n-grams gave only a slight improvement over this (SYN-ALL).

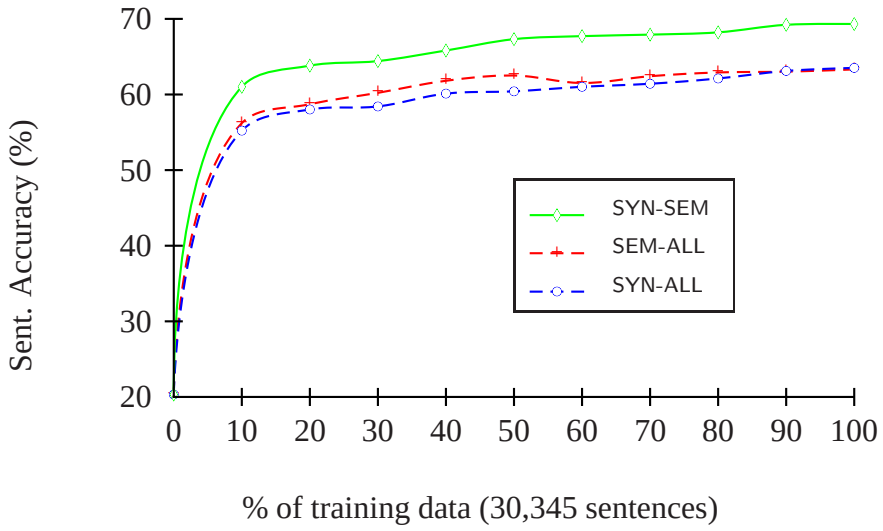


Figure 5.2: Learning Curves (Definitions)

The effect of smoothing by superordinate semantic classes (SEM-Class), shows a modest improvement. The syntactic model already contains a back-off to lexical-types, we hypothesize that the semantic classes behave in the same way. Surprisingly, as we add more data, the very top level of the semantic class hierarchy performs almost as well as the more detailed levels. The features using the valency dictionary (SP) also provide a considerable improvement over the basic dependencies.

Combining all the semantic features (SEM-ALL) provides a clear improvement, suggesting that the information is heterogeneous. Finally, combing the syntactic and semantic features gives the best results by far (SYN-SEM: SYN-ALL + SEM-Dep + SEM-Class + SEM-L2 + SP). The definitions sentences are harder syntactically, and thus get more of a boost from the semantics. The semantics still improve performance for the example sentences.

The semantic class based sense features used here are based on manual annotation, and thus show an upper bound on the effects of these features. This is not an absolute upper bound on the use of sense information — it may be possible to improve further through feature engineering. The learning curves (Fig 5.2) have not yet flattened out. We can still improve by increasing the size of the training data.

5.4 Discussion

Bikel (2000) combined sense information and parse information using a subset of SemCor (with WordNet senses and Penn-II treebanks) to produce a combined model. This model did not use semantic dependency relations, but only syntactic dependencies augmented with heads, which suggests that the deeper structural semantics provided by the HPSG parser is important. Xiong et al. (2005) achieved only a very minor improvement over a plain syntactic model, using features based on both the correlation between predicates and their arguments, and between predicates and the hypernyms of their arguments (using HowNet). However, they do not investigate generalizing to different levels than a word's immediate hypernym. Resently, Agirre et al. (2008) shows that semantic classes help to obtain significant improvement in both parsing and PP attachment tasks. They tested using English dataset: Penn Treebank, SemCor and WordNet.

Pioneering work by Toutanova et al. (2005) and Baldridge and Osborne. (2007) on parse selection for an English HPSG treebank used simpler semantic features without sense information, and got a far less dramatic improvement when they combined syntactic and semantic information.

The use of hand-crafted lexical resources such as the **Goi-Taikei** ontology is sometimes criticized on the grounds that such resources are hard to produce and scarce. While it is true that valency lexicons and sense hierarchies are hard to produce, they are of such value that they have already been created for all of the languages we know of which have large treebanks. In fact, there are more languages with WordNets than large treebanks.

In future work we intend to confirm that we can get improved results with raw sense disambiguation results not just the gold standard annotations and test the results on other sections of the **Hinoki** corpus. To get the high-quality raw sense disambiguation results, we propose a word sense disambiguation method in Chapter 6.

5.5 Conclusions

We have shown that sense-based semantic features combined with ontological information are effective for parse selection. Training and testing on the definition subset

of the **Hinoki** corpus, a combined model gave a 5.6% improvement in parse selection accuracy over a model using only syntactic features (63.8% → 69.4%). Similar results (76.2% → 79.2%) were found with example sentences.

Chapter 6

Word Sense Disambiguation using Disambiguated Superordinate Semantic Classes

In this chapter, we propose a new method for word sense disambiguation (WSD) using superordinate semantic classes. We separate WSD into two stages. In the first stage, we determine superordinate semantic classes, then in the second stage we determine fine-grained word senses using the results of the first stage. In the second stage, by using superordinate semantic classes, we show an improvement over the best published method of Japanese dictionary-based lexical-sample task of **SENSEVAL-2** 2. In addition, we show the effectiveness of superordinate semantic classes for unseen words [1].

6.1 Introduction

In this chapter, we propose a method for word sense disambiguation (WSD) using Superordinate Semantic Classes. Many words have multiple meanings, and they change depending on the context. WSD has been shown to be useful in a variety of NLP applications including parse selection (Fujita et al., 2007) and machine translation (Chan et al., 2007).

[1] We reported this in Fujita et al. (2008)

There is much previous research on WSD. In the SENSEVAL-2 Japanese lexical task, supervised systems using a large number of shallow features did the best (Murata et al., 2003). More recently, unsupervised approaches such as extended Lesk have been shown to do well (Baldwin et al., 2008), although they are beaten by supervised approaches using both semantic and syntactic features (Tanaka et al., 2007).

However, we would like to use the WSD results to improve the accuracy of our parsing, so we cannot use the results of syntactic analysis to restrict the senses. So, in this chapter, we propose a WSD method that does not use syntactic information.

Now, we consider the reason why supervised WSD is difficult. One important reason is the difficulty of constructing enough training data for large sense inventories. When there are several tens or hundreds of thousands of word senses, it is very difficult to get enough training data for all the words. In addition, the number of classes makes it hard to train standard machine learning tools.

Because of that, we separate WSD into 2 stages. In the first stage, we guess higher-level (superordinate) semantic classes, such as *person, place, thing, event*. In the second stage, we deduce detailed word senses using the superordinate semantic classes as constraints.

It is easier to disambiguate the superordinate semantic classes because the number of higher level classes is much less than the number of word senses, therefore we can get enough accuracy using relatively less training data. In addition, many word senses can already be decided just from the superordinate semantic class. This approach is similar to Kohomban and Lee (2005), who used **WordNet** (Fellbaum, 1998) unique beginners (25 for nouns and 15 for verbs) which effectively divide **WordNet** senses into coarser superordinate classes.

In the next section, we describe the resources which we use. Then, in Section 6.3, we describe the superordinate semantic class disambiguation. In Section 6.4, we describe WSD using the superordinate semantic classes. In Section 6.5 and Section 6.6, we discuss and describe future work, before concluding in Section 6.7.

6.2 Resources

We use the **Hinoki** corpus (See Section 2.4) to train both the superordinate semantic class models (in Section 6.3) and the full WSD models (in Section 6.4).

INDEX	ライター *raitâ* “lighter/writer/raita”	
POS	noun Lexical-type `noun-lex`	
FAMILIARITY	6.2 [1–7]	
SENSE 1	DEFINITION	**点**火$_1$ 器 。特に$_1$ 、 たばこ$_2$ に 火$_1$ を 点ける$_{17}$ ため の 用具$_1$ 。 a device for lighting things, especially cigarettes.
	EXAMPLE	彼$_3$ は ライター$_1$ で たばこ$_2$ に 火$_1$ を 点け$_{17}$ た 。 He lit the cigarette with his lighter
	HYPERNYM	用具$_1$ *yougu* “device”
	SEM. CLASS	⟨`915:household appliance`⟩ (⊂ ⟨`706:inanimate`⟩)
	IWANAMI	53815,0-0-1-0 (L ≃ R(2/3), L ⊃ R(1/3))

Where Sem. Classes come from **Goi-Taikei**.; ⊂ shows subsumption (not necessarily direct).

Figure 6.1: Entry for ライター$_1$ *raitâ* “lighter” from **Lexeed**

The **Hinoki** corpus was annotated with both syntactic parses and semantic information (HPSG parses and sense tags from **Lexeed**, (Tanaka et al., 2006), see Section 2.4), but in this section, we don’t use the syntactic information, as we wish to use the WSD results in parse selection in future work.

We described resources which we use in Chapter 2. we also show a (simplified) example of an **Lexeed** entry in Figure 6.1.

All words in the 28,000 word fundamental vocabulary of **Hinoki** are tagged with word senses of **Lexeed**, which are in turn linked to the **Goi-Taikei** semantic classes. Any words outside of this vocabulary are untagged. For example, the word たばこ *tabako* “cigarette” (of example sentence in Figure 6.1) is tagged as sense 2 in the example sentence, with the meaning “cigarette” not “tobacco plant” and this has the semantic class ⟨`862:cigarette`⟩. Each word was sense annotated by five annotators. We use the majority choice as correct sense in case of disagreements (Tanaka et al., 2006).

Table 6.1 shows the number of word senses per semantic class. That is, it shows the effect of constraining words senses using the higher semantic classes. For example, of all the polysemic senses (48,180), 56.7% word senses will be completely

Table 6.1: Number of word senses per semantic class (at each level)

# WS per class	ALL Semantic Classes		Lvl 5 (392)		Lvl 4 (136)		Lvl 3 (30)		Lvl 2 (9)	
	#	%	#	%	#	%	#	%	#	%
1	32,167	66.8	27,316	56.7	20,775	43.1	16,944	35.2	10,582	22.0
2	11,606	24.1	14,078	29.2	15,852	32.9	17,106	35.5	18,236	37.8
3	2,769	5.7	3,897	8.1	5,244	10.9	6,084	12.6	7,344	15.2
4	900	1.9	1,264	2.6	2,080	4.3	2,628	5.5	3,680	7.6
≥ 5	738	1.5	1,625	3.4	4,229	8.8	5,418	11.2	8,338	17.3
Total	48,180	100	48,180	100	48,180	100	48,180	100	48,180	100

We use only one class for each word sense even if it's linked to multiple semantic classes.

disambiguated by the superordinate semantic class at level 5. In addition, even if they can't be completely disambiguated, the number of choices is reduced, for example, 29.2 % have only two choices.

6.3 Superordinate Semantic Class Disambiguation

In this section we describe the construction of the superordinate semantic class selection model. In order to investigate which is the best level of superordinate semantic classes to use in Disambiguating word senses, we investigated levels 2 to 5 of **Goi-Taikei**. The word senses are binned into only 9 classes at level 2, 30 classes at level 3, 136 classes at level 4, and 392 classes at level 5.

6.3.1 Mapping Word Sense to Superordinate Semantic Class

Beacuse the semantic classes are in a hierarchy, we can simply generalize them into superordinate classes. For example, in the case of the example sentence of ライター$_1$ *raitâ* "lighter" (Figure 6.1), the semantic classes for たばこ$_2$*tabako* "cigarette" is ⟨862:cigarette⟩, while 火$_1$*hi* "fire" is ⟨2312:burning/combustion⟩ and 点ける$_{17}$*tsukeru* "light" is ⟨2004:operation⟩. At level 3, the superordinate class for

(21)	たばこ$_2$	に	火$_1$	を	点け$_{17}$	た
	cigarette	DAT	fire	ACC	light	TENSE
Sem.	⟨862:cigarette⟩	-	⟨2312:burning/	-	⟨2004:operation⟩	-
Class			combustion⟩	-		
Lvl 5	⟨893:equipment/	-	⟨2306:material	-	⟨1920:labor⟩	-
	tool⟩		phenomenon⟩			
Lvl 4	⟨760:artifact⟩	-	⟨2305:non-living	-	⟨1560:act/conduct⟩	-
			phenomenon⟩			
Lvl 3	⟨706:inanimate⟩	-	⟨2304:natural	-	⟨1236:human act.⟩	-
			phenomenon⟩			
Lvl 2	⟨533:objects⟩	-	⟨1235:events⟩	-	⟨1235:events⟩	-

Where phen. is abbreviation of phenomenon, and act. is activity.

⟨862:cigarette⟩ is ⟨706:inanimate⟩, ⟨2312:burning/combustion⟩ is ⟨2304:natural phenomena⟩ and ⟨2004:operation⟩ is ⟨1236:human activity⟩. So we replace word senses into superordinate semantic classes as shown in (21)[2], which shows the actual semantic class for each content word and the superordinate terms at levels 2 to 5. The more specific sense (the hyponym), is shown lower here, thus ⟨2004:operation⟩ ⊂ ⟨1920:labor⟩ ⊂ ⟨1560:act/conduct⟩ ⊂ ⟨1236:human activity⟩ ⊂ ⟨1235:events⟩.

6.3.2 Problems in Mappings

Because **Lexeed** and **Goi-Taikei** were developed separately, there are some inconsistencies in the hierarchies. Generally, **Lexeed** is more fine-grained, but occasionally a single **Lexeed** sense will be linked to multiple semantic classes in **Goi-Taikei**. In this case we use the first listed class (which should is the most frequent class (Ikehara et al., 1997)).

For example, in **Lexeed**, the simplified definition of word 牛 *ushi* "beef/cow" is "A kind of mammal. It's milk and meat are edible.". The **Lexeed** sense inventory does not distinguish between the animal and its meat or milk. However, **Goi-Taikei** links 牛 *ushi* "beef/cow" to both ⟨537:beast⟩ and ⟨843:meat and eggs⟩. Thus, for example,

[2]We use the following abbreviations: ACC: accsative postposition; DAT: dative postposition; LOC: locative postposition.

in (22), the word 牛 *ushi* “beef/cow” should be tagged with ⟨537:beast⟩ (at level 3, ⟨534:animate⟩). In contrast, in (23), it should be tagged with ⟨843:meat and eggs⟩ (at level 3, ⟨706:inanimate⟩). But in this experiment, both are tagged with the first class, that is ⟨537:beast⟩ (at level 3, ⟨534:animate⟩)[3]. Note that both ⟨537:beast⟩ and ⟨843:meat and eggs⟩ are merged into ⟨533:objects⟩ at level 2.

(22) 農家　　で　牛　を　飼う
farm family LOC *cow* ACC *keep*
“A farm family keeps cows.”

(23) スーパー　　で　牛　を　買う
supermarket LOC *beef* ACC *buy*
“I buy beef in supermarket.”

This is a problem with the granularity of **Lexeed**, which conflates the animal and meat senses of 牛 *ushi* “beef/cow” in a single entry.

Table 6.2 shows the number of semantic classes per word sense at each level. Even at level 5, more than 70% **Lexeed** word senses have only one superordinate semantic class.

Table 6.2: Number of Semantic Classes per word sense

classes	Lvl 2		Lvl 3		Lvl 4		Lvl 5		Class	
/sense	No	(%)	No	(%)	No	(%)	No	(%)	No	(%)
1	39,654	86.0	36,928	80.1	35,075	76.1	32,496	70.5	30,558	66.3
2	6,101	13.2	8,409	18.2	9,858	21.4	11791	25.6	13,102	28.4
3	323	0.7	683	1.5	1,018	2.2	1,517	3.3	1,955	4.2
4	19	0.0	55	0.1	100	0.2	199	0.4	345	0.7
5	4	0.0	18	0.0	21	0.0	41	0.1	62	0.1
≥ 6	0	0.0	8	0.0	29	0.1	57	0.1	79	0.2
Total	46,101		46,101		46,101		46,101		46,101	

[3]To solve this problem, we are annotating **Hinoki** by correct **Goi-Taikei**’s semantic classes.

6.3.3 Data used

We trained and tested on the dictionary definition (Def.) and example (Ex.) sentences and Kyoto Corpus (KC) in the **Hinoki** corpus. In this chapter, we assume that morphological analysis has been done, and we use the results of morphological analysis as inputs.

We divided the data into training and test data. Table 6.3 shows the size of the data sets for training and test. Target words are those open class words tagged with **Lexeed** senses.

Note that several word and word senses appeared in the test data which did not appear in the training data (in the case of Def. 19 words and 389 senses are missing in the training data, for Ex. 1,038 words and 1614 senses, and for KC 137 words and 267 senses). Generalizing to superordinate semantic classes alleviates this data sparseness problem.

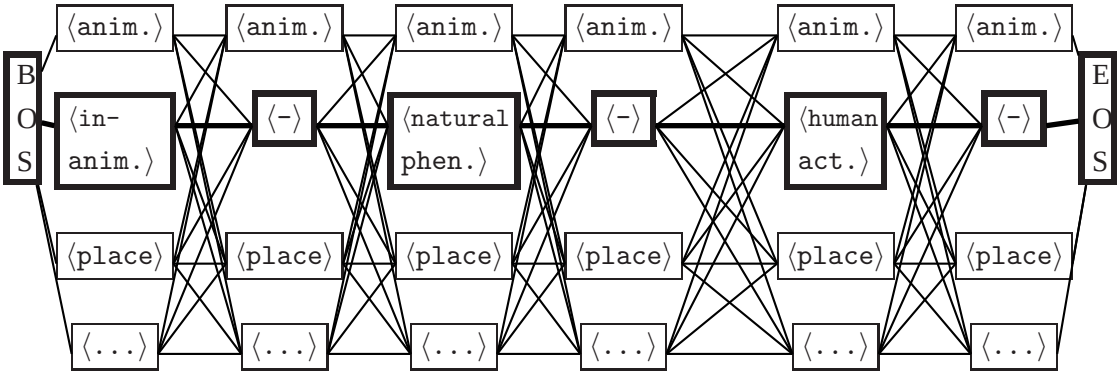
Table 6.3: Data Sets for Superordinate Classes (All Words)

Corpus	Set	# Sents	# Target Words	# All Words
Def.	Train	67,202	175,709	613,216
	Test	4,942	15,436	54,276
Ex.	Train	106,528	133,616	432,514
	Test	8,942	11,043	41,019
KC	Train	35,440	211,567	947,298
	Test	2,000	12,123	53,703

6.3.4 Method

Machine Learning Method We take a sequence labeling approach to make superordinate semantic class disambiguation models, because our goal is to get a wide coverage and robust word sense tagger, not only for a few target words. Ciaramita and Altun (2006) applied Perceptron-trained Hidden Markov Model (HMM) to estimate supersenses of **WordNet**. But we use Conditional Random Fields: CRF (Suzuki et al.,

Lattice (paths of possible superordinate semantic classes at Level 3)



Where `phen.` is abbreviation of phenomenon, `act.` is activity, and `anim.` is animate.
The **bold line** shows the correct path.

Input

w	たばこ	に	火	を	点け	た
b	たばこ	に	火	を	点ける	た
	cigarette	DAT	fire	ACC	light	TENCE
*p*1	noun	particle	noun	particle	verb	aux verb
*p*2	noun-com	particle-cm	noun-com	particle-cm	verb-independent	aux verb-*
*p*3	noun-com-*	particle-cm-com	noun-com-*	particle-cm-com	verb-independent-*	aux verb-*

We use the following abbreviations: cm: casemarker; aux: auxiliary; com: common(general).

Figure 6.2: Simplified Example of Input Information and (Ideal) Lattice of Possible Superordinate Semantic Classes (Level 3)

2006).We select CRF because it allows relaxation of the strong independence assumptions made by HMMs and has performed well for similar sequential labeling problems such as part-of-speech (POS) tagging (Lafferty et al., 2001; Kudo et al., 2004) and named entity recognition (Suzuki et al., 2006).

We can't apply CRF directly to the full WSD problem because the number of classes of senses is too large. But, by restricting ourselves to superordinate semantic classes the number of classes is reduced enough to enable us to train.

Features Now, we describe the features used by CRF. We use uni-gram, bi-gram and combinations of morphological information: that is the word itself (w), base form (b), main category of POS ($p1$), sub-category of POS ($p2$) and sub-sub-category of POS ($p3$). We make features from two words on either side of the target word. That is, in the case of the target is ith word/morpheme, we use the information from the $i-2$th to $i+2$th morphemes.

Hard Dictionary Constraints Figure 6.2 shows a simplified example of input and used information and lattice of possible superordinate semantic classes at Level 3. We select the best path (which has the best score) of superordinate semantic classes. The learner considers all superordinate semantic classes, and thus may predict a class that is not used in the dictionary for this word. For example, at level 3, the only possible classes for たばこ *tabako* "tobacco plant/cigarette" are ⟨`534:animate`⟩ and ⟨`706:inanimate`⟩ according to the entry in **Lexeed**. But the system may guess a different class, such as ⟨`388:place`⟩. To fix such impossible errors, we relabel any words marked with classes not found in the lexicon with the most frequent possible semantic class. This happens between 1–7% of the time, depending on the level.

6.3.5 Results and Discussion

Table 6.4 shows the results of superordinate semantic class disambiguation using CRF. The system actually chooses the semantic classes for all of the words including monosemous words.But in Table 6.4, we show the results for polysemous words: that is the target words shown in Table 6.3.

The baseline (BL) method selects the most frequent semantic class from all possible semantic classes for all senses of the target word. As we can see in Table 6.4, CRF gives much better results than the baseline, especially at deep levels. But CRF needs much time and memory. So, we get some scores (underlined) without $p2$ (the POS subcategry), on average, the scores without $p2$ drop down 0.1-0.2 % from the scores with $p2$.

In Table 6.4, the results labeled with *Hard* shows the results with the hard dictionary constraints applied (see Section 6.3.4). The *Hard* results are better than raw *CRF* for all combinations, and we will use these results in the next section.

Table 6.4: Results of Superordinate Semantic Class Disambiguation

Corpus		Definitions					Examples					Kyoto Corpus				
	Level	noun	verb	adj	misc	total	noun	verb	adj	misc	total	noun	verb	adj	misc	total
BL	2	89.7	96.3	81.7	98.0	91.3	84.5	93.4	87.2	100.0	87.4	89.9	94.4	70.5	80.9	90.3
	3	84.2	85.1	68.5	96.1	83.6	78.5	84.4	74.0	95.2	80.2	84.1	83.4	55.8	79.4	83.3
	4	77.7	83.7	71.3	94.1	79.3	74.9	80.9	72.2	95.2	76.7	80.9	79.9	63.9	77.9	80.3
	5	70.9	70.6	60.0	60.8	70.1	68.7	67.0	57.5	52.4	67.7					
CRF	2	96.0	97.4	88.2	88.2	96.0	85.2	95.8	89.7	85.7	88.7	93.1	94.3	87.7	58.8	93.0
	3	93.9	90.5	80.1	88.2	92.0	81.9	89.1	77.6	81.0	84.0	91.3	86.2	81.8	58.8	89.8
	4	92.5	89.2	78.7	88.2	90.6	79.8	87.2	76.4	81.0	82.0	89.7	84.6	76.5	57.4	88.2
	5	88.7	82.5	74.3	84.3	85.9	76.6	80.6	69.8	81.0	77.6					
Hard	2	96.3	97.5	89.0	96.1	96.3	90.2	95.8	89.9	100.0	92.0	96.6	96.0	89.1	80.9	96.2
	3	94.4	90.9	81.1	94.1	92.5	87.3	89.3	78.1	95.2	87.6	95.1	88.7	83.5	79.4	93.4
	4	93.0	89.8	79.7	96.1	91.2	85.4	87.4	76.9	95.2	85.7	93.6	87.3	78.2	79.4	91.9
	5	89.5	83.5	75.2	88.2	86.7	82.9	80.9	70.3	95.2	81.9					
Targets		9,575	4,895	915	51	15,436	7,189	3,426	407	21	11,043	9,303	2467	285	68	12,123

Where underlined figures were obtained with a simplified model not using *p*2 (sub category of POS) as a feature.

Note that this may not be the desired behaviour for a completely open system: senses may be missing in **Lexeed**, and allowing senses not in the lexicon could be beneficial. However, in this experiment, words can only be tagged with exisiting senses, so we thus restrict them.

6.4 Word Sense Disambiguation (WSD)

In this section, to investigate the effectiveness of superordinate semantic classes for WSD, we show 2 types of data. First, we describe the sense level WSD experiment using the superordinate semantic classes which were extracted in Section 6.3.

Secondly, we show the effects on unseen words: words which appeared in the test data which did not appear in the training data (Section 6.4.2).

6.4.1 Comparison with SENSEVAL-2 Japanese Task

First, we show the effectiveness of superordinate semantic classes on full WSD.

The best published result for the Japanese dictionary-based lexical-sample task of SENSEVAL-2 is given by Murata et al. (2003). We therefore reimplemented their system for comparison. We call this reimplemented system, **CRL**′. Murata et al. (2003) used SVM (Chang and Lin, 2001) as a learner with the following features (See Murata et al. (2003) for more details.): uni/bi/tri-gram characters which precede and follow the target word; Morphological features extracted from the results of morphological analysis; syntactic features from a shallow dependency parser; coocurrence features foremd from all morphemes in the same sentence; and Universal Decimal Classification (UDC) codes.

However, our implementation of **CRL**′ differs from Murata et al. (2003) in two places: we do not use the syntactic features or the UDC codes. The reason that we didn't use syntactic features is that we believe that the results of WSD are useful for syntactic parsing, so we don't use syntactic parsing as pre-processing. We didn't use UDC features because the UDC codes are not tagged in the **Hinoki** Corpus. A further difference is that , they used JUMAN/RWC for morphological analysis, but we used ChaSen.

Our system, **NEW**, also uses SVM, and adds the superordinate semantic classes

(which were extracted in Section 6.3) as features to **CRL**′. These are added on the word itself, and the two words on either side. We experiment with superordinate semantic classes generalized to different upper levels.

For example, in Figure 6.2, if we guess ⟨1236:human activity⟩ as the superordinate semantic class corresponding to 5th word 付け*tsuke* "light" at level 3, we use both ⟨1236:human activity⟩ and ⟨1235:events⟩ as features.

Data for WSD

For the fine grained word sense disambiguation experiment, we use the same target words as SENSEVAL-2, in order to give a more meaningful comparison. There are 100 target words: 50 nouns and 50 verbs. The test documents were the same as in SENSEVAL-2, with all text coming from newspaper articles (these are not part of the training data). Table 6.5 shows the amount of training and test data used in this experiment.

Results and Discussion

Table 6.6 shows the results of the full WSD. The baseline (BL) method selects the word sense occurring most frequently in the training corpus. The higher baseline system (BL2) uses the most frequent sense restricted by the disambiguated superordinate semantic classes.

NEW also uses the disambiguated superordinate semantic classes. All results are significantly better than the baseline (BL). And most results of **NEW** are better than **CRL**′, even at upper levels.

In addition, Table 6.6 shows that even if we just use the most frequent sense restricted by the disambiguated superordinate semantic classes (BL2), we can get high accuracy. In general, the more specific the superordinate classes, the higher the accuracy, even though the accuracy for disambiguating the more specific classes is lower.

The improvement is smallest for the Kyoto Corpus data. We hypothesize that this is because it has more unknown senses — none of the words not in **Lexeed**'s fundamental vocabulary are tagged. In particular, proper nouns are not tagged. This means 25% of the words (mainly noun phrases) have no sense information. In contrast, all the words in the Example and Definition sub-corpora are in the fundamental vocabulary. We are

Table 6.5: Data Sets for WSD (Senseval 100 words)

Corpus	Set	noun	verb
Def.	Trains	6512	11409
	Test	745	1151
Ex.	Trains	4448	7888
	Test	317	826
KC	Trains	11140	10744
	Test	763	610

Table 6.6: Results of WSD (by SVM)

Corpus		Definitions			Examples			Kyoto Corpus		
Level		noun	verb	ave	noun	verb	ave	noun	verb	ave
BL		74.5	56.8	63.8	63.7	56.2	58.3	69.2	62.1	66.1
CRL′		81.1	65.3	71.5	79.5	68.5	71.6	80.9	67.0	74.7
BL2	2	76.8	59.9	66.5	66.9	58.8	61.0	69.9	63.4	67.0
	3	80.8	60.6	68.5	69.1	60.5	62.8	75.0	65.4	70.7
	4	80.9	61.6	69.2	71.0	61.3	64.0	76.7	**68.0**	72.8
	5	**83.4**	**67.4**	**73.7**	76.3	65.2	68.3			
NEW	2	81.3	65.6	71.8	79.5	68.3	71.4	81.3	67.0	74.9
	3	81.5	66.1	72.2	79.5	68.5	71.6	**81.5**	67.0	**75.1**
	4	81.6	66.3	72.3	79.5	68.8	71.7	81.3	67.0	74.9
	5	81.7	67.2	72.9	**80.1**	**69.2**	**72.3**			

Table 6.7: Accuracy for words which didn't appear in training data (Zero Frequency)

Corpus		Definitions					Examples					Kyoto Corpus				
	Level	noun	verb	adj	misc	total	noun	verb	adj	misc	total	noun	verb	adj	misc	total
first sense		27.8	0.0	0.0	0.0	26.3	27.4	0.0	0.0	0.0	27.4	36.5	0.0	0.0	23.1	29.9
NEW′	2	55.6	0.0	0.0	0.0	52.6	46.5	0.0	0.0	0.0	46.3	43.3	20.0	40.0	30.8	39.4
	3	**61.1**	0.0	0.0	0.0	**57.9**	**48.8**	0.0	0.0	0.0	**48.7**	**53.8**	33.3	20.0	46.2	**49.6**
	4	55.6	0.0	0.0	0.0	52.6	47.8	0.0	0.0	0.0	47.7	40.4	**40.0**	**60.0**	**53.8**	42.3
	5	50.0	0.0	0.0	**100.0**	52.6	47.3	0.0	0.0	0.0	47.2					
# Targets		18	0	0	1	19	1,035	3	0	0	1,038	104	15	5	13	137

currently the remaining words in the kyoto Corpus with **Goi-Taikei** semantic classes to give us the data to test this hypothesis.

6.4.2 Effect on Unseen Words

Because of the huge numbers of words and senses, it is very difficult to get enough training data. Sometimes, we can get no training data at all for some words. In such case, most supervised WSD methods (including **CRL′**) doesn't work. But in our method, at least, we can guess superordinate semantic classes for words which didn't appear in training data.

Table 6.7 shows the accuracy for such words (frequency is 0). In this case, we compare the first sense (in **Lexeed**) baseline with the first sense restricted by the disambiguated superordinate semantic classes (**NEW′**). The accuracy of **NEW′** is much better than first sense baseline. Disambiguating superordinate semantic classes gives a much more robust WSD system. It's interesting to note that the superordinate semantic classes at Level 3 give the best results overall, with an accuracy of 49%, compared to the baseline of 27%.

6.5 Discussion

We showed that disambiguating superordinate semantic classes is an effective way of WSD, even though we use superordinate classes from a different resource. This is

important, as the sense inventories used in a task are not always in a full hierarchy (e.g., the Japanese SENSEVAL-2 task gave word senses from a dictionary with no associated hierarchy). We expect we could get even better results using a hierarchy built around the **Lexeed** word senses.

Further, we have shown that a quite large superordinate class inventory (level 5 with 393 results) gave the best results on several test sets. This suggests that work on English using the unique beginners could possibly be improved by specializing even further in the initial step.

It could be that the **Lexeed** semantic classes are too fine-grained for reliable sense disambiguation. Navigli (2006) shows that this is true for the English **WordNet**—clustering senses allows for more reliable manual and automatic annotation. Bond et al. (2004) argue that, in comparison to **Goi-Taikei**, the finer granularity of **Lexeed** is necessary for question answering, but it may still be the case that not all of the sense distinctions are meaningful. Fujita et al. (2007) use gold-standard sense information to improve parse selection, and found that the superordinate senses at level two were the most effective to reduce data sparseness for parse selection.

6.6 Future Work

In future work we intend to confirm that we can get improved results in other languages such as English using various levels of superordinate senses in **WordNet** (Agirre et al., 2008).

Then we intend to make a superordinate semantic class tagger using CRF like MeCab (Kudo et al., 2004). That is, in this chapter, for the experiment, we used the packaged CRF based machine learner, but if we save the possible pairs of entries and superordinate semantic classes into a dictionary, we will not have to fix impossible errors (That is we can get *Hard* data from the beginning). We hope that this will improve the accracy even further. Alternatively, for morphological analysis, we may get part-of-speech tag and sense tag together. In addition, we would like to further experiment with limiting the number of states, more features and guessing superordinate semantic classses at even deeper levels.

Finally, to get more training data for superordinate semantic class disambiguation, we intend to use untagged corpus, by applying the method proposed by Tsuboi et al.

(2008).

6.7 Conclusion

In this chapter, we proposed the method for word sense disambiguation (WSD) using superordinate semantic classes. At the first stage, we guess superordinate semantic classes, then at the second stage we guess word senses using the results of 1st stage.

At the first stage, we applied CRF to superordinate semantic class disambiguation. As a result, it gave us very high accuracy. At the second stage, we got higher accuracy for WSD than published best method of Japanese dictionary-based lexical-sample task of **SENSEVAL-2**. In addition, we showed the effectiveness of superordinate semantic classes for unseen words.

In conclusion, our proposed WSD method using superordinate semantic classes is very effective.

Chapter 7

Conclusion

7.1 Summary

In this thesis, we introduced some resources: **Goi-Taikei**, its bilingual valency pattern dictionary, **Hinoki**, and **Lexeed**, which have rich information and then are related to each other. First, we construct these resources by hand as shown in Chapter 2. Then, we proposed a method to extend them effectively, and proved the usefulness through several task-based evaluations.

In Chapter 3, we presented a method of extending the coverage of the bilingual valency dictionary, by assigning valency information and selectional restrictions to entries in a bilingual dictionary. The method exploits existing bilingual valency dictionaries and is based on two basic assumptions: words with similar meaning have similar subcategorization frames and selectional restrictions; and words with the same translations have similar meanings. A prototype system allowed 6,327 new patterns to be built, using only simple human judgement (pre-filter). Of those more than 51% were usable as is, and more than 36% were usable with minor revisions, giving 87.7% potentially useful patterns. The cost, including human revisions, is less than 6 minutes per pattern. Furthermore, even before applying human revisions, adding the created patterns to a Japanese-to-English machine translation system improved the translation for 32% of sentences using these verbs, and degraded it for only 16%; a substantial improvement in quality.

In Chapter 4, we presented a method that uses alternation data to add new entries to an existing bilingual valency dictionary. The new entries have detailed information

about argument structure and selectional restrictions. If the existing lexicon has only one half of the alternation, then our method constructs new Japanese entries with 69% of the verbs having one or more correct entries. We also showed that it is possible to simultaneously add entries to a second language with a reduced accuracy of 56% if your exisiting lexicon has such information. In this section we focused on one class of alternations, but it is applicable to any alternation.

In Chapter 5, we showed that sense-based semantic features combined with ontological information are effective for parse selection. Training and testing on the definition subset of the **Hinoki** corpus, a combined model gave a 5.6% improvement in parse selection accuracy over a model using only syntactic features (63.8% $\rightarrow$ 69.4%). Similar results (76.2% $\rightarrow$ 79.2%) were found with example sentences.

In Chapter 6, to get sense information automatically, we proposed a method for word sense disambiguation (WSD) using superordinate semantic classes. We separated this method into two stages. In the first stage, we estimate superordinate semantic classes. We did this using CRFs, and were able to disambiguate with a very high accuracy.

In the second stage we estimate word senses using the results of the first stage. We got higher accuracy for WSD than published best method of Japanese dictionary-based lexical-sample task of **SENSEVAL-2**. In addition, we showed the effectiveness of superordinate semantic classes for unseen words.

As shown above, though the most recent research direction is on statistical methods, rich resources (dictionary, ontology, treebank, sensebank, etc.) are effective for deeper natural language processing. We showed the effectiveness through task-based evaluations, machine translation, parse selection and word sense disambiguation.

7.2 Future Work

There are several directions for future research.

Construction of Rich Information Resources Construction of Rich Information Resources In this thesis, we introduced several manual or semi-automatic methods of constructing rich information resources. Especially for the bilingual valency dictionary, we investigated effective methods to expand it.

In future work, we want to expand the resources in several ways. For **Lexeed**, we want to import entries or senses from other machine readable dictionaries or online resources such as Wikipedia[1] and Wiktionary[2]. From these we can extract at least a lemma and it's definition. Then, by using link information of online dictionaries, we can extract other information such as examples, frequency, and access frequency.

For **Hinoki**, we plan to expand the target domain into open domains such as Blog, e-mail. We have already begun to expand the sensebank over blog data. We illustrate the rough plan of expanding the target domain in Figure 7.1.

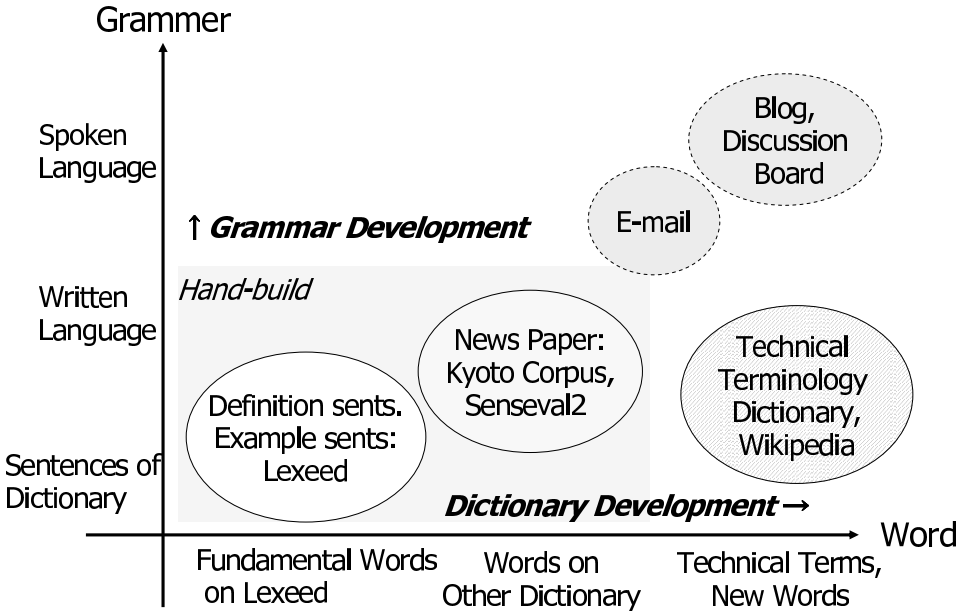


Figure 7.1: Plan to Expand Resources: from closed world to open domain, from hand-build to semi-automatic

For **Goi-Taikei**, we want to reliably add words to semantic classes. One way is to use parse results of definitions in other dictionaries like Bond et al. (2004). Now, we are extracting unknown words from Wikipedia, then trying to estimate the semantic

[1]http://en.wikipedia.org/wiki/Main_Page

[2]http://ja.wiktionary.org/wiki/

classes through a rule based method. On the other hand, NTT is planing to make **Goi-Taikei** open. We can then expect to improve **Goi-Taikei**'s coverage and keep it up-to-date by getting users feedback.

Usage of the Resources As shown in Chapter 5 and 6, we are combining symbolic and statistical approaches to parse selection and word sense disambiguation. In future work, we want to go ahead with combining approaches to more various natural language processing tasks; especially for machine translation. In this thesis, we used the rule-based machine translation system **ALT-J/E**, but great progress has been made in learning statistical models from annotated corpora, and some (online) statistical machine translation systems[3] are now available. However, statistical machine translation is not strong for out of domain data. According to Koehn and Monz (2006), for in-domain data, statistical approaches are stronger, but for out-of-domain data, rule-based system Systran[4] becomes stronger. It shows that because dictionaries and translation rules are relatively domain independent, they help to make systems robust. So in future work, we want to export the bilingual valency dictionary into a statistical machine translation system: we need to investigate the best way to export the dictionary.

We also intend to make a superordinate semantic class tagger. If we can provide a packaged **Goi-Taikei** semantic class tagger, it will help the open **Goi-Taikei** to gain wide acceptance. Finally, we intend to confirm that we can get improved results with raw sense disambiguation results not just the gold standard annotations.

7.3 Conclusion

In this thesis, we first introduced various rich information resources which we have used or constructed: **Goi-Taikei**, its bilingual valency (pattern) dictionary, **Hinoki**, and **Lexeed**. We also compared the these resources with other similar resources. Because they were mainly built by hand, constructing such rich resources was both time consuming andcostly. To extend such rich resources efficiently, we proposed some methods to extend them using the hand-made rich resources as seeds.

[3]http://www.google.com/language_tools

[4]http://www.systransoft.com

First, we proposed various methods to extend the valency dictionary: using simpler bilingual dictionaries and linguists analyses of alternations. This not only extended but also added more information into the bilingual valency dictionary. The evaluation of the extended resource's quality was done with both a translation task-based evaluation and a direct evaluation by lexicographers. Through these evaluations, we showed the effectiveness of our methods.

We then investigated the usage of rich information by applying it to parse selection (ranking), and to word sense disambiguation. Through these experiments, we showed the importance and usefulness of semantic information in statistical approaches to natural language processing tasks.

Appendix A

Data on the distribution of Goi-Taikei's Semantic Classes

We show some data about **Goi-Taikei**'s classes in Tables A.1, A.2, A.3, A.4 and A.5.

Table A.1 shows that the most frequent 30 classes in **Goi-Taikei**'s Japanese Word Dictionary. As shown in Table A.1, **Goi-Taikei** has a lot of entries about names of places or humans.

Table A.2 shows that the distribution of semantic classes over newspaper text. In the **Hinoki** project, all words of the first half of Kyoto Corpus (newspaper text, including 19,013 sentences, 522,884 words, 298,974 contents words) are tagged with **Goi-Taikei**'s semantic classes (for more details of **Hinoki**, see Section 2.4.2). In the case of this corpus, we tagged all contents words (not only common noun but also proper nouns, verbs, adjectives and adverbs) using **Goi-Taikei**'s common noun ontology. But in Table A.2, we showed the nouns only.

Table A.1: Most Frequent 30 Semantic Classes in Japanese Dictionary

Class	Lvl	Token	Sample Word
⟨464:jurisdiction⟩	4	93,141	都市 *toshi* “city”
⟨48:male/man⟩	7	38,798	男性 *dansei* “male”
⟨5:human⟩	4	29,654	人 *hito* “person”
⟨471:land⟩	6	17,855	土地 *tochi* “land”
⟨49:female/woman⟩	7	11,846	女性 *josei* “female”
⟨459:zone/area/district⟩	4	8,411	地域 *chiiki* “region”
⟨364:executive agency/ administrative body⟩	5	7,567	政府 *seifu* “government”
⟨414:station⟩	6	5,876	ホーム *hômu* “home”
⟨495:rivers and streams⟩	7	5,094	水系 *suikei* “water system”
⟨499:wetlands⟩	7	5,091	湖沼 *koshou* “lake”
⟨1035:method⟩	6	4,454	対策 *taisaku* “action”
⟨413:platform/loading platform⟩	5	3,854	ターミナル *tâminaru* “terminal”
⟨973:electrical component⟩	7	3,546	抵抗 *teikou* “resistance”
⟨465:city⟩	4	3,532	住宅 *juutaku* “houce”
⟨2498:structure⟩	5	3,338	制度 *seido* “”
⟨374:enterprise/ corporation/industry⟩	6	2,669	企業 *kigyo* “company”
⟨1020:logic⟩	6	2,248	手続き *tetuzuki* “procedure”
⟨2595:unit⟩	4	2,084	一部 *ichibu* “part”
⟨428:work place⟩	4	2,059	会社 *kaisha* “company”
⟨712:matter/material (bodies)⟩	5	1,805	物質 *bushitu* “matter”
⟨2586:number⟩	4	1,787	一つ *hitotu* “one”
⟨2596:calculated value⟩	4	1,753	金利 *kinri* “interest rate”
⟨2435:pattern, method⟩	5	1,705	体制 *taisei* “system”
⟨2591:weights and measures⟩	5	1,656	最大 *saidai* “maximum”
⟨971:computer⟩	7	1,613	パソコン *pasokon* “personal computer”
⟨367:public institution⟩	5	1,550	病院 *byouin* “hospital”
⟨1008:knowledge, intelligence⟩	6	1,529	情報 *jouhou* “information”
⟨962:machinery⟩	5	1,394	システム *shisutemu* “system”
⟨2592:degree/extent/measure⟩	5	1,359	高さ *takasa* “hight”
⟨507:sea/ocean⟩	6	1,219	海 *umi* “sea”

Table A.2: Top 30 Semantic Classes over Newspaper Text (first half of Kyoto Corpus): Noun Only

Class	Lvl	Token	Sample Word
⟨2586:number⟩	4	13,805	二十八 *28* “28”
⟨2595:unit⟩	4	8,201	年 *nen* “year”
⟨47:men and women/gender⟩	6	7,259	村山 *Murayama* “family name”
⟨464:jurisdiction⟩	4	6,064	高知 *Kouch* “Kouchi Prefecture”
⟨385:nation⟩	4	5,959	ロシア *roshia* “Russia”
⟨2682:day⟩	6	3,018	日 *nichi* “sun”
⟨1022:circumstance/thing/ matter/affair⟩	6	3,006	こと *koto* “thing”
⟨48:male/man⟩	7	2,757	富市 *Tomiichi* “name”
⟨374:enterprise/ corporation/industry⟩	6	2,339	会社 *kaisha* “company”
⟨459:zone/area/district⟩	4	2,153	南部 *nanbu* “south”
⟨380:political party⟩	6	2,093	社会党 *shakaitou* “Socialist Party”
⟨260:politician⟩	7	2,079	首相 *shusho* “Prime Minister”
⟨2535:aspect/condition/phase (other)⟩	5	1,755	可能 *kanou* “possibility”
⟨43:honorific title/term of respect⟩	6	1,753	氏 *shi* “Mr.”
⟨364:executive agency/ administrative body⟩	5	1,738	内閣 *naikaku* “cabinet”
⟨1680:sport⟩	6	1,715	サッカー *sakkâ* “soccer”
⟨2679:year⟩	6	1,494	今年 *kotoshi* “this year”
⟨2680:month⟩	6	1,443	二月 *nigatu* “February”
⟨2509:circumstance/situation⟩	5	1,205	よう *you* “like”
⟨2600:part⟩	5	1,122	部 *bu* “division”
⟨363:establishment/institution⟩	4	1,070	議会 *gikai* “assembly”
⟨465:city⟩	4	1,033	首都 *shuto* “capital”
⟨2508:aspect/condition/phase⟩	4	1,028	的 *teki* “target”
⟨2456:purpose⟩	5	1,014	方針 *houshin* “policy”
⟨2692:the time⟩	6	904	午前 *gozen* “morning”
⟨2695:period (natural and human activity, etc.⟩	6	895	時期 *jiki* “season”
⟨378:society⟩	6	894	会 *kai* “meeting”
⟨323:chief/president/manager⟩	6	893	議長 *gichou* “chairperson”
⟨2608:extent/degree⟩	5	874	重大 *jyûdai* “important”
⟨2623:interior⟩	6	841	内部 *naibu* “inside”
Total (Noun)	-	213,276	

Tables A.3, A.4 shows that the distribution of semantic classes merged into superordinate semantic classes at level 2 and 3. In Tables A.3, A.4, we don't restrict by it's POS: that is they are including all contents words. Tables A.3, A.4 shows that the semantic classes are heterogeneously-distributed. From Table A.3, ⟨1235:event⟩ and ⟨2422:abstract relationship⟩ appear at a high rate. And Table A.4 shows that the majority of children of ⟨1235:event⟩ is ⟨1236:human activity⟩.

Table A.3: Distribution of Semantic Classes in Newspaper Text (The first half of Kyoto Corpus): Merged into Superordinate Semantic Classes at Level 2

Class	Lvl	Token	(%)	Sample Word
⟨1:common noun⟩	0	46	0	あれこれ *arekore* "this and that"
⟨2:concrete⟩	1	81	0	万物 *banbutsu* "all things"
⟨3:agent⟩	2	49,248	16.5	私 *watashi* "I"
⟨388:place⟩	2	16,058	5.4	本陣 *honjin* "headquarters"
⟨533:object⟩	2	10,581	3.5	ゴンドラ *gondora* "gondola"
⟨1000:abstract⟩	1	4	0	もの *mono* "thing"
⟨1001:abstract thing⟩	2	19,382	6.5	条例 *jourei* "regulation"
⟨1235:event⟩	2	96,551	32.3	任官 *ninkan* "appointment"
⟨2422:abstract relationship⟩	2	107,023	35.8	背後 *haigo* "back"
Total		298,974	100	

Table A.4: Distribution of Semantic Classes in Newspaper Text (The first half of Kyoto Corpus): Marged into Superordinate Semantic Classes at Level 3

Class	Lvl	Token	(%)	Sample Word
⟨1:common noun⟩	0	46	0	あれこれ *arekore* “this and that”
⟨2:concrete⟩	1	81	0	万物 *banbutsu* “all things”
⟨3:agent⟩	2	318	0.1	主体 *shutai* “subject”,
⟨4:person⟩	3	29,785	10	私 *watashi* “I”
⟨362:organizations⟩	3	19,145	6.4	チェコ *cheko* “Czech”
⟨388:place⟩	2	163	0.1	ところ *tokoro* “place”
⟨389:facility⟩	3	3,918	1.3	本陣 *honjin* “headquarters”
⟨458:region⟩	3	10,574	3.5	南部 *nanbu* “south”
⟨468:natural place⟩	3	1,403	0.5	ビーチ *bîchi* “beach”
⟨533:object⟩	2	1	0	物 *mono* “thing”
⟨534:animate⟩	3	2,148	0.7	金魚 *kingyo* “goldfish”
⟨706:inanimate⟩	3	8,432	2.8	ボイラー *boirâ* “boiler”
⟨1000:abstract⟩	1	4	0	もの *mono* “thing”
⟨1002:mental thing⟩	3	12,534	4.2	知 *chi* “wisdom”
⟨1154:abstract thing (behavior)⟩	3	6,848	2.3	条例 *jourei* “regulation”
⟨1235:event⟩	2	286	0.1	七不思議 *7fushigi* “seven wonders”
⟨1236:human activity⟩	3	68,067	22.8	消し止める *keshi-tomeru* “put out”
⟨2054:phenomena⟩	3	24,757	8.3	改まる *aratamaru* “be renewed”
⟨2304:natural phenomena⟩	3	3,441	1.2	腐る *kusaru* “go bad”
⟨2422:abstract relationship⟩	2	5	0	ずれ *zure* “difference”
⟨2423:existence/being⟩	3	3,105	1	留保 *ryûho* “reservation”
⟨2432:kind OR system⟩	3	2,433	0.8	上位 *joui* “higher rank”
⟨2443:connected to/related to⟩	3	9,107	3	交互 *kougo* “alternation”
⟨2483:nature/disposition⟩	3	4,454	1.5	深い *fukai* “profound”
⟨2507:state⟩	3	31,134	10.4	真空 *shinkû* “vacuum”
⟨2564:shape⟩	3	397	0.1	鋭角 *eikaku* “acute angle”
⟨2585:amount⟩	3	32,459	10.9	あれだけ *aredake* “that much”
⟨2610:location⟩	3	4,949	1.7	背後 *haigo* “back”
⟨2670:time⟩	3	18,980	6.3	週間 *shûkan* “week”
Total		298,974	100	

Of **Goi-Taikei**'s classes, 547 classes don't appear in the newspaper text (The first half of Kyoto Corpus). We show some samples of semantic classes which don't appear in that newspaper text in Table A.5. Table A.5 lists the top 10 classes in order of the number of tokens in **Goi-Taikei**'s Japanese dictionary. Note that some classes of these classes have children which appear in the Corpus. Forexample, ⟨963:general machinery⟩ doesn't appear in the target text, but its children ⟨964:motor⟩, ⟨965:implement⟩, ⟨966:communicator⟩ and ⟨967:machine part⟩ appear 23 times collectively. Therefore, we marked Table A.5 whose children appear or doesn't appear. In the column Children Appear, *Yes* means that children of the class appear, and *No* means that children of the class doesn't appear.

Table A.5: Samples of Semantic Classes which don't appear in Newspaper Text (The first half of Kyoto Corpus): Top 30 classes in **Goi-Taikei**'s Japanese Dictionary

Class	Lvl	Token (Dict.)	Sample Word	Children Appear
⟨504:springs and wells⟩	7	437	温泉 *onsen* "hot spring"	Yes
⟨1082:sentence⟩	7	134	文 *bun* "sentence"	No
⟨726:charcoal⟩	8	104	石炭 *sekitan* "coal"	No
⟨791:mineral oil/petroleum⟩	7	95	石油 *sekiyu* "oil"	No
⟨165:servant/ retainer/employee⟩	8	78	臣 *shin* "vassal"	No
⟨213:good person/ virtuous person⟩	8	62	信者 *shinja* "believer"	No
⟨978:optical component⟩	7	62	セクター *sekutâ* "sector"	Yes
⟨1529:explanatory notes⟩	9	52	注釈 *chuushaku* "note"	No
⟨966:communicator⟩	7	51	ベルト *beruto* "belt"	No
⟨200:lazy person⟩	8	49	寄生虫 *kiseichu* "parasite"	No
⟨1039:poetry⟩	6	49	句 *ku* "phrase"	Yes
⟨963:general machinery⟩	6	48	機械 *kikai* "machine"	Yes
⟨1243:madness⟩	6	47	狂気 *kyouki* "madness"	No
⟨704:bark, peel/rind/skin⟩	6	45	樹脂 *jushi* "resin"	No
⟨315:prostitute⟩	7	45	売春婦 *baishunfu* "prostitute"	No
⟨900:oven⟩	8	45	窯 *kama* "oven"	No
⟨164:feudal lord⟩	8	45	ロード *lodo* "road"	No
⟨628:mole, wart⟩	8	44	たこ *tako* "lump"	No
⟨206:flirt/a lustful and promiscuous person⟩	10	44	サド *sado* "sadism"	No
⟨314:gangster⟩	7	43	不良 *furyou* "inferiority"	No
⟨196:coward/weakling⟩	8	43	弱者 *jakusha* "weak"	No
⟨290:shipping agent/carrier⟩	8	42	強力 *kyouryoku* "great strength"	No
⟨775:stone⟩	7	41	ブロック *burokku* "block"	No
⟨1104:figure, table, score⟩	7	40	図表 *zuhyo* "chart"	Yes
⟨118:companion⟩	10	39	仲間 *nakama* "friend"	No
⟨1090:Cn- and Jn-style readings of Cn characters⟩	7	39	訓 *kun* "Jn reading of a Cn character"	No
⟨1091:grapheme (linguistic)⟩	6	38	英字名 *eijina* "symbolic name"	Yes
⟨2618:border⟩	4	37	きれめ *kireme* "slit"	Yes
⟨333:the roles of people⟩	5	36	関係者 *kankeisha* "person concerned"	Yes
⟨699:flower⟩	7	34	花粉 *kafun* "pollen"	No
⟨623:membrane⟩	7	34	網膜 *moumaku* "retina"	No

Appendix B

Classification of English Alternations for Japanese S = O Alternation

Vi			Vt		
Japanese		English	Japanese		English
Type: **S = O**					
開く	aku	open, be open	開ける	akeru	open
空く	aku	open, become empty	空ける	akeru	open, empty
明く	aku	open, be open	明ける	akeru	open, dawn
当たる	ataru	touch, hit, be hit	当てる	ateru	hit
当て嵌まる	atehamaru	apply (a rule), be applicable	当て嵌める	atehameru	apply
浴びる	abiru	pour(over oneself), bathe	浴びせる	abiseru	pour(over another), bathe, pour on
荒立つ	aradatsu	be aggravated, be rough or aggravated or worse	荒立てる	aradateru	aggravate
癒える	ieru	heal	癒す	iyasu	heal
痛む	itamu	hurt, be hurt	痛める	itameru	injure, hurt
燻る	iburu	smoke	燻す	ibusu	fumigate, smoke
卑しむ	iyashimu	despise	卑しめる	iyashimeru	despise
浮かぶ	ukabu	float, float up	浮かべる	ukaberu	set afloat, float up, float
浮く	uku	float	浮かす	ukasu	float

Vi		Vt	
Japanese	English	Japanese	English
Type: **S = O**			
動く ugoku	move	動かす ugokasu	move
うだる udaru	boil	茹でる uderu	boil
移る utsuru	move, move (a to b)	移す utsusu	move, move (a to b), remove
写る utsuru	project, be photographed	写す utsusu	project, film
裏返る uragaeru	be turn inside out	裏返す uragaesu	turn inside out
売れる ureru	sell (well), sell, be sold	売る uru	sell
起きる okiru	wake up, get up	起こす okosu	wake up, raise
溺れる oboreru	drown	溺らす oborasu	drown
折れる oreru	break	折る oru	break
終わる owaru	end, finish	終える oeru	end, finish
帰る kaeru	return (home), go back	帰す kaesu	return (home), send back
帰る kaeru	return	返す kaesu	return
返る kaeru	return	返す kaesu	return, return something
掛かる kakaru	hang down, cost, take (e.g. time, money, etc), easel	掛ける kakeru	hang, spend, wear
屈む kagamu	bend, stoop	屈める kagameru	bend, stoop
隠れる kakureru	hide	隠す kakusu	hide
欠ける kakeru	lack, be lacking	欠く kaku	lack
重なる kasanaru	pile up, piled up, be piled up	重ねる kasaneru	pile up
傾げる katageru	lean	傾ぐ katagu	lean
固まる katamaru	harden	固める katameru	harden
傾く katamuku	lean, incline toward	傾ける katamukeru	lean, incline
角立つ kadodatsu	be sharp	角立てる kadodateru	be sharp
涸れる kareru	dry up	涸らす karasu	dry up
乾く kawaku	dry, get dry	乾かす kawakasu	dry, dry (clothes, etc.)

Vi			Vt		
Japanese		English	Japanese		English
Type: **S** = **O**					
換わる	kawaru	change, take the place of	換える	kaeru	change, exchange
替わる	kawaru	change, take the place of	替える	kaeru	change, exchange
代わる	kawaru	take the place of, change	代える	kaeru	change, exchange
変わる	kawaru	change	変える	kaeru	change
聞き違う	kikichigau	mishear	聞き違える	kikichigaeru	mishear
切り替わる	kirikawaru	change completely	切り替える	kirikaeru	change
腐る	kusaru	spoil, rot	腐らす	kusarasu	spoil
腐れる	kusareru	spoil	腐らす	kusarasu	spoil
砕ける	kudakeru	be smashed, break	砕く	kudaku	smash, break
覆る	kutsugaeru	capsize, topple over	覆す	kutsugaesu	capsize, overturn
繰り上がる	kuriagaru	move up (date or rank)	繰り上げる	kuriageru	move up
凍る	kooru	freeze	凍らす	koorasu	freeze
焦げる	kogeru	be scorched, burn	焦がす	kogasu	scorch, burn
零れる	koboreru	spill, be spilt, overflow	零す	kobosu	spill
凝る	koru	be devoted to, stiffen, grow stiff	凝らす	korasu	devote to, stiffen, concentrate
転がる	korogaru	roll, roll over	転がす	korogasu	roll, roll over
転げる	korogeru	roll, roll over	転がす	korogasu	roll
壊れる	kowareru	break, be broken	壊す	kowasu	break
下がる	sagaru	drop, hang down	下げる	sageru	lower, hang down, hang
裂ける	sakeru	tear, be torn, split	裂く	saku	tear
覚める	sameru	awake, wake	覚ます	samasu	arouse, wake, awaken
冷める	sameru	get cool, cool, become cool	冷ます	samasu	cool
復習う	sarau	review	復習える	saraeru	review
仕上がる	shiagaru	be finished	仕上げる	shiageru	finish up
沈む	shizumu	sink	沈める	shizumeru	sink
死に掛かる	shinikakaru	be dying	死に掛ける	shinikakeru	be dying
閉まる	shimaru	be closed, close	閉める	shimeru	close
湿る	shimeru	get wet, be wet	湿す	shimesu	wet

Vi			Vt		
Japanese		English	Japanese		English
Type: **S = O**					
過ぎる	sugiru	pass	過ごす	sugosu	make pass, pass
透く	suku	be transparent	透ける	sukeru	be transparent
透ける	sukeru	be transparent	透く	suku	be transparent
済む	sumu	end, be settled, finish	済ます	sumasu	end, settle, finish
澄む	sumu	be clear, clear, clear (e.g. weather)	澄ます	sumasu	make clear, clarify, clear
擦れる	sureru	rub	擦る	suru	rub
摩れる	sureru	rub	擦る	suru	rub
狭まる	sebamaru	get narrow, narrow	狭める	sebameru	narrow
備わる	sonawaru	be furnished with	備える	sonaeru	be furnished with
染まる	somaru	be dyed, dye	染める	someru	dye
背く	somuku	turn away, run counter to	背ける	somukeru	turn away, turn one's face away
反る	soru	bend, be warped, warp	反らす	sorasu	bend, warp
立つ	tatsu	stand	立てる	tateru	raise, stand, stand up
貯まる	tamaru	collect	貯める	tameru	collect
垂れ下がる	taresagaru	hang	垂れ下げる	taresageru	hang (a curtain)
垂れる	tareru	drop, hang	垂らす	tarasu	drop, suspend
縮む	chidimu	shrink	縮める	chidimeru	reduce, shrink, shorten
縮れる	chidireru	curl, be wavy	縮らす	chidirasu	curl
散る	chiru	scatter, be scattered, fall	散らす	chirasu	scatter
疲れる	tsukareru	tire, get tired	疲らす	tsukarasu	tire
続く	tsuduku	continue, be continued	続ける	tsudukeru	continue
窄まる	tsubomaru	get narrow, close	窄める	tsubomeru	narrow, close
窄む	tsubomu	get narrower, close	窄める	tsubomeru	narrow, close
溶ける	tokeru	melt, dissolve	溶かす	tokasu	melt, dissolve
止まる	tomaru	stop, come to a halt	止める	tomeru	stop
灯る	tomoru	burn	灯す	tomosu	burn, light
慰む	nagusamu	cheer up, comfort	慰める	nagusameru	cheer, comfort
並ぶ	narabu	line up, form a line	並べる	naraberu	line up

Vi			Vt		
Japanese		English	Japanese		English
Type: **S = O**					
鳴る	naru	ring, sound	鳴らす	narasu	ring, sound
煮える	nieru	boil, be boiled	煮る	niru	boil
煮立つ	nitatsu	boil or simmer	煮立てる	nitateru	boil or simmer
伸びる	nobiru	extend, be stretched, stretch	伸ばす	nobasu	extend, stretch, lengthen
生える	haeru	grow	生やす	hayasu	grow
剥がれる	hagareru	peel off, come unstuck from	剥がす	hagasu	peel off, tear off
始まる	hajimaru	start, begin	始める	hajimeru	start, begin
嵌まる	hamaru	fit	嵌める	hameru	fit
早まる	hayamaru	hasten, be hasty	早める	hayameru	hasten
腫れる	hareru	swell	腫らす	harasu	swell
晴れる	hareru	clear up, clear away, be sunny	晴らす	harasu	clear up, clear away, dispel
冷える	hieru	cool, cool down, grow cold	冷やす	hiyasu	cool, cool down
低まる	hikumaru	become lower, lower	低める	hikumeru	lower
開ける	hirakeru	open	開く	hiraku	open
翻る	hirugaeru	wave, turn over	翻す	hirugaesu	wave, change
広がる	hirogaru	spread, widen	広げる	hirogeru	spread, widen
広まる	hiromaru	spread	広める	hiromeru	spread, broaden
殖える	fueru	increase	殖やす	fuyasu	increase
増える	fueru	increase	増やす	fuyasu	increase
深まる	fukamaru	deepen	深める	fukameru	deepen
膨らむ	fukuramu	swell, expand	膨らます	fukuramasu	inflate, expand, swell
降る	furu	rain, precipitate	降らす	furasu	rain upon, send
ぶら下がる	burasagaru	hang down, hang from	ぶら下げる	burasageru	suspend, hang
減る	heru	decrease, decrease (in size or number)	減らす	herasu	decrease, abate
曲がる	magaru	bend, turn	曲げる	mageru	bend
巻き上がる	makiagaru	roll up	巻き上げる	makiageru	roll up

Vi			Vt		
Japanese		English	Japanese		English
Type: **S** = **O**					
交ざる	mazaru	mix, be mixed	交ぜる	mazeru	mix, be mixed
纏まる	matomaru	be settled, conclude, be collected	纏める	matomeru	settle, conclude, put in order
回る	mawaru	turn, go round	回す	mawasu	turn, turn round
向く	muku	face, point towards	向ける	mukeru	face, point towards, turn towards
剥ける	mukeru	peel	剥く	muku	peel
燃える	moeru	burn	燃やす	moyasu	burn
戻る	modoru	return, turn back	戻す	modosu	return, restore
漏れる	moreru	leak, leak out	漏らす	morasu	let leak, leak
焼ける	yakeru	burn	焼く	yaku	burn, bake
和らぐ	yawaragu	soften	和らげる	yawarageru	soften
茹だる	yudaru	boil	茹でる	yuderu	boil
緩む	yurumu	loosen, be loose, become loose	緩める	yurumeru	loosen
揺れる	yureru	shake	揺る	yuru	shake
横たわる	yokotawaru	lie down	横たえる	yokotaeru	lay down, lie down
弱まる	yowamaru	weaken, abate	弱める	yowameru	weaken
沸く	waku	boil	沸かす	wakasu	boil
湧く	waku	gush	湧かす	wakasu	gush
渉る	wataru	extend	渉す	watasu	extend
割れる	wareru	split, break	割る	waru	split, break, divide
Type: Passive					
揚がる	agaru	be fried, rise	揚げる	ageru	fry, lift
暖まる	atatamaru	be warmed, get warm, warm oneself	暖める	atatameru	warm
集まる	atsumaru	gather, be gathered	集める	atsumeru	collect, gather
余る	amaru	remain, be left	余す	amasu	save, leave
改まる	aratamaru	be renewed	改める	aratameru	renew, change

Vi			Vt		
Japanese		English	Japanese		English
Type: Passive					
荒れる	areru	be devastated, be stormy	荒らす	arasu	devastate, lay waste
傷む	itamu	be damaged	傷める	itameru	damage
苛立つ	iradatsu	be excited, be irritated	苛立てる	iradateru	excite, irritate
埋もれる	uzumoreru	be buried	埋める	uzumeru	bury
映る	utsuru	be reflected	映す	utsusu	reflect, project
埋まる	umaru	be buried	埋める	umeru	bury, bury (e.g. one's face in hands)
売り切れる	urikireru	be sold out	売り切る	urikiru	sell out
潤う	uruou	be moistened	潤す	uruosu	moisten
植わる	uwaru	be planted	植える	ueru	plant
治まる	osamaru	be ruled, be at peace	治める	osameru	rule, govern
収まる	osamaru	subside, be obtained	収める	osameru	suppress, obtain
納まる	osamaru	be reached, be obtained	納める	osameru	reach, obtain
驚く	odoroku	be surprised	驚かす	odorokasu	surprise
おぶさる	obusaru	be carried	おぶう	obuu	carry
孵る	kaeru	be hatched	孵す	kaesu	hatch
掠れる	kasureru	be grazed	掠る	kasuru	graze
片付く	kataduku	be in order, be tidied, put in order	片付ける	katadukeru	tidy up, tidy
絡む	karamu	twine round, be tangled	搦める	karameru	bind, tangle
聴こえる	kikoeru	be heard	聴く	kiku	hear
聞こえる	kikoeru	be audible, be heard	聞く	kiku	hear
決まる	kimaru	be decided	決める	kimeru	decide
清まる	kiyomaru	become pure, be purified	清める	kiyomeru	purify
窮まる	kiwamaru	go to extremes, be mastered, reach limit, terminate	窮める	kiwameru	attain, master, take to limit, carry to extremes
挫ける	kujikeru	be discouraged, be crushed	挫く	kujiku	break, crush
曇る	kumoru	be clouded	曇らす	kumorasu	cloud

Vi		Vt	
Japanese	English	Japanese	English
Type: Passive			
くるまる kurumaru	be wrapped up	包む kurumu	wrap
焦がれる kogareru	be scorched	焦がす kogasu	scorch
こなれる konareru	be digested	こなす konasu	digest
壊われる kowareru	be broken	壊わす kowasu	break
割ける sakeru	be separated	割く saku	separate
刺さる sasaru	stick into, be stuck, stick	刺す sasu	thrust into, stick, pierce
定まる sadamaru	be decided, be fixed, become settled	定める sadameru	decide, fix
捌ける sabakeru	be sold, be in order	捌く sabaku	sell, handle
絞まる shimaru	be constricted, be strangled	絞める shimeru	constrict, strangle
締まる shimaru	tigheten, be tied, be shut	締める shimeru	tighten, tie
育つ sodatsu	grow up, be brought up, raise	育てる sodateru	bring up, be brought up
揃う sorou	match, be arranged, become complete	揃える soroeru	arrange, put things in order
助かる tasukaru	be saved, be rescued	助ける tasukeru	save, rescue, help
溜まる tamaru	be accumulated, collect	溜める tameru	accumulate, amass
撓む tawamu	be bent	撓める tawameru	bend
ちぎれる chigireru	be torn off	ちぎる chigiru	tear, cut up fine
縮まる chidimaru	shrink, be shortened	縮める chidimeru	reduce, shorten
散らかる chirakaru	be scattered, be in disorder	散らかす chirakasu	scatter, scatter around
掴まる tsukamaru	be caught	掴える tsukamaeru	catch
掴まる tsukamaru	be caught	掴む tsukamu	catch
漬かる tsukaru	be soaked in, be soaked, be pickled	漬ける tsukeru	soak
付く tsuku	adhere to, be attached, catch fire, adjoin	付ける tsukeru	attach, turn on

Vi			Vt		
Japanese		English	Japanese		English
Type: Passive					
伝わる	tsutawaru	be handed down, be conveyed	伝える	tsutaeru	convey
潰れる	tsubureru	be crushed, be smashed	潰す	tsubusu	crush, smash
詰まる	tsumaru	be stuffed, be blocked	詰める	tsumeru	stuff, block, pack
詰む	tsumu	be stuffed, become fine	詰める	tsumeru	stuff, pack
整う	totonou	be prepared	整える	totonoeru	prepare, put in order
捕らわれる	torawareru	be caught	捕らえる	toraeru	catch
蕩ける	torokeru	be bewitched	蕩かす	torokasu	bewitch
治る	naoru	be cured	治す	naosu	cure
直る	naoru	be fixed, be cured	直す	naosu	repair, cure
無くなる	nakunaru	be lost, disappear	無くなす	nakunasu	lose
悩む	nayamu	be distressed, be worried	悩ます	nayamasu	distress, afflict
煮詰まる	nitsumaru	be boiled down	煮詰める	nitsumeru	boil down
抜ける	nukeru	come off, be thrown, come out	抜く	nuku	remove, throw, extract
延びる	nobiru	be prolonged	延ばす	nobasu	prolong, lengthen
乗る	noru	ride, be placed on, get on	乗せる	noseru	give a ride, place on
外れる	hazureru	come off, be disconnected	外す	hazusu	take off, disconnect, unfasten
吹き飛ぶ	fukitobu	be blown off	吹き飛ばす	fukitobasu	blow off
塞がる	fusagaru	be blocked, be plugged up	塞ぐ	fusagu	block, stop up
凹む	hekomu	become hollow, be dented	凹ます	hekomasu	dent
隔たる	hedataru	be distant, be separated	隔てる	hedateru	separate
紛れる	magireru	be diverted	紛らす	magirasu	divert
捲れる	makureru	be tucked up, be turned up (inside out)	捲る	makuru	tuck up, verb suffix to indicate reckless abandon
負ける	makeru	be defeated, lose	負かす	makasu	defeat
まぶれる	mabureru	be smeared	まぶす	mabusu	smear

Vi			Vt		
Japanese		English	Japanese		English
Type: Passive					
迷う	mayou	be puzzled, be lost	迷わす	mayowasu	puzzle, lose
見える	mieru	be seen, be visible	見る	miru	see
乱れる	midareru	be disorderd, be disordered, get confused	乱す	midasu	put in disorder, disorder, throw out of order
満ちる	michiru	be filled, be full	満たす	mitasu	fill, satisfy
見付かる	mitsukaru	be found	見付ける	mitsukeru	find, be familiar
蒸れる	mureru	be steamed, be stuffy	蒸す	musu	steam
蒸れる	mureru	be steamed, be stuffy	蒸らす	murasu	steam, cook by steam
揉める	momeru	be wrinkled	揉む	momu	wrinkle
休まる	yasumaru	be rested	休める	yasumeru	set at ease, rest
破れる	yabureru	get broken, be torn, get torn	破る	yaburu	break, tear
敗れる	yabureru	be defeated	敗る	yaburu	defeat
汚れる	yogoreru	be stained, get dirty, become dirty, be dirty	汚す	yogosu	stain, disgrace, soil, dirty
喜ぶ	yorokobu	be delighted, be pleased	喜ばす	yorokobasu	please, delight
分かれる	wakareru	branch off, be divided	分ける	wakeru	divide
煩う	wazurau	be worried, worry about	煩わす	wazurawasu	worry, trouble
Type: Synthetic					
遊ぶ	asobu	play	遊ばす	asobasu	let play, let one play
生きる	ikiru	live	生かす	ikasu	revive, make live
輝く	kagayaku	shine	輝かす	kagayakasu	make shine, light up
被さる	kabusaru	get coverd	被せる	kabuseru	cover
絡まる	karamaru	become entwined	絡む	karamu	entwine
枯れる	kareru	get hoarse, dry up, wither	枯らす	karasu	make hoarse, exhaust, let wither, let dry
傷付く	kizutsuku	get hurt, be hurt	傷付ける	kizutsukeru	hurt, wound

Vi			Vt		
Japanese		English	Japanese		English
Type: Synthetic					
静まる	shizumaru	get quiet, quieten, quieten down	静める	shizumeru	quiet, calm, appease
滑る	suberu	slip, glide	滑らす	suberasu	let slip, let something slip
通る	tooru	pass, go through	通す	toosu	let pass, allow through
飛ぶ	tobu	fly	飛ばす	tobasu	let fly, skip over
泣く	naku	cry	泣かす	nakasu	make cry, make someone cry
無くなる	nakunaru	get lost, lose, disappear	無くす	nakusu	lose, remove, lose something
馴れる	nareru	become domesticated	馴らす	narasu	domesticate
逃げる	nigeru	escape	逃がす	nigasu	let escape, let loose
濁る	nigoru	get muddy, become muddy	濁す	nigosu	muddy, make muddy
温まる	nukumaru	get warm	温める	nukumeru	warm
温もる	nukumoru	get warm	温める	nukumeru	warm
濡れる	nureru	get wet	濡らす	nurasu	wet, dampen
寝る	neru	sleep, go to bed	寝かす	nekasu	make sleep, put to sleep
逃れる	nogareru	escape	逃す	nogasu	let escape, let loose
光る	hikaru	shine	光らす	hikarasu	make shine
膨れる	fukureru	swell	膨らます	fukuramasu	make swell
ふやける	fuyakeru	get soaked	ふやかす	fuyakasu	soak
丸まる	marumaru	be round	丸める	marumeru	make round
持つ	motsu	have	持たす	motasu	let have
漏る	moru	leak	漏らす	morasu	let leak
揺れる	yureru	sway	揺るがす	yurugasu	make sway
捩れる	yojireru	get twisted	捩る	yojiru	twist
煩らう	wazurau	worry	煩らわす	wazurawasu	make worried

Vi			Vt		
Japanese		English	Japanese		English
Type: Diff Head					
合う	au	match, fit	合わす	awasu	bring together, join together
仰のく	aonoku	look up	仰のける	aonokeru	turn up (one's face or a card)
仰向く	aomuku	look upward	仰向ける	aomukeru	turn up (ones face)
赤らむ	akaramu	become red	赤らめる	akarameru	blush
挙がる	agaru	rise, become prosperous	挙げる	ageru	raise
上がる	agaru	rise, enter	上げる	ageru	raise, give
飽きる	akiru	be fed up, get tired of	飽かす	akasu	weary, glut
明ける	akeru	dawn	明かす	akasu	spend(the night)
甘える	amaeru	fawn upon	甘やかす	amayakasu	spoil
現われる	arawareru	appear	現わす	arawasu	show
表われる	arawareru	appear	表わす	arawasu	express
合わさる	awasaru	get together	合わす	awasu	join together
合わさる	awasaru	get together	合わせる	awaseru	join together
併さる	awasaru	get together	併せる	awaseru	unite
泡立つ	awadatsu	bubble	泡立てる	awadateru	beat
怒る	ikaru	get angry	怒らす	ikarasu	anger someone
至る	itaru	reach	致す	itasu	bring about
弥増さる	iyamasaru	become still greater	弥増す	iyamasu	increase (all the more)
入れ代わる	irekawaru	change places	入れ替える	irekaeru	replace
入れ違う	irechigau	pass each other	入れ違える	irechigaeru	misplace
受かる	ukaru	pass(an exam), pass	受ける	ukeru	take(an exam), undertake
薄まる	usumaru	become thin, become weak	薄める	usumeru	make thin, dilute
失せる	useru	disappear	失う	ushinau	lose
俯く	utsumuku	look down	俯く	utsumukeru	cast down
俯く	utsumuku	look downward	俯ける	utsumukeru	turn upside down
映る	utsuru	be reflected	写す	utsusu	copy

Vi			Vt		
Japanese		English	Japanese		English
Type: Diff Head					
生まれる	umareru	be born	生む	umu	give birth
縁付く	enduku	marry	縁付ける	endukeru	marry off
遅れる	okureru	be late for, be late	遅らす	okurasu	delay, retard
起こる	okoru	happen, occur	起こす	okosu	arouse, raise
興る	okoru	rise	興す	okosu	revive
修まる	osamaru	be controled, govern oneself	修める	osameru	control, study
押し詰まる	oshitsumaru	approach the year end	押し詰める	oshitsumeru	pack (in box)
教わる	osowaru	learn	教える	oshieru	teach
落ち着く	ochitsuku	calm down	落ち着ける	ochitsukeru	quiet
落ちる	ochiru	fall, fail (e.g. exam)	落とす	otosu	drop
脅える	obieru	be frightend	脅かす	obiyakasu	threaten
思い浮かぶ	omoi-ukabu	occur to	思い浮かべる	omoi-ukaberu	be reminded of
及ぶ	oyobu	reach	及ぼす	oyobosu	influence
折り重なる	orikasanaru	lie on top of one another	折り重ねる	orikasaneru	fold back
下りる	oriru	get off, go down, alight (e.g. from bus)	下ろす	orosu	let off, lower, take down
懸かる	kakaru	be suspended from	懸ける	kakeru	hang
懸け隔たる	kake-hedataru	far apart	懸け隔てる	kake-hedateru	put distance between
駆ける	kakeru	gallop	駆る	karu	spur on
叶う	kanau	come true	叶える	kanaeru	grant (request, wish)
絡まる	karamaru	twine round	絡める	karameru	bind
消える	kieru	go out, disappear	消す	kesu	extinguish
利く	kiku	be effective	利かす	kikasu	use
着る	kiru	wear	着せる	kiseru	dress, put on clothes
切れる	kireru	be cut off	切らす	kirasu	run short of
食い合う	kuiau	fit together	食い合わす	kuiawasu	clench
食う	kuu	eat	食わす	kuwasu	feed

Vi			Vt		
Japanese		English	Japanese		English
Type: Diff Head					
くすぶる	kusuburu	smoke	くすべる	kusuberu	fumigate
崩れる	kuzureru	collapse	崩す	kuzusu	destroy
下る	kudaru	go down, descend, get down	下す	kudasu	lower
下る	kudaru	go down	下さる	kudasaru	bestow
くっつく	kuttsuku	adhere to	くっつける	kuttsukeru	attach
窪まる	kubomaru	be low (as a hollow)	凹める	kubomeru	hollow out
眩む	kuramu	grow dizzy	晦ます	kuramasu	dazzle
狂う	kuruu	go mad	狂わす	kuruwasu	drive mad
苦しむ	kurushimu	suffer	苦しめる	kurushimeru	torment, afflict
くるまる	kurumaru	be wrapped up	くるめる	kurumeru	lump together
黒まる	kuromaru	blacken	黒める	kuromeru	make something black
加わる	kuwawaru	join, join in	加える	kuwaeru	add, add to, append
肥える	koeru	get fat, become fertile, grow fat	肥やす	koyasu	fertilize, make fertile
こじれる	kojireru	get worse	こじらす	kojirasu	aggravate
言付かる	kotodukaru	be entrusted with	言付ける	kotodukeru	send word
事寄せる	kotoyoseru	pretend	事寄す	kotoyosu	find an excuse
篭もる	komoru	be full of	込める	komeru	include
懲りる	koriru	learn by experience	懲らす	korasu	chastise
逆立つ	sakadatsu	bristle, stand up	逆立てる	sakadateru	ruffle up, ruffle
先立つ	sakidatsu	lead	先立てる	sakidateru	have go ahead
授かる	sazukaru	receive, be gifted	授ける	sazukeru	grant
騒ぐ	sawagu	be excited	騒がす	sawagasu	agitate
従う	shitagau	go along with	従える	shitagaeru	take along with
知れる	shireru	become known	知る	shiru	come to know
焦れる	jireru	be impatient, get impatient	焦らす	jirasu	irritate
吸い付く	suitsuku	stick to	吸い付ける	suitsukeru	attract
透く	suku	be transparent	透かす	sukasu	look through

Vi			Vt		
Japanese		English	Japanese		English
Type: Diff Head					
竦む	sukumu	crouch, cower	竦める	sukumeru	duck(head), shrug
透ける	sukeru	be transparent	透かす	sukasu	look through
進む	susumu	make progress	進める	susumeru	advance
刷り上がる	suriagaru	be off the press	刷り上げる	suriageru	finish printing
擦り切れる	surikireru	wear out	擦り切る	surikiru	cut by rubbing
摩り切れる	surikireru	wear out	摩り切る	surikiru	cut by rubbing
擦り剥ける	surimukeru	abrade	擦り剥く	surimuku	skin (one's knee)
座る	suwaru	sit, squat	据える	sueru	set
ずれる	zureru	slip, slide	ずらす	zurasu	shift, put off
添う	sou	go along with, accompany	添える	soeru	add to
殺げる	sogeru	be hollow	殺ぐ	sogu	diminish
聳える	sobieru	rise	聳やかす	sobiyakasu	raise
逸れる	soreru	deviate, stray from subject	逸らす	sorasu	divert, turn away
絶える	taeru	die out	絶やす	tayasu	exterminate
倒れる	taoreru	fall down, collapse	倒す	taosu	bring down, knock down, throw down
高まる	takamaru	rise	高める	takameru	raise
立ち上がる	tachiagaru	stand up	立ち上げる	tachiageru	boot (a computer)
建つ	tatsu	stand	建てる	tateru	build
近付く	chikaduku	approach	近付ける	chikadukeru	allow to come near, bring near
近寄る	chikayoru	approach	近寄せる	chikayoseru	bring close to
力付く	chikara-duku	by force	力付ける	chikara-dukeru	encourage
違う	chigau	deviate, differ	違える	chigaeru	break (one's word), alter, change
潰える	tsuieru	collapse	費やす	tsuiyasu	waste
費える	tsuieru	collapse	費やす	tsuiyasu	spend
使える	tsukaeru	be usable	使う	tsukau	use

Vi			Vt		
Japanese		English	Japanese		English
Type: Diff Head					
番う	tsugau	mate	番える	tsugaeru	fix (an arrow to a string)
突き刺さる	tsukisasaru	stick into	突き刺す	tsukisasu	stab
突き通る	tsukitooru	penetrate	突き通す	tsukitoosu	pierce
就く	tsuku	settle in	就ける	tsukeru	put
着く	tsuku	arrive at	着ける	tsukeru	arrive
突っ立つ	tsuttatsu	stand up	突っ立てる	tsuttateru	stab
約まる	tsudumaru	shrink	約める	tsudumeru	reduce
勤まる	tsutomaru	be fit for	勤める	tsutomeru	serve
務まる	tsutomaru	be fit for	務める	tsutomeru	act as
繋がる	tsunagaru	be tied together	繋げる	tsunageru	connect
積み重なる	tsumi-kasanaru	accumulate	積み重ねる	tsumi-kasaneru	be piled up
強まる	tsuyomaru	grow strong, get strong	強める	tsuyomeru	strenghten, strengthen
連なる	tsuranaru	lie in a row, extend	連ねる	tsuraneru	line up, link
照れる	tereru	be shy	照らす	terasu	shine on
出る	deru	come out, go out	出す	dasu	remove, send out
遠ざかる	toozakaru	withdraw, go far off	遠ざける	toozakeru	shun, keep away
遠退く	toonoku	become distant	遠退ける	toonokeru	keep at a distance
とがる	togaru	become sharp	尖らす	togarasu	sharpen
届く	todoku	arrive, reach	届ける	todokeru	deliver
留まる	todomaru	be fixed	留める	todomeru	stop
泊まる	tomaru	stay(at), stay, stay at (e.g. hotel)	泊める	tomeru	lodge, give shelter to
富む	tomu	grow rich, be rich	富ます	tomasu	enrich
点る	tomoru	burn	点す	tomosu	light
採れる	toreru	be produced	採る	toru	take
取れる	toreru	come off	取る	toru	take
退く	doku	retreat	退かす	dokasu	remove
流れる	nagareru	flow, stream	流す	nagasu	wash away, pour, drain

Vi			Vt		
Japanese		English	Japanese		English
Type: Diff Head					
亡くなる	nakunaru	die	亡くす	nakusu	lose someone, lose someone, wife, child, etc
亡くなる	nakunaru	die	亡くなす	nakunasu	lose someone, wife, child, etc
懐く	natsuku	become attached to, be attached, become emotionally attached	懐ける	natsukeru	win over, gain affection
靡く	nabiku	bend	靡かす	nabikasu	seduce
慣れる	nareru	get used to, grow accustomed to	慣らす	narasu	tame, accustom
匂う	niou	be fragrant	匂わす	niowasu	give out an odor, scent or perfume
賑わう	nigiwau	prosper	賑わす	nigiwasu	make prosperous
濁ごる	nigoru	become muddy	濁ごす	nigosu	make muddy
鈍る	niburu	become less capable	鈍らす	niburasu	blunt
似る	niru	resemble	似せる	niseru	imitate, copy
抜ける	nukeru	come off, come out	抜かす	nukasu	leave out, omit
脱げる	nugeru	come off	脱ぐ	nugu	take off, take off clothes
温む	nurumu	become lukewarm	温める	nurumeru	make less hot
捻じれる	nejireru	twist	捻じる	nejiru	screw
寝る	neru	sleep	寝せる	nekaseru	put to sleep
退く	noku	retreat, get out of the way	退ける	nokeru	expel, repel, remove
残る	nokoru	remain	残す	nokosu	leave, leave over, leave (behind, over)
伸し上がる	noshiagaru	stand on tiptoe	伸し上げる	noshiageru	promote
載る	noru	appear in print, appear (in print)	載せる	noseru	publish, place on
入る	hairu	enter	入れる	ireru	put in
はぐれる	hagureru	go astray	はぐらかす	hagurakasu	put off

Vi			Vt		
Japanese		English	Japanese		English
Type: Diff Head					
剥げる	hageru	come off	剥がす	hagasu	peel off, tear off
剥げる	hageru	come off	剥ぐ	hagu	peel off, tear off
跳ね返る	hanekaeru	rebound	跳ね返す	hanekaesu	reject
張り付く	haritsuku	cling	張り付ける	haritsukeru	attach to a flat surface with glue
貼り付く	haritsuku	cling	張り付ける	haritsukeru	attach to a flat surface with glue
ばれる	bareru	come to light, leak out (a secret)	ばらす	barasu	expose
引き下がる	hikisagaru	withdraw	引き下げる	hikisageru	pull down
引き締まる	hikishimaru	become tense	引き締める	hikishimeru	tighten
引き立つ	hikitatsu	become active	引き立てる	hikitateru	favour
引っ掛かる	hikkakaru	be caught in	引っ掛ける	hikkakeru	hang on
ひっくり返る	hikkuri-kaeru	be overturned	ひっくり返す	hikkuri-kaesu	turn over
引っ込む	hikkomu	draw back	引っ込める	hikkomeru	pull back, draw in
引っ込む	hikkomu	withdraw	引っ込ます	hikkomasu	pull back
閃く	hirameku	flash (of thunder)	閃かす	hiramekasu	brandish
封じ込む	fuujikomu	entrap	封じ込める	fuujikomeru	shut in
吹く	fuku	blow, blow (wind, etc)	吹かす	fukasu	puff, smoke (a cigarette)
蒸ける	fukeru	become ready to eat (as a result of s	蒸かす	fukasu	steam
伏す	fusu	lie down	伏せる	fuseru	lay down
降り込む	furikomu	rain upon	降り込める	furikomeru	rain (or snow), keeping people indoor
震う	furuu	shake	震える	furueru	shiver
ぶつかる	butsukaru	appear, hit	ぶつける	butsukeru	display, strike
ほぐれる	hogureru	come untied	ほぐす	hogusu	untie
細る	hosoru	become thin	細める	hosomeru	make narrow
解ける	hodokeru	loosen, come untied, come apart	解く	hodoku	solve, untie, take apart, unfasten

Vi			Vt		
Japanese		English	Japanese		English
Type: Diff Head					
仄めく	honomeku	be seen dimly	仄めかす	honomekasu	hint at
滅びる	horobiru	go to ruin, perish, be ruined	滅ぼす	horobosu	destroy
ぼける	bokeru	fade	暈す	bokasu	shade off
巻き付く	makitsuku	twine around	巻き付ける	makitsukeru	wreathe (e.g rope)
跨る	matagaru	sit astride	跨ぐ	matagu	straddle
間違う	machigau	be wrong, make a mistake	間違える	machigaeru	mistake, make an error, err
惑う	madou	be puzzled	惑わす	madowasu	bewilder
見る	miru	see	見せる	miseru	show
結び付く	musubi-tsuku	be connected or related	結び付ける	musubi-tsukeru	combine
儲かる	moukaru	be profitable	儲ける	moukeru	earn, get
もげる	mogeru	come off	もぎる	mogiru	pluck off
持ち上がる	mochiagaru	lift	持ち上げる	mochiageru	raise
盛り上がる	moriagaru	rouse	盛り上げる	moriageru	pile up
焼き付く	yakitsuku	scorch	焼き付ける	yakitsukeru	burn or bake into
役立つ	yakudatsu	be useful	役立てる	yakudateru	put to use
宿る	yadoru	lodge (at), lodge	宿す	yadosu	conceive, keep
歪む	yugamu	warp	歪める	yugameru	bend
揺る	yuru	shake	揺らす	yurasu	rock
揺れる	yureru	shake	揺らす	yurasu	rock
寄る	yoru	approach, visit	寄せる	yoseru	let come near, collect
渡る	wataru	cross over	渡す	watasu	hand over, pass over
Type: Diff Structure					
切れる	kireru	be cut off, be cut, cut well	切る	kiru	cut
突き抜ける	tsukinukeru	pierce through	突き抜く	tsukinuku	pierce
照る	teru	shine	照らす	terasu	illuminate, shine on
振り向く	furimuku	turn one’s face	振り向ける	furimukeru	turn

Bibliography

Eneko Agirre, Timothy Baldwin, and David Martinez. Improving Parsing and PP attachment Performance with Sense Information. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: ACL/HLT-2008*, pp. 317–325, 2008.

Yasuhiro Akiba, Megumi Ishii, Hussein Almuallim, and Shigeo Kaneda. Learning English Verb Selection Rules from Hand-made Rules and Translation Examples. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-95*, pp. 206–220, Leuven, 1995.

Yasuhiro Akiba, Hiromi Nakaiwa, Satoshi Shirai, and Yoshifumi Ooyama. Interactive Generalization of a Translation Example Using Queries Based on a Semantic Hierarchy. In *Proceedings of the 12th IEEE International Conference on Tools with Artificial Intelligence: ICTAI-2000*, pp. 326–332, 2000.

Shigeaki Amano and Tadahisa Kondo. Estimation of mental lexicon size with word familiarity database. In *Proceedings of the 5th International Conference on Spoken Language Processing: ICSLP-98*, volume 5, pp. 2119–2122, 1998.

Shigeaki Amano and Tadahisa Kondo. *Nihongo-no Goi-Tokusei (Lexical properties of Japanese)*. Sanseido, 1999. (in Japanese).

Ulrich Apel. WaDokuJT - A Japanese-German Dictionary Database. In *Proceedings of Papillon 2002 Workshop (CDROM)*, 2002. (Lexicon at: `http://www.babbletower.net/`).

Jason Baldridge and Miles Osborne. Active learning and logarithmic opinion pools

for hpsg parse selection. *Natural Language Engineering*, Vol. 13, No. 1, pp. 1–32, 2007.

Timothy Baldwin, Francis Bond, and Ben Hutchinson. A Valency Dictionary Architecture for Machine Translation. In *Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-99*, pp. 207–217, Chester, UK, 1999.

Daniel M. Bikel. A Statistical Model for Parsing and Word-Sense Disambiguation. In *ACL-2000 Student Research Workshop*, pp. 1–7, Hong Kong, 2000.

Francis Bond and Satoshi Shirai. Practical and Efficient Organization of a Large Valency Dictionary. In *NLPRS-97 Workshop on Multilingual Information Processing*, Phuket, 1997.

Francis Bond, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. Design and Construction of a machine-tractable Japanese-Malay Dictionary. In *Proceedings of the 8th Machine Translation Summit: MT Summit VIII*, pp. 53–58, Santiago de Compostela, Spain, 2001.

Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. Bootstrapping a WordNet using multiple existing WordNets. In *Proceedings of the 6th International Conference on Language Resources and Evaluation: LREC-2008*, Marrakech, 2008.

J. W. Breen. Building an electronic Japanese-English dictionary. Japanese Studies Association of Australia Conference, 1995.

J. W. Breen. JMDict: a Japanese-multilingual dictionary. In *COLING-2004 Workshop on Multilingual Linguistic Resources*, pp. 71–78, Geneva, 2004.

Ben Bullock. Alternative sci.lang.japan frequently asked questions, 1999.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics: ACL-2007*, pp. 33–40, June 2007.

Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Massimiliano Ciaramita and Yasemin Altun. Broad-Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing: EMNLP-2006*, pp. 594–602, Sydney, Australia, July 2006.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. Minimal Recursion Semantics. An introduction. *Research on Language and Computation*, Vol. 3, No. 4, pp. 281–332, 2005.

Jean-Marc Desperrier. Analysis of the Results of a Collaborative Project for the Creation of a Japanese-French Dictionary. In *Proceedings of the Papillon 2002 Workshop (CDROM)*, 2002. (Lexicon at: `http://dico.fj.free.fr/dico.php`).

Mike Dillinger. Dictionary Development Workflow for MT: Design and Management. In *Proceedings of the 8th Machine Translation Summit: MT Summit VIII*, pp. 83–88, Santiago de Compostela, 2001.

Bonnie J. Dorr. Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. *Machine Translation Journal*, Vol. 12, No. 4, pp. 271–322, 1997.

Bonnie J. Dorr, Gina-Anne Levow, and Dekang Lin. Construction of a Chinese-English Verb Lexicon for Machine Translation. *Machine Translation Journal*, 17 (1–2), 2002.

EDR. Concept dictionary. Technical report, Japan Electronic Dictionary Research Institute, Ltd, April 1990.

Katrin Erk, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics: ACL-2003*, 2003.

Christine Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

Hisanori Furumaki and Hozumi Tanaka. The Consideration of <n-suru> for Construction of the Dynamic Lexicon. In *9th Annual Meeting of The Association for Natural Language Processing*, pp. 298–301, 2003. (in Japanese).

Ralph Grishman, Catherine Macleod, and Adam Myers. *COMLEX Syntax Reference Manual*. Proteus Project, NYU, 1998. (`http://nlp.cs.nyu.edu/comlex/refman.ps`).

Tom Gruber. Ontology. *Encyclopedia of Database Systems*, 2008.

Masahiko Haruno and Takefumi Yamazaki. High-performance Bilingual Text Alignment using Statistical and Dictionary Information. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics: ACL-96*, pp. 131–138, 1996.

Munpyo Hong, Young-Kil Kim, Sang-Kyu Park, and Youn-Jik Lee. Semi-automatic Construction of Korean-Chinese Verb Patterns based on Translation Equivalency. In *COLING-2004 Workshop on Multilingual Linguistic Resources*, pp. 87–92, Geneva, 2004.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. NAIST Text Corpus : Annotating Predicate-Argument and Coreference Relations. *IPSJ SIG Notes*, 2007, No. 7, pp. 71–78, 2007. ISSN 09196072. URL `http://ci.nii.ac.jp/naid/110006202767/`. (in Japanese).

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. Toward an MT System without Pre-Editing — Effects of New Methods in **ALT-J/E** —. In *Proceedings of the Third Machine Translation Summit: MT Summit III*, pp. 101–106, Washington DC, 1991.

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, Francis Bond, and Yoshie Omi. Automatic Determination of Semantic Attributes for User Defined Words in Japanese-to-English Machine Translation. *Journal of Natural Language Processing*, Vol. 2, No. 1, pp. 3–17, 1995. (in Japanese).

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. *Goi-Taikei — A Japanese Lexicon*. Iwanami Shoten, Tokyo, 1997. 5 volumes/CD-ROM.

IPA. IPAL (basic adjectives). Lexicon, Information-Technology Promotion Agency, Tokyo, Japan, 1994. (ftp://ftp.mgt.ipa.go.jp/pub/ipal).

IPA. IPAL (basic verbs). Lexicon, Information-Technology Promotion Agency, Tokyo, Japan, 1987. (ftp://ftp.mgt.ipa.go.jp/pub/ipal).

Wesley Jacobsen. *Transitivity in the Japanese Verbal System*. PhD thesis, University of Chicago, 1981. (Reproduced by the Indiana University Linguistics Club, 1982).

Christopher Johnson, Miriam Petruck, Collin Baker, Michael Ellsworth, Josef Ruppenhofer, and Charles Fillmore. Framenet: Theory and practice, 2002.

Toshiyuki Kanamaru, Masaki Murata, Kow Kuroda, and Hitoshi Isahara. Obtaining Japanese Lexical Units for Semantic Frames from Berkeley FrameNet using a Bilingual Corpus. In *Proceedings of the 6th International Workshop on Linguistically Interpreted Corpora (LINC-2005)*, pp. 11–20, Jeju Island, Korea, 2005.

Kaname Kasahara, Kazumitsu Matsuzawa, and Tsutomu Ishikawa. A Method for Judgment of Semantic Similarity between Daily-used Words by using Machine Readable Dictionaries. *Transactions of IPSJ*, Vol. 38, No. 7, pp. 1272–1283, 1997. (in Japanese).

Daisuke Kawahara and Sadao Kurohashi. Japanese Case Frame Construction by Coupling the Verb and its Closest Case Component. In *Proceedings of First International Conference on Human Language Technology Research: HLT 2001*, pp. 204–210, San Diego, 2001.

Daisuke Kawahara and Sadao Kurohashi. Gradual Fertilization of Case Frames. *Journal of Natural Language Processing*, Vol. 12, No. 2, pp. 109–132, 3 2005. (in Japanese).

Daisuke Kawahara and Sadao Kurohashi. Case Frame Compilation from the Web using High-performance Computing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation: LREC-2006*, pp. 1344–1347, 2006.

Adam Kilgarriff. Inheriting Verb Alternations. In *Proceedings of the Sixth Conference of the European Chapter of the ACL: EACL-1993*, pp. 213–221, Utrecht, 1993.

Haruhiko Kindaichi and Yasaburou Ikeda. *Gakken Japanese Dictionary 2nd edition*. Gakken Co., Ltd., 1988.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. Extending Verbnet with Novel Verb Classes. In *Proceedings of the 5th International Conference on Language Resources and Evaluation: LREC-2006*, Genoa, Italy, 2006.

Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 423–430, 2003.

Philipp Koehn and Christof Monz. Manual and Automatic Evaluation of Machine Translation between European Languages. In *HTL-NAACL-2006 Workshop on Statistical Machine Translation*, pp. 102–121, 2006.

Upali Sathyajith Kohomban and Wee Sun Lee. Learning Semantic Classes for Word Sense Disambiguation. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics: ACL-2005*, pp. 34–41, June 2005.

Kokken. *Bunrui Goi Hyo Database CD-ROM (Word List by Semantic Principles, Revised and Enlarged Edition)*. Studies in Corpus Linguistics. Dainippon-tosho, 2004.

Anna Korhonen. Semantically Motivated Subcategorization Acquisition. In *ACL-2002 Workshop on Unsupervised Lexical Acquisition*, Philadelphia, USA, 2002.

Taku Kudo and Hideto Kazawa. *Web Japanese N-gram Version 1*. Gengo Shigen Kyokai, 2007.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing: EMNLP-2004*, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Sadao Kurohashi and Makoto Nagao. Building a Japanese parsed corpus — while improving the parsing system. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, chapter 14, pp. 249–260. Kluwer Academic Publishers, 2003.

Kyoto University. *JUMAN version 6.0 Manual*, 2008.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *18th International Conference on Machine Learning*, pp. 282–289, 2001.

Beth Levin. *English Verb Classes and Alternations*. University of Chicago Press, Chicago, London, 1993.

Hang Li and Naoki Abe. Generalizing Case Frames Using a Thesaurus and the MDL principle. *Computational Linguistics*, Vol. 24, No. 2, pp. 217–244, 1998.

Robert Malouf. A Comparison of Algorithms for Maximum Entropy Parameter Estimation. In *Proceedings of the 6th Conference on Computational Natural Language Learning: CoNLL-2002*, Taipei, Taiwan, 2002.

Robert Malouf and Gertjan van Noord. Wide Coverage Parsing with Stochastic Attribute Value Grammars. In *IJCNLP-04 Workshop: Beyond shallow analyses - Formalisms and statistical modeling for deep analyses*. JST CREST, March 2004.

Christopher D. Manning. Automatic Acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics: ACL-93*, pp. 235–242, 1993.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert Macintyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The Penn Treebank: Annotating predicate argument structure. In *In ARPA Human Language Technology Workshop*. Morgan Kaufmann, 1994.

Diana McCarthy. Using Semantic Preferences to Identify Verbal Participation in Role Switching Alternations. In *Proceedings of the first Conference of the North American Chapter of the ACL: ANLP-NAACL-2000*, Seattle, WA, 2000.

Masaaki Murata, Masao Utiyama, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. CRL at Japanese dictionary-based task of SENSEVAL-2. *Journal of Natural Language Processing*, Vol. 10, No. 3, pp. 115–143, 2003. (in Japanese).

Hiromi Nakaiwa and Satoru Ikehara. Intrasentential Resolution of Japanese Zero Pronouns in a Machine Translation System using Semantic and Pragmatic Constraints. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-95*, pp. 96–105, Leuven, 1995.

Roberto Navigli. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics: COLING/ACL-2006*, pp. 105–112, July 2006.

Eric Nichols, Francis Bond, and Daniel Flickinger. Robust Ontology Acquisition from Machine-Readable Dictionaries. In *Proceedings of the International Joint Conference on Artificial Intelligence: IJCAI-2005*, pp. 1111–1116, Edinburgh, 2005.

Minoru Nishio, Etsutaro Iwabuchi, and Shizuo Mizutani. *Iwanami Kokugo Jiten Dai Go Han [Iwanami Japanese Dictionary Edition 5]*. Iwanami Shoten, Tokyo, 1994. (in Japanese).

Naoyuki Nomura and Kazunori Muraki. An Empirical Architecture for Verb Subcategorization Frame. In *Proceedings of the 16th International Conference on Computational Linguistics: COLING-96*, pp. 640–645, Copenhagen, 1996.

Stephan Oepen and Jan Tore Lønning. Discriminant-Based MRS Banking. In *Proceedings of the 5th International Conference on Language Resources and Evaluation: LREC-2006*, Genoa, Italy, May 2006.

Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christoper D. Manning. LinGO Redwoods: A Rich and Dynamic Treebank for HPSG. *Research on Language and Computation*, Vol. 2, No. 4, pp. 575–596, 2004.

Akira Oishi and Yuji Matsumoto. Detecting the Organization of Semantic Subclasses of Japanese Verbs. *International Journal of Corpus Linguistics*, Vol. 2, No. 1, pp. 65–89, 1997.

Kyonghee Paik, Francis Bond, and Satoshi Shirai. Using Multiple Pivots to Align Korean and Japanese Lexical Resources. In *NLPRS-2001 Workshop on Language Resources in Asia*, pp. 63–70, Tokyo, 2001.

Martha Palmer, Dan Gildea, and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1), 2005.

Carl Pollard and Ivan A. Sag. *Head Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, 1994.

Kiyoaki Shirai. Construction of a word sense tagged corpus for SENSEVAL-2 Japanese dictionary task. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation: LREC-2002*, pp. 605–608, 2002.

Satoshi Shirai. Toward Collecting All Valency Patterns, –from the Viewpoint of Japanese-to-English Machine Translation–. In *Symposium on Sharing and reusing linguistic resources*, 1999. (in Japanese).

Shogakukan and Peking Shomoinshokan, editors. *Ri-Zhong Cidian [Japanese-Chinese Dictionary]*. Shogakukan, 1987.

Melanie Siegel and Emily M. Bender. Efficient Deep Processing of Japanese. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computational Linguistics*, Taipei, 2002.

Jun Suzuki, Erik McDermott, and Hideki Isozaki. Training Conditional Random Fields with Multivariate Evaluation Measures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 217–224, Sydney, Australia, July 2006. Association for Computational Linguistics.

Kouichi Takeuchi, Construction of Verb Dictionary Based on Lexical Conceptual Structure. In *Proceedings of Proceedings of the 10th Annual Meeting of The Association for Natural Language Processing: NLP-2004*, pp. 576–579, 2004. (in Japanese).

Kumiko Tanaka, Kyoji Umemura, and Hideya Iwasaki. Construction of a Bilingual Dictionary Intermediated by a Third Language. *Transactions of the Information Processing Society of Japan*, Vol. 39, No. 6, pp. 1915–1924, 1998. (in Japanese).

Kristina Toutanova, Christopher D. Manning, Dan Flickinger, and Stephan Oepen. Stochastic HPSG Parse Disambiguation using the Redwoods Corpus. *Research on Language and Computation*, Vol. 3, No. 1, pp. 83–105, 2005.

Yuta Tsuboi, Hisashi Kashima, Shinsuke Mori, Hiroki Oda, and Yuji Matsumoto. Training Conditional Random Fields using Incomplete Annotations. In *Proceedings of the 22nd International Conference on Computational Linguistics: COLING-2008*, 2008.

Takehito Utsuro, Takashi Miyata, and Yuji Matsumoto. Maximum Entropy Model Learning of Subcategorization Preference. In *Proceedings of the 5th Workshop on Very Large Corpora*, pp. 246–260, 1997.

Deyi Xiong, Qun Liu Shuanglong Li and, Shouxun Lin, and Yueliang Qian. Parsing the Penn Chinese Treebank with Semantic Knowledge. In Robert Dale, Jian Su Kam-Fai Wong and, and Oi Yee Kwong, editors, *Natural Language Processing — IJCNLP 005: Second International Joint Conference Proceedings*, pp. 70–81. Springer-Verlag, 2005.

Mitsuko Yamura-Takei, Miho Fujiwara, Makoto Yoshie, and Teruaki Aizawa. Automatic Linguistic Analysis for Language Teachers: The Case of Zeros. In *Proceedings of the 19th International Conference on Computational Linguistics: COLING-2002*, pp. 1114–1120, Taipei, 2002.

Akio Yokoo, Hiromi Nakaiwa, Satoshi Shirai, and Satoru Ikehara. Skeleton-Flesh Type Semantic Structure Dictionaries for Japanese-to-English Machine Translation. In *Proceedings of the 48th Annual Convention of the IPSJ*, pp. 139–140, 1994. (in Japanese).

List of Publications

List of Major Publications

Refereed Journal Publications

Sanae Fujita and Francis Bond. An Investigation into the Nature of Verbal Alternations and their Use in the Creation of Bilingual Valency Entries. *Journal of Natural Language Processing*, Vol. 12, No. 3, pp. 67–89, 2005. (in Japanese).

Sanae Fujita and Francis Bond. A Method of Creating New Valency Entries. *Machine Translation Journal*, Vol. 21, No. 1, pp. 1–28, 2007.

International Conferences

Sanae Fujita and Francis Bond. A Method of Adding New Entries to a Valency Dictionary by Exploiting Existing Lexical Resources. In *Proceedings of the Ninth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-2002*, pp. 42–52, Keihanna, Japan, 2002.

Sanae Fujita and Francis Bond. Extending the Coverage of a Valency Dictionary. In *Proceedings of COLING-2002 workshop on Machine Translation in Asia*, pp. 67–73, Taipei, 2002.

Francis Bond and Sanae Fujita. Evaluation of a Method of Creating New Valency Entries. In *Proceedings of the 9th Machine Translation Summit: MT Summit IX*, pp. 16–23, New Orleans, 2003.

Sanae Fujita and Francis Bond. A Method of Creating New Bilingual Valency Entries

using Alternations. In *Proceedings of COLING-2004 Workshop on Multilingual Linguistic Resources*, pp. 41–48, Geneva, 2004.

Sanae Fujita and Francis Bond. An Automatic Method of Creating New Valency Entries using Plain Bilingual Dictionaries. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-2004*, pp. 55–64, Baltimore, 2004.

Sanae Fujita, Francis Bond, Stephan Oepen, and Takaaki Tanaka. Exploiting Semantic Information for HPSG Parse Selection. In *Proceedings of ACL-2007 Workshop on Deep Linguistic Processing*, pp. 25–32, Prague, Czech Republic, June 2007.

List of Other Publications

Refereed Journal Publications

Francis Bond, Sanae Fujita, and Takaaki Tanaka. The Hinoki Syntactic and Semantic Treebank of Japanese. *Language Resources and Evaluation*, Vol. 40, No. 3–4, pp. 253–261, 2006. (Special issue on Asian language technology).

International Conferences

Francis Bond, Eric Nichols, Sanae Fujita, and Takaaki Tanaka. Acquiring an Ontology for a Fundamental Vocabulary. In *Proceedings of the 20th International Conference on Computational Linguistics: COLING-2004*, pp. 1319–1325, Geneva, 2004.

Sanae Fujita and Francis Bond. Evaluation of a Method of Creating New Valency Entries. *AAMT Journal*, No. 35, pp. 19–20, 2004. (in Japanese).

Takaaki Tanaka, Francis Bond, Stephan Oepen, and Sanae Fujita. High Precision Treebanking – Blazing Useful Trees Using POS Information. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics: ACL-2005*, pp. 330–337, 2005.

Takaaki Tanaka, Francis Bond, Stephan Oepen, and Sanae Fujita. High Precision Treebanking – Blazing Useful Trees Using POS Information. In *Proceedings of the*

11th Annual Meeting of The Association for Natural Language Processing: NLP-2005, pp. 994–997, Takamatsu, 2005.

Sanae Fujita, Takaaki Tanaka, Francis Bond, and Hiromi Nakaiwa. An Implemented Description of Japanese: The Lexeed Dictionary and the Hinoki Treebank. In *Proceedings of COLING/ACL-2006 Interactive Presentation Sessions*, pp. 65–68, Sydney, 2006.

Eric Nichols, Francis Bond, Takaaki Tanaka, Sanae Fujita, and Daniel Flickinger. Robust Ontology Acquisition from Multiple Sources. In *Proceedings of COLING-2006 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pp. 10–17, Sydney, 2006.

Takaaki Tanaka, Francis Bond, and Sanae Fujita. The Hinoki Sensebank — A Large-Scale Word Sense Tagged Corpus of Japanese —. In *Proceedings of the COLING/ACL-2006 Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pp. 62–69, Sydney, 2006.

Takaaki Tanaka, Francis Bond, Timothy Baldwin, Sanae Fujita, and Chikara Hashimoto. Word Sense Disambiguation Incorporating Lexical and Structural Semantic Information. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning: EMNLP-CoNLL-2007*, pp. 477–485, 2007.

Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. A Japanese Predicate Argument Structure Analysis using Decision Lists. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing: EMNLP-2008*, pp. 522–531, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

Timothy Baldwin, Su Nam Kim, Francis Bond, Sanae Fujita, David Martinez, and Takaaki Tanaka. MRD-based Word Sense Disambiguation: Further Extending Lesk. In *Proceedings of the Third International Joint Conference on Natural Language Processing: IJCNLP-2008*, pp. 775–780, Hyderabad, India, 2008.

Others

Francis Bond, Timothy Baldwin, and Sanae Fujita. Detecting Alternation Instances in a Valency Dictionary. In *Proceedings of the 8th Annual Meeting of The Association for Natural Language Processing: NLP-2002*, pp. 519–522. The Association for Natural Language Processing, 2002.

Sanae Fujita and Francis Bond. K$\bar{o}$taikankei-wo riy$\bar{o}$shitaketsug$\bar{o}$kajisho-no kakutokuh$\bar{o}$h$\bar{o}$ [A Method of Creating a Valency Dictionary using Alternations]. In *Proceedings of the 9th Annual Meeting of The Association for Natural Language Processing: NLP-2003*, pp. 361–364, 2003. (in Japanese).

Francis Bond, Sanae Fujita, Chikara Hashimoto, Kaname Kasahara, Shigeko Nariyama, Eric Nichols, Akira Ohtani, Takaaki Tanaka, and Shigeaki Amano. The Hinoki Treebank: A Treebank for Text Understanding. In *IEICE Technical Report: 2004-NLC-159*, pp. 83–90, 2004. (in Japanese).

Francis Bond, Sanae Fujita, Chikara Hashimoto, Kaname Kasahara, Shigeko Nariyama, Eric Nichols, Akira Ohtani, Takaaki Tanaka, and Shigeaki Amano. The Hinoki Treebank: Working toward Text Understanding. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora: LINC-04*, pp. 7–10, Geneva, 2004.

Francis Bond, Sanae Fujita, Chikara Hashimoto, Kaname Kasahara, Shigeko Nariyama, Eric Nichols, Akira Ohtani, Takaaki Tanaka, and Shigeaki Amano. The Hinoki Treebank: A Treebank for Text Understanding. In *Proceedings of the First International Joint Conference on Natural Language Processing: IJCNLP-2004*, pp. 554–559, Hainan Island, 2004.

Francis Bond, Sanae Fujita, Chikara Hashimoto, Shigeko Nariyama, Eric Nichols, Akira Ohtani, and Takaaki Tanaka. Development of the Hinoki Treebank Based on a Precise Grammar. In *IEICE Technical Report: 2004-NLC-159*, pp. 91–98, 2004. (in Japanese).

Sanae Fujita and Francis Bond. Evaluation of a Method of Creating New Valency Entries. *AAMT Journal*, No. 35, pp. 19–20, 2004. (in Japanese).

Kaname Kasahara, Hiroshi Sato, Francis Bond, Takaaki Tanaka, Sanae Fujita, Tomoko Kanasugi, and Shigeaki Amano. Construction of a Japanese Semantic Lexicon: Lexeed. SIG NLC-159, IPSJ, Tokyo, 2004. (in Japanese).

Francis Bond, Sanae Fujita, Chikara Hashimoto, Kaname Kasahara, Shigeko Nariyama, Eric Nichols, Akira Ohtani, Takaaki Tanaka, and Shigeaki Amano. The Hinoki Treebank: A Treebank for Text Understanding. *Lecture Notes in Computer Science, Natural Language Processing: IJCNLP 2004*, Vol. 3248, pp. 158–167, 2005.

Francis Bond, Sanae Fujita, Takaaki Tanaka, and Hiromi Nakaiwa. Nihongo-no tōgo/imi kōpasu Hinoki [Japanese Synactic-Semantic Copus — Hinoki]. In *Proceedings of the 11th Annual Meeting of The Association for Natural Language Processing: NLP-2005*, pp. 490–498, Takamatsu, 2005. (in Japanese).

Takaaki Tanaka, Francis Bond, Sanae Fujita, and Chikara Hashimoto. Word Sense Disambiguation Incorporating Lexical and Structural Semantic Information. In *Proceedings of the 13th Annual Meeting of The Association for Natural Language Processing: NLP-2007*, pp. 1086–1089, 2007. (in Japanese).

Sanae Fujita, Francis Bond, Stephan Oepen, and Takaaki Tanaka. Parse Ranking with Lexical and Structural Semantics. In *Proceedings of the 13th Annual Meeting of The Association for Natural Language Processing: NLP-2007*, pp. 1090–1093, 2007. (in Japanese).

Sanae Fujita, Francis Bond, and Akinori Fujino. Word Sense Disambiguation using Disambiguated Superordinate Semantic Classes. In *Proceedings of the 14th Annual Meeting of The Association for Natural Language Processing: NLP-2008*, pp. 568–571, 2008. (in Japanese).