Acquiring an Ontology for a Fundamental Vocabulary

Francis Bond* and **Eric Nichols** ** and **Sanae Fujita*** and **Takaaki Tanaka***
* {bond,fujita,takaaki}@cslab.kecl.ntt.co.jp ** eric-n@is.naist.jp
* NTT Communication Science Laboratories,
Nippon Telegraph and Telephone Corporation
** Nara Advanced Institute of Science and Technology

Abstract

In this paper we describe the extraction of thesaurus information from parsed dictionary definition sentences. The main data for our experiments comes from Lexeed, a Japanese semantic dictionary, and the Hinoki treebank built on it. The dictionary is parsed using a head-driven phrase structure grammar of Japanese. Knowledge is extracted from the semantic representation (Minimal Recursion Semantics). This makes the extraction process language independent.

1 Introduction

In this paper we describe a method of acquiring a thesaurus and other useful information from a machine-readable dictionary. The research is part of a project to construct a **fundamental vocabulary knowledge-base of Japanese**: a resource that will include rich syntactic and semantic descriptions of the core vocabulary of Japanese. In this paper we describe the automatic acquisition of a thesaurus from the dictionary definition sentences. The basic method has a long pedigree (Copestake, 1990; Tsurumaru et al., 1991; Rigau et al., 1997). The main difference from earlier work is that we use a mono-stratal grammar (Head-Driven Phrase Structure Grammar: Pollard and Sag (1994)) where the syntax and semantics are represented in the same structure. Our extraction can thus be done directly on the semantic output of the parser.

In the first stage, we extract the thesaurus backbone of our ontology, consisting mainly of **hypernym** links, although other links are also extracted (e.g., **domain**). We also link our extracted thesaurus to an existing ontology of Japanese: the Goi-Taikei ontology (Ikehara et al., 1997). This allows us to use tools that exploit the Goi-Taikei ontology, and also to extend it and reveal gaps.

The immediate application for our ontology is in improving the performance of stochastic models for parsing (see Bond et al. (2004) for further discussion) and word sense disambiguation. However, this paper discusses only the construction of the ontology.

We are using the Lexeed semantic database of Japanese (Kasahara et al. (2004), next section), a machine readable dictionary consisting of headwords and their definitions for the 28,000 most familiar open class words of Japanese, with all the definitions using only those 28,000 words (and some function words). We are parsing the definition sentences using an HPSG Japanese grammar and parser and treebanking the results into the Hinoki treebank (Bond et al., 2004). We then train a statistical model on the treebank and use it to parse the remaining definition sentences, and extract an ontology from them.

In the next phase, we will sense tag the definition sentences and use this information and the thesaurus to build a model that combines syntactic and semantic information. We will also produce a richer ontology — by combining information for word senses not only from their own definition sentences but also from definition sentences that use them (Dolan et al., 1993), and by extracting selectional preferences. Once we have done this for the core vocabulary, we will look at ways of extending our lexicon and ontology to less familiar words.

In this paper we present the details of the ontology extraction. In the following section we give more information about Lexeed and the Hi-

**Some of this research was done while the second author was visiting the NTT Communication Science Laboratories

noki treebank.We then detail our method for extracting knowledge from the parsed dictionary definitions (§ 3). Finally, we discuss the results and outline our future research (§ 4).

2 Resources

2.1 The Lexeed Semantic Database of Japanese

The Lexeed Semantic Database of Japanese is a machine readable dictionary that covers the most common words in Japanese (Kasahara et al., 2004). It is built based on a series of psycholinguistic experiments where words from two existing machine-readable dictionaries were presented to multiple subjects who ranked them on a familiarity scale from one to seven, with seven being the most familiar (Amano and Kondo, 1999). Lexeed consists of all open class words with a familiarity greater than or equal to five. The size, in words, senses and defining sentences is given in Table 1.

Table 1: The Size of Lexeed

Headwords	28,300
Senses	46,300
Defining Sentences	81,000

The definition sentences for these sentences were rewritten by four different analysts to use only the 28,000 familiar words and the best definition chosen by a second set of analysts. Not all words were used in definition sentences: the defining vocabulary is 16,900 different words (60% of all possible words were actually used in the definition sentences). An example entry for the word ドライバー *doraibā* "driver" is given in Figure 1, with English glosses added. The underlined material was not in Lexeed originally, we extract it in this paper. *doraibā* "driver" has a familiarity of 6.55, and three senses. The first sense was originally defined as just the synonym *nejimawashi* "screwdriver", which has a familiarity below 5.0. This was rewritten to the explanation: "A tool for inserting and removing screws".

2.2 The Hinoki Treebank

In order to produce semantic representations we are using an open source HPSG grammar of Japanese: JACY (Siegel and Bender, 2002), which we have extended to cover the dictionary definition sentences (Bond et al., 2004). We have treebanked 23,000 sentences using the [incr tsdb()] profiling environment (Oepen and Carroll, 2000) and used them to train a parse ranking model for the PET parser (Callmeier, 2002) to selectively rank the parser output. These tools, and the grammar, are available from the Deep Linguistic Processing with HPSG Initiative (DELPH-IN: `http://www.delph-in.net/`).

We use this parser to parse the defining sentences into a full meaning representation using minimal recursion semantics (MRS: Copestake et al. (2001)).

3 Ontology Extraction

In this section we present our work on creating an ontology. Past research on knowledge acquisition from definition sentences in Japanese has primarily dealt with the task of automatically generating hierarchical structures. Tsurumaru et al. (1991) developed a system for automatic thesaurus construction based on information derived from analysis of the terminal clauses of definition sentences. It was successful in classifying hyponym, meronym, and synonym relationships between words. However, it lacked any concrete evaluation of the accuracy of the hierarchies created, and only linked words not senses. More recently Tokunaga et al. (2001) created a thesaurus from a machine-readable dictionary and combined it with an existing thesaurus (Ikehara et al., 1997).

For other languages, early work for English linked senses exploiting dictionary domain codes and other heuristics (Copestake, 1990), and more recent work links senses for Spanish and French using more general WSD techniques (Rigau et al., 1997). Our goal is similar. We wish to link each word sense in the fundamental vocabulary into an ontology. The ontology is primarily a hierarchy of **hyponym** (is-a) relations, but also contains several other relationships, such as **abbreviation**, **synonym** and **domain**.

We extract the relations from the semantic output of the parsed definition sentences. The output is written in Minimal Recursion Semantics (Copestake et al., 2001). Previous work has successfully used regular expressions, both for

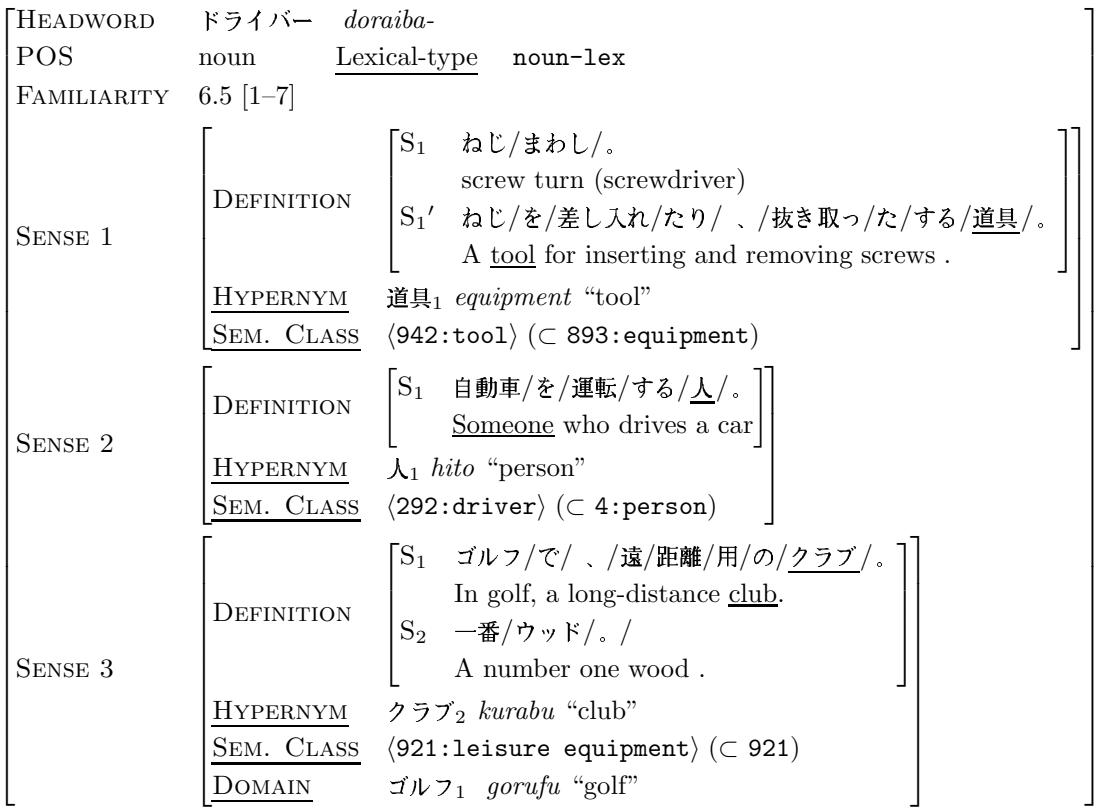

Figure 1: Entry for the Word *doraiba-* "driver" (with English glosses)

English (Barnbrook, 2002) and Japanese (Tsurumaru et al., 1991; Tokunaga et al., 2001). Regular expressions are extremely robust, and relatively easy to construct. However, we use a parser for four reasons. The first is that it makes our knowledge acquisition more language independent. If we have a parser that can produce MRS, and a machine readable dictionary for that language, the knowledge acquisition system can easily be ported. The second reason is that we can go on to use the parser and acquisition system to acquire knowledge from non-dictionary sources. Fujii and Ishikawa (2004) have shown how it is possible to identify definitions semi automatically, however these sources are not as standard as dictionaries and thus harder to parse using only regular expressions. The third reason is that we can more easily acquire knowledge beyond simple hypernyms, for example, identifying synonyms through common definition patterns as proposed by Tsuchiya et al. (2001). The final reason is that we are ultimately interested in language understanding, and thus wish to develop a parser. Any effort spent in building and refining regular expressions is not reusable, while creating and improving a grammar has intrinsic value.

3.1 The Extraction Process

To extract hypernyms, we parse the first definition sentence for each sense. The parser uses the stochastic parse ranking model learned from the Hinoki treebank, and returns the MRS of the first ranked parse. Currently, just over 80% of the sentences can be parsed.

An MRS consists of a bag of labeled elementary predicates and their arguments, a list of scoping constraints, and a pair of relations that provide a hook into the representation — a label, which must outscope all the handles, and an index (Copestake et al., 2001). The MRSs for the definition sentence for *doraibā*$_2$ and its English equivalent are given in Figure 2. The hook's label and index are shown first, followed by the list of elementary predicates. The figure omits some details (message type and scope have been suppressed).

$\langle h_0, x_1\{h_0 : prpstn\_rel(h_1)$ $\quad$ $\langle h_1, x_1\{h : prpstn\_rel(h_0)$
$h_1 : hito(x_1)$ $\quad$ $h_1 : person(x_1)$
$h_2 : u_def(x_1, h_1, h_6)$ $\quad$ $h_2 : some(x_1, h_1, h_6)$
$h_3 : jidosha(x_2)$ $\quad$ $h_3 : car(x_2)$
$h_4 : u_def(x_2, h_3, h_7)$ $\quad$ $h_4 : indef(x_1, h_3, h_7)$
$h_5 : unten(u_1, x_1, x_2)\}\rangle$ $\quad$ $h_5 : drive(u_1, x_1, x_2)\}\rangle$
「自動車を運転する人」 $\quad$ *somebody who drives a car*

Figure 2: Simplified MRS represntations for *doraibā*$_2$

In most cases, the first sentence of a dictionary definition consists of a fragment headed by the same part of speech as the headword. Thus the noun *driver* is defined as an noun phrase. The fragment consists of a **genus term** (*somebody*) and **differentia** (*who drives a car*).[1] The genus term is generally the most semantically salient word in the definition sentence: the word with the same index as the index of the hook. For example, for sense 2 of the word ドライバー *doraibā*, the hypernym is 人 *hito* "person" (Figure 2). Although the actual hypernym is in very different positions in the Japanese and English definition sentences, it is the hook in both the semantic representations.

For some definition sentences (around 20%), further parsing of the semantic representation is necessary. The most common case is where the index is linked to a coordinate construction. In that case, the coordinated elements have to be extracted, and we build two relationships. Other common cases are those where the relationship between the headword and the genus is given explicitly in the definition sentence: for example in (1), where the relationship is given as `abbreviation`. We initially process the relation, 略 *ryaku* "abbreviation", yielding the coordinate structure. This in turn gives two words: アルプス *arupusu* "alps" and 日本アルプス *nihon arupusus* "Japanese Alps". Our system thus produces two relations: `abbreviation`(ア,アルプス) and `abbreviation`(ア,日本アルプス). As can be seen from this example, special cases can embed each other, which makes the use of regular expressions difficult.

(1) ア: アルプス 、または 日本アルプス
a: arupusu , matawa nihon-arupusu
a: alps , or japan alps
の 略
no ryaku
ADN abbreviation
a: an abbreviation for the Alps or the Japanese Alps

The extent to which non-hypernym relations are included as text in the definition sentences, as opposed to stored as separate fields, varies from dictionary to dictionary. For knowledge acquisition from open text, we can not expect any labeled features, so the ability to extract information from plain text is important.

We also extract information not explicitly labeled, such as the domain of the word, as in Figure 3. Here the adpositional phrase representing the domain has wide scope — in effect the definition means "In golf, [a driver$_3$ is] a club for playing long strokes". The phrase that specifies the domain should modify a non-expressed predicate. To parse this, we added a construction to the grammar that allows an NP fragment heading an utterance to have an adpositional modifier. We then extract these modifiers and take the head of the noun phrase to be the domain. Again, this is hard to do reliably with regular expressions, as an initial NP followed by で could be a copula phrase, or a PP that attaches anywhere within the definition — not all such initial phrases restrict the domain. Most of the domains extracted fall under a few superordinate terms, mainly sport, games and religion. Other, more general domains, are marked explicitly in Lexeed as features. Japanese equivalents of the following words have a sense marked as being in the domain golf: *approach, edge, down, tee, driver, handicap, pin, long shot*.

We summarize the links acquired in Table 2, grouped by coarse part of speech. The first three lines show hypernym relations: implicit hypernyms (the default); explicitly indicated hypernyms, and implicitly indicated hyponyms.

[1] Also know as superordinate and discriminator or restriction.

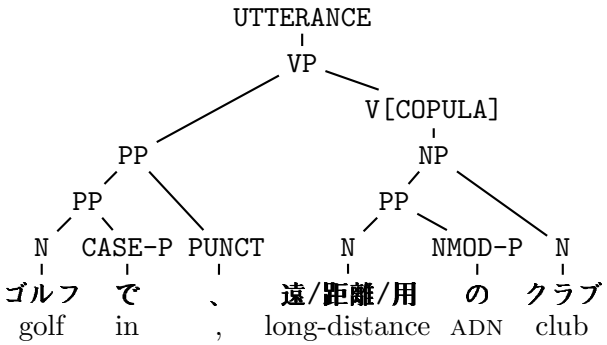

Figure 3: Parse for Sense_3 of Driver

The second three names show other relations: abbreviations, names and domains. Implicit hypernyms are by far the most common relations: fewer than 10% of entries are marked with an explicit relationship.

Relation Type	Noun	Verbal Noun	Verb	Other
Implicit	21,245	5,467	6,738	5,569
Hypernym	230	5		9
Hyponym	194	5		5
Abbreviation	423	35		76
Name	121			5
Domain	922	170		141

Table 2: Acquired Knowledge

3.2 Verification with Goi-Taikei

We verified our results by comparing the hypernym links to the manually constructed Japanese ontology Goi-Taikei. It is a hierarchy of 2,710 semantic classes, defined for over 264,312 nouns (Ikehara et al., 1997). Because the semantic classes are only defined for nouns (including verbal nouns), we can only compare nouns. Senses are linked to Goi-Taikei semantic classes by the following heuristic: look up the semantic classes C for both the headword (w_i) and the genus term(s) (w_g). If at least one of the index word's semantic classes is subsumed by at least one of the genus' semantic classes, then we consider their relationship confirmed (1).

$$\exists(c_h, c_g) : \{c_h \subset c_g; c_h \in C(w_h); c_g \in C(w_g)\} \quad (1)$$

In the event of an explicit hyponym relationship indicated between the headword and the genus, the test is reversed: we look for an instance of the genus' class being subsumed by the headword's class ($c_g \subset c_h$). Our results are summarized in Table 3. The total is 58.5% (15,888 confirmed out of 27,146). Adding in the named and abbreviation relations, the coverage is 60.7%. This is comparable to the coverage of Tokunaga et al. (2001), who get a coverage of 61.4%, extracting relations using regular expressions from a different dictionary.

3.3 Extending the Goi-Taikei

In general we are extracting pairs with more information than the Goi-Taikei hierarchy of 2,710 classes. For 45.4% of the confirmed relations both the headword and its genus term were in the same Goi-Taikei semantic class. In particular, many classes contain a mixture of class names and instance names: 豚肉 *buta niku* "pork" and 肉 *niku* "meat" are in the same class, as are ドラム *doramu* "drum" and 打楽器 *dagakki* "percussion instrument", which we can now distinguish. This conflation has caused problems in applications such as question answering as well as in fundamental research on linking syntax and semantics (Bond and Vatikiotis-Bateson, 2002).

An example of a more detailed hierarchy deduced from Lexeed is given in 4. All of the words come from the same Goi-Taikei semantic class: ⟨`842:condiment`⟩, but are given more structure by the thesaurus we have induced. There are still some inconsistencies: `ketchup` is directly under `condiment`, while `tomato sauce` and `tomato ketchup` are under `sauce`. This reflects the structure of the original machine readable dictionary.

4 Discussion and Further Work

From a language engineering point of view, we found the ontology extraction an extremely useful check on the output of the grammar/parser. Treebanking tends to focus on the syntactic structure, and it is all too easy to miss a malformed semantic structure. Parsing the semantic output revealed numerous oversights, especially in binding arguments in complex rules and lexical entries.

It also reveals some gaps in the Goi-Taikei coverage. For the word ドライバー *doraibā* "driver" (shown in Figure 1), the first two hypernyms are confirmed. However, ドライバーin

Relation		Noun		Verbal Noun
Implicit	56.66%	(12,037/21,245)	64.55%	(3,529/5,467)
Hypernym	56.52%	(134/230)	0%	(0/5)
Hyponym	94.32%	(183/194)	100%	(5/5)
Subtotal	57.01%	(12,354/21,669)	64.52%	(3,534/5,477)

Table 3: Links Confirmed by Comparison with Goi-Taikei

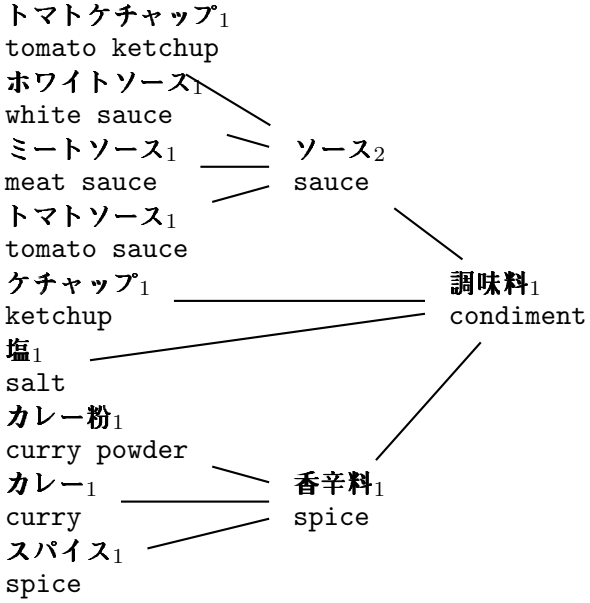

Figure 4: Refinement of the class `condiment`.

GT only has two semantic classes: ⟨`942:tool`⟩ and ⟨`292:driver`⟩. It does not have the semantic class ⟨`921:leisure equipment`⟩. Therefore we cannot confirm the third link, even though it is correct, and the `domain` is correctly extracted.

Further Work

There are four main areas in which we wish to extend this research: improving the grammar, extending the extraction process itself, further exploiting the extracted relations and creating a thesaurus from an English dictionary.

As well as extending the coverage of the grammar, we are investigating making the semantics more tractable. In particular, we are investigating the best way to represent the semantics of explicit relations such as 一種 *isshu* "a kind of".[2]

We are extending the extraction process by adding new explicit relations, such as 丁寧語 *teineigo* "polite form". For word senses such as driver$_3$, where there is no appropriate Goi-Taikei class, we intend to estimate the semantic class by using the definition sentence as a vector, and looking for words with similar definitions (Kasahara et al., 1997).

We are extending the extracted relations in several ways. One way is to link the hypernyms to the relevant word sense, not just the word. If we know that クラブ *club* "kurabu" is a hypernym of ⟨`921:leisure equipment`⟩, then it rules out the card suit "clubs" and the "association of people with similar interests" senses. Other heuristics have been proposed by Rigau et al. (1997). Another way is to use the thesaurus to predict which words name explicit relationships which need to be extracted separately (like *abbreviation*).

5 Conclusion

In this paper we described the extraction of thesaurus information from parsed dictionary definition sentences. The main data for our experiments comes from Lexeed, a Japanese semantic dictionary, and the Hinoki treebank built on it. The dictionary is parsed using a head-driven phrase structure grammar of Japanese. Knowledge is extracted from the semantic representation. Comparing our results with the Goi-Taikei hierarchy, we could confirm 60.73% of the relations extracted.

Acknowledgments

The authors would like to thank Colin Bannard, the other members of the NTT Machine Translation Research Group, NAIST Matsumoto Laboratory, and researchers in the DELPH-IN community, especially Timothy Baldwin, Dan Flickinger, Stephan Oepen and Melanie Siegel.

[2]These are often **transparent** nouns: those nouns which are transparent with regard to collocational or selection relations between their dependent and the external context of the construction, or transparent to number agreement (Fillmore et al., 2002).

References

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

Geoff Barnbrook. 2002. *Defining Language — A local grammar of definition sentences*. Studies in Corpus Linguistics. John Benjamins.

Francis Bond and Caitlin Vatikiotis-Bateson. 2002. Using an ontology to determine English countability. In *19th International Conference on Computational Linguistics: COLING-2002*, volume 1, pages 99–105, Taipei.

Francis Bond, Sanae Fujita, Chikara Hashimoto, Kaname Kasahara, Shigeko Nariyama, Eric Nichols, Akira Ohtani, Takaaki Tanaka, and Shigeaki Amano. 2004. The Hinoki treebank: A treebank for text understanding. In *Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*. Springer Verlag. (in press).

Ulrich Callmeier. 2002. Preprocessing and encoding techniques in PET. In Stephan Oepen, Dan Flickinger, Jun-ichi Tsujii, and Hans Uszkoreit, editors, *Collabarative Language Engineering*, chapter 6, pages 127–143. CSLI Publications, Stanford.

Ann Copestake, Alex Lascarides, and Dan Flickinger. 2001. An algebra for semantic construction in constraint-based grammars. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, Toulouse, France.

Ann Copestake. 1990. An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary. In *Proceedings of the First International Workshop on Inheritance in Natural Language Processing*, pages 19–29, Tilburg. (ACQUILEX WP NO. 8.).

William Dolan, Lucy Vanderwende, and Stephen D. Richardson. 1993. Automatically deriving structured knowledge from on-line dictionaries. In *Proceedings of the Pacific Association for Computational Linguistics*, Vancouver.

Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato. 2002. Seeing arguments through transparent structures. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 787–91, Las Palmas.

Atsushi Fujii and Tetsuya Ishikawa. 2004. Summarizing encyclopedic term descriptions on the web. In *20th International Conference on Computational Linguistics: COLING-2004*, Geneva. (this volume).

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

Kaname Kasahara, Kazumitsu Matsuzawa, and Tsutomu Ishikawa. 1997. A method for judgment of semantic similarity between daily-used words by using machine readable dictionaries. *Transactions of IPSJ*, 38(7):1272–1283. (in Japanese).

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

Stephan Oepen and John Carroll. 2000. Performance profiling for grammar engineering. *Natural Language Engineering*, 6(1):81–97.

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

German Rigau, Jordi Atserias, and Eneko Agirre. 1997. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of joint EACL/ACL 97*, Madrid.

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

Takenobu Tokunaga, Yasuhiro Syotu, Hozumi Tanaka, and Kiyoaki Shirai. 2001. Integration of heterogeneous language resources: A monolingual dictionary and a thesaurus. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, NLPRS2001*, pages 135–142, Tokyo.

Masatoshi Tsuchiya, Sadao Kurohashi, and Satoshi Sato. 2001. Discovery of definition patterns by compressing dictionary sentences. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, NLPRS2001*, pages 411–418, Tokyo.

Hiroaki Tsurumaru, Katsunori Takesita, Itami Katsuki, Toshihide Yanagawa, and Sho Yoshida. 1991. An approach to thesaurus construction from Japanese language dictionary. In *IPSJ SIGNotes Natural Language*, volume 83-16, pages 121–128. (in Japanese).