The Hinoki Treebank
A Treebank for Text Understanding

Francis Bond[1], Sanae Fujita[1], Chikara Hashimoto[2], Kaname Kasahara[1], Shigeko Nariyama[3], Eric Nichols[3], Akira Ohtani[4], Takaaki Tanaka[1], and Shigeaki Amano[1]

[1] NTT Communication Science Laboratories,
Nippon Telegraph and Telephone Corporation
<{bond, sanae, kaname, takaaki, amano}@cslab.kecl.ntt.co.jp>
[2] Kobe Shoin Women's University
<chashi@sils.shoin.ac.jp>
[3] Nara Advanced Institute of Science and Technology
<{eric-n, shigeko}@is.naist.ac.jp>
[4] Osaka Gakuin University
<ohtani@utc.osaka-gu.ac.jp>

Abstract. In this paper we describe the motivation for and construction of a new Japanese lexical resource: the Hinoki treebank. The treebank is built from dictionary definition sentences, and uses an HPSG grammar to encode the syntactic and semantic information. We then show how this treebank can be used to extract thesaurus information from definition sentences in a language-neutral way using minimal recursion semantics.

1 Introduction

In this paper we describe the construction of a new lexical resource: the Hinoki treebank. We present the motivation for its construction, and a preliminary application. The ultimate goal of our research is natural language understanding — we aim to create a system that can parse text into some useful semantic representation. Ideally this would be such that the output can be used to actually update our semantic models. This is an ambitious goal, and this paper does not present a completed solution, but rather a road-map to the solution, with some progress along the way. The mid-term goal is to build a thesaurus from dictionary definition sentences and use it to enhance a stochastic parse ranking model that combines syntactic and semantic information.

Recently, significant improvements have been made in combining symbolic and statistical approaches to various natural language processing tasks. For example, in parsing, symbolic grammars are being combined with stochastic models (Toutanova et al., 2002). Statistical techniques have also been shown to be useful for word sense disambiguation (Stevenson, 2003). However, to date, there have been no combinations of sense information together with symbolic grammars and statistical models. Klein and Manning (2003) show that much of the gain in statistical parsing using lexicalized models comes from the use of a small set of

function words. General relations between words are not so useful, presumably because the data is too sparse: in the Penn treebank normally used to train and test statistical parsers *stocks* and *skyrocket* never appear together. They note that this should motivate the use of similarity and/or class based approaches: the superordinate concepts capital ($\supset$ *stocks*) and move upward ($\supset$ *skyrocket*) frequently appear together. However, there has been little success reported on using ontologies with statistical parsers, despite the long history of their succesful use with rule-based systems (Ikehara et al., 1991; Mahesh et al., 1997).

We hypothesize that there are two major reasons for this lack of success. The first reason is that there simply is no single resource that combines syntactic and semantic tagging in a single corpus, so it is impossible to train statistical models using both sources of information. The second is that it is still not clear exactly what kind of semantic information is useful in parsing or how to obtain it.

Our proposed solution to these problems has three phases. In the first phase, we are building a treebank using the Japanese semantic database Lexeed (Kasahara et al., 2004). This is a hand built self-contained lexicon: it consists of headwords and their definitions for the most familiar 28,000 words of Japanese, with all the definitions using only those 28,000 words (and some function words). This set is large enough to include most basic level words and covers over 75% of the common word tokens in a sample of Japanese newspaper text. We then train a statistical model on the treebank and use it to help us induce a thesaurus. In phase two, we will tag the definition sentences with senses and use this information and the thesaurus to build a model that combines syntactic and semantic information. We will also produce a richer ontology — for example extracting qualia structures (Pustejovsky, 1995) and selectional preferences. In phase three, we will look at ways of extending our lexicon and ontology to less familiar words.

In this paper we discuss preliminary results from phase one. In particular, we introduce the construction of the treebank, building the statistical models and inducing the thesaurus. The technologies we are using in phase one are not new, the novelty is in the combination.

In the following section we give more information about Lexeed. Then, in Section 3 we discuss the creation of the treebank: Hinoki. The design is inspired by the Redwoods treebank of English (Oepen et al., 2002) a dynamic treebank closely linked to an HPSG analysis. Hinoki uses the JACY Japanese grammar (Siegel and Bender, 2002).

We describe the use of the Lexeed corpus and the grammar used in the treebank to create a stochastic parse ranker and a thesaurus (§ 4). Finally, we outline in more detail our path to the goal of understanding Japanese (§ 5)

2 The Lexeed Semantic Database of Japanese

The Lexeed Semantic Database of Japanese aims to cover the most common words in Japanese (Kasahara et al., 2004). It was built based on a series of psycholinguistic experiments where words from two existing machine-readable dictionaries were presented to subjects and they were asked to rank them on a

INDEX	ドライバー *doraibā*	
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/を/差し入れ/たり/ 、 抜き取っ/た/する/道具/。 A tool for inserting and removing screws .
	SEM. CLASS	⟨`942:tool`⟩ (⊂ 893:equipment)
SENSE 2	DEFINITION	S_1 自動車/を/運転/する/人/。 Someone who drives a car
	SEM. CLASS	⟨`292:driver`⟩ (⊂ 4:person)
SENSE 3	DEFINITION	S_1 ゴルフ/で/ 、/遠/距離/用/の/クラブ/。 In golf, a long-distance club. S_2 一番/ウッド/。/ A number one wood .
	SEM. CLASS	⟨`921:leisure equipment`⟩ (⊂ 921)
	DOMAIN	ゴルフ$_1$ *gorufu* “golf”

Fig. 1. Entry for the Word *doraibā* “driver” (with English glosses)

familiarity scale from one to seven, with seven being the most familiar (Amano and Kondo, 1999).

Lexeed consists of all words with a familiarity greater than or equal to five. There are 28,000 words in all. Many words have multiple senses, there were 46,347 different senses. 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. In the final configuration, 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 third sense has two defining sentences. There are 1.7 defining sentences/sense overall.

3 The Hinoki Treebank

The structure of our treebank is based on the Redwoods treebank of English (Oepen et al., 2002). The treebank is built up from the parse output of an HPSG grammar. We chose this structure for several reasons. The most important is that the representation is very rich. The treebank records the complete

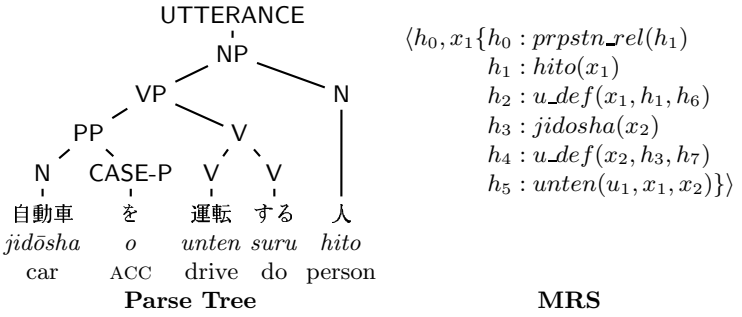

Fig. 2. Parse Tree and Simplified MRS for ドライバー$_2$ *doraibā* "driver"

syntacto-semantic analysis provided by the HPSG grammar, along with an annotator's choice of the most appropriate parse. From this record, all kinds of information can be extracted at various levels of granularity. In particular, traditional syntactic structure (e.g. in the form of labeled trees), dependency relations between words and full meaning representations using minimal recursion semantics (MRS: Copestake et al. (1999)). An example of the labeled tree and MRS views for ドライバー$_2$ *doraibā* "driver" is given in Figure 2.

Another important reason was the availability of a reasonably robust existing HPSG grammar of Japanese (JACY), and a wide range of open source tools for developing the grammars. We made extensive use of the LKB (Copestake, 2002), a grammar development environment, in order to extend JACY to the domain of defining sentences. We also used the extremely efficient PET parser (Callmeier, 2000), which handles grammars developed using the LKB, to parse large test sets for regression testing, treebanking and finally knowledge acquisition. Most of our development was done within the [incr tsdb()] profiling environment (Oepen and Carroll, 2000). In addition to its well documented facilities for comparing different versions of a grammar (or the same grammar using different parsers), it has facilities for annotating treebanks, updating them and training stochastic models using them. These models can then be used by PET to selectively rank the parser output.

3.1 Creating and Maintaining the Treebank

The construction of the treebank is a two stage process. First, the corpus is parsed (in our case using JACY with the PET parser), and then the annotator selects the correct analysis (or occasionally rejects all analyses). Selection is done through a choice of discriminants. The system selects features that distinguish between different parses, and the annotator selects or rejects the features until only one parse is left. The number of decisions for each sentence is proportional to $\log_2$ of the number of parses, although sometimes a single decision can reduce the number of remaining parses by more or less than half. In general, even a sentence with 5,000 parses only requires around 12 decisions.

Because the disambiguating choices made by the annotators are saved, it is possible to update the treebank when the grammar changes (Oepen et al., 2004). Although the trees depend on the grammar, re-annotation is only necessary in cases where either the parse has become more ambiguous, so new decisions have to be made, or existing rules or lexical items have changed so much that the system cannot reconstruct the parse.

One concern that has been raised with Redwoods style treebanking is the fact that the treebank is tied to a particular implementation of a grammar. The ability to update the treebank alleviates this concern to a large extent. A more serious concern is that it is only possible to annotate those trees that the grammar can parse. Sentences for which no analysis had been implemented in the grammar, or which fail to parse due to processing constraints are left unannotated. This makes grammar coverage an urgent issue. In the next section we discuss how we extended the grammar coverage in order to build the treebank.

3.2 Extending the Grammar

Testing JACY on the full set of 81,000 defining sentences from Lexeed gave a coverage of 39.3%, using the inbuilt unknown word mechanism. This was trivially extended to 46.2% by adding some orthographic variants.

We decided to test JACY's usability by attempting to extend its coverage on the Lexeed defining sentences to over 80% in 4 weeks. Six people were involved in this task; none of whom were involved in the original JACY development. Three of the six had little experience with HPSG.

We expected dictionary definitions to be a relatively easy domain. Barnbrook (2002, p87) showed that for English defining sentences, some eight sentence types covered over 92% of all entries. Japanese defining sentences showed similar regularity. In addition, there is little reference to outside context, and Lexeed has a fixed defining vocabulary.

Because we also wanted to experiment on treebanking in the same time-frame, we restricted ourselves to considering only the first defining sentence for each sense of all words with a familiarity greater than or equal to 6.0. This came to some 10,000 sentences, with an average length of 10.1 words/sentence. Finally, because we wanted to enter full syntactic information for all of the words in Lexeed, we switched off the unknown word processing. This gave us an initial coverage of around 10%.

We were able to bring the coverage to over 80% within the four weeks. The results are shown in Figure 3 which shows both the increase in coverage and change in the number of parser analyses.

The first big increase in coverage (to 55%) came from automatically expanding the lexicon. Tuning the lexicon and rules led to some incremental gains, mainly from relaxing the constraints on some existing rules. We also added some new rules,[5] for example a rule to parse compound verbs. This bought us to over

[5] We benefited greatly from some advice from the JACY developers.

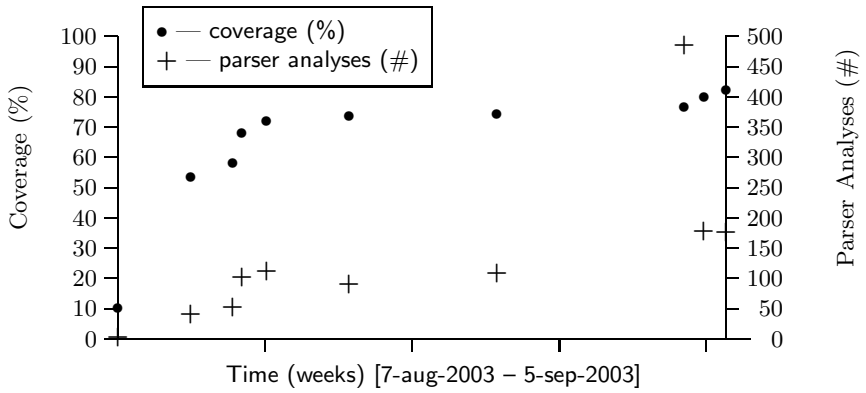

Fig. 3. Evolution of coverage

70%, at which point we started treebanking. Up to this point, none of the rules we added had been specific to the definition domain.

After two weeks treebanking, we made several small improvements and added a new domain-specific rule for definitions such as ***driver**: In golf, a long distance club.*. In this case the phrase *in golf* does not modify anything internal to the definition, but is effectively external to it. To handle this, we added a construction which effectively adds a constructionally defined predicate above a noun phrase, if and only if there is an extra adverbial to modify the phrase and the noun phrase is the highest constituent (the root): ***driver**: In golf, [driver means] a club for ....* Although the implementation of the construction was specific to Japanese, the idea is not at all language specific, and the resulting semantic representation is language neutral.

Keeping the number of analyses as low as possible is very important from the point of view of building the treebank. All extra ambiguity means more work in selecting the best parse. However, as coverage increases, real ambiguity also increased unavoidably. Occasionally, a rule would cause massive spurious ambiguity. In our first attempt to allow adverbial modification of root noun phrase fragments, we allowed adverbial modification of any noun phrase fragment, which sent the ambiguity skyrocketing to around 500 parses per sentence. The final ambiguity at the end of the four weeks was around 180 parses/sentence. This means that on average each tree requires 7-8 decisions to disambiguate it fully.

3.3 Current Status

We have now treebanked all of the 10,000 sentences with familiarity ≥ 6 which could be parsed (8,000), and over 15,000 of the second and subsequent sentences. Of these sentences, 95% were able to be resolved to one correct parse. Around 5% had no correct parse, mainly due to two errors — one in the construction

of the semantic representation for determiners and one in the way coordinate constructions are constructed. The annotator could not settle on a single correct parse for fewer than 1% of the sentences.

All the words in Lexeed's basic vocabulary have been entered in JACY. The current vocabulary size is around 32,000 words. We are now working with the main developers to reduce the average ambiguity and increase the coverage. The handling of numeral classifiers has also been improved (Bender and Siegel, 2004). The extended JACY grammar is available for download from `www.dfki.uni-sb.de/~siegel/grammar-download/JACY-grammar.html`.

4 Applications

The treebanked data and grammar have been tested in two ways. The first is to train a stochastic model for parse selection. The second is to use the parsed data to extract a thesaurus.

4.1 Stochastic Parse Ranking

Using the treebanked data, we built a stochastic parse ranking model with [incr tsdb()]. The ranker uses a maximum entropy learner to train a PCFG over the parse derivation trees, with the current node as a conditioning feature. The correct parse is selected 61.7% of the time (training on 4,000 sentences and testing on another 1,000; evaluated per sentence). More feature-rich models using parent and grandparent nodes along with semantic features have been proposed and implemented with an English grammar and the Redwoods treebank (Oepen et al., 2002). We intend to include such features, as well as to add our own extensions to train on constituent weight and semantic class.

4.2 Knowledge Acquisition

In addition to our work on the development of a Japanese language HPSG treebank, we are using the the corpus of dictionary definition sentences for knowledge acquisition. Past research in knowledge acquisition from definition sentences in Japanese has primarily dealt with the task of automatically generating hierarchy 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 hyperonym, hyponym, and synonym relationships; however, it lacked any concrete evaluation of the accuracy of the hierarchies created. More recently Tokunaga et al. (2001) created an ontology from a machine-readable dictionary and combined it with an existing thesaurus.

Our method differs from the aforementioned in two main respects: first, prior research has been limited to nouns, where our method handles all parts of speech; second, we are fully parsing the input, not just using regular expressions.

It is the use of full syntactic analysis with a well defined semantics (Minimal Recursion Semantics, Copestake et al. (1999)) that is the most important. There are three reasons. The first is that it makes our knowledge acquisition somewhat language independent: if we have a parser that can produce MRS, and a dictionary for that language, the algorithm can easily be ported. The second reason is that we can go on to use the same system to acquire knowledge from non-dictionary sources, which will not be as regular as dictionaries and thus harder to parse using only regular expressions. Third, we can more easily acquire knowledge beyond simple hypernyms, for example, identifying synonyms through common definition patterns as proposed by Tsuchiya et al. (2001).

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, 82% of the sentences can be parsed. In most cases, the word with the highest scope in the MRS representation will be the hypernym. For example, for *doraibā*$_1$ the hypernym is 道具 *tool* "dōgu" and for *doraibā*$_2$ the hypernym is 人 *hito* "person" (see Figure 1). Although the actual hypernym is in very different positions in the Japanese and English definition sentences, it takes the highest scope in both their semantic representations.

For some definition sentences (around 20%), further parsing of the semantic representation is necessary. For example, ドライバー$_3$ is defined as ***driver**: In golf, a long distance club*. In this case *in golf* has the highest scope: the hypernym is the complement of the empty copula. Again, this semantic representation is not language dependent, so we do not have to recreate the knowledge extraction system for new languages. Further, as we expand the scope of the knowledge acquisition the parsing can give us more information: for example that this sense of *driver* is used in the domain of golf.

We evaluate the extracted pairs by comparison with an existing thesaurus: the Goi-Taikei (Ikehara et al., 1997). Currently 58.5% of the pairs extracted for nouns are linked to nodes in the Goi-Taikei ontology (Bond et al., 2004). Some examples are given in Table 1. The remaining entries are words whose definition requires more parsing (15%) or those where one or both words could not be found in the Goi-Taikei.

Word	Hypernym	Word Node	Hypernym Node
ドライバー$_1$ "driver"	人"person"	worker	person
ドライバー$_2$	道具"equipment"	tool	equipment
ドライバー$_3$	クラブ"club"	leisure equipment	leisure equipment

Table 1. Sense Disambiguation using Hypernyms

In general, we are extracting pairs with more information than the Goi-Taikei hierarchy of 2,710 classes. 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 *percussion instrument* "drum" and 打楽器 *dagakki* "percussion instrument", which we can now distinguish.

5 Conclusion and Further Work

In this paper we have explained the motivation for the construction of a new lexical resource: the Hinoki treebank, and described its initial construction. We have further showed how it can be used to develop a language independent system for acquiring thesauruses from machine-readable dictionaries.

The first step in our path toward developing a system capable of fully understanding Japanese is to treebank all the defining sentences in Lexeed. This means that we must improve the coverage of the grammar, so that we can parse all sentences. When we have completed this task we will retrain our statistical model and use the new grammar to relearn the hypernym relations with higher precision.

In phase two we will add the knowledge of hypernyms into the stochastic model, and look at learning other information from the parsed defining sentences — in particular syntactic lexical-types, semantic association scores, meronyms, synonyms and antonyms.

In phase three, we will use the acquisition models learned in phase two, to extend our model to words not in Lexeed, using definition sentences from machine readable dictionaries or where they appear within normal text. In this way, we can grow an extensible lexicon and thesaurus from Lexeed.

Acknowledgements We would like to thank Colin Bannard, Timothy Baldwin, Emily Bender, Ulrich Callmeier, Ann Copestake, Dan Flickinger, Stephan Oepen, Yuji Matsumoto and especially Melanie Siegel for their support and encouragement.

References

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

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

Emily M. Bender and Melanie Siegel. Implementing the syntax of Japanese numeral classifiers. In *Proceedings of the IJC-NLP-2004*. Springer-Verlag, 2004. (this volume).

Francis Bond, Eric Nichols, Sanae Fujita, and Takaaki Tanaka. Acquiring an ontology for a fundamental vocabulary. In *COLING 2004*, Geneva, 2004. (to appear).

Ulrich Callmeier. PET - a platform for experimentation with efficient HPSG processing techniques. *Natural Language Engineering*, 6(1):99–108, 2000.

Ann Copestake. *Implementing Typed Feature Structure Grammars*. CSLI Publications, 2002.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. Minimal recursion semantics: An introduction. (manuscript `http://www-csli.stanford.edu/~aac/papers/newmrs.ps`), 1999.

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/CDROM.

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 *Third Machine Translation Summit: MT Summit III*, pages 101–106, Washington DC, 1991. (`http://xxx.lanl.gov/abs/cmp-lg/9510008`).

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).

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*, pages 423–430, 2003. URL `http://www.aclweb.org/anthology/P03-1054.pdf`.

Kavi Mahesh, Sergei Nirenburg, Stephen Beale, Evelyne Viegas, Victor Raskin, and Boyan Onyshkevych. Word sense disambiguation: Why statistics when you have these numbers? In *Seventh International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-97*, pages 151–159, Santa Fe, 1997.

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

Stephan Oepen, Dan Flickinger, and Francis Bond. Towards holistic grammar engineering and testing — grafting treebank maintenance into the grammar revision cycle. In *Beyond Shallow Analyses — Formalisms and Satitistical Modelling for Deep Analysis (Workshop at IJCNLP-2004)*, Hainan Island, 2004. (`http://www-tsujii.is.s.u-tokyo.ac.jp/bsa/`).

Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christoper D. Manning. LinGO redwoods: A rich and dynamic treebank for HPSG. In *Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT2002)*, Sozopol, Bulgaria, 2002.

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

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.

Mark Stevenson. *Word Sense Disambiguation*. CSLI Publications, 2003.

Takenobu Tokunaga, Yasuhiro Syotu, Hozumi Tanaka, and Kiyoaki Shirai. 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, 2001.

Kristina Toutanova, Christoper D. Manning, and Stephan Oepen. Parse ranking for a rich HPSG grammar. In *Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT2002)*, Sozopol, Bulgaria, 2002.

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

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