The Hinoki Sensebank
— A Large-Scale Word Sense Tagged Corpus of Japanese —

Takaaki Tanaka, Francis Bond and Sanae Fujita
{takaaki,bond,fujita}@cslab.kecl.ntt.co.jp
NTT Communication Science Laboratories,
Nippon Telegraph and Telephone Corporation

Abstract

Semantic information is important for precise word sense disambiguation system and the kind of semantic analysis used in sophisticated natural language processing such as machine translation, question answering, etc. There are at least two kinds of semantic information: lexical semantics for words and phrases and structural semantics for phrases and sentences.

We have built a Japanese corpus of over three million words with both lexical and structural semantic information. In this paper, we focus on our method of annotating the lexical semantics, that is building a word sense tagged corpus and its properties.

1 Introduction

While there has been considerable research on both structural annotation (such as the Penn Treebank (Taylor et al., 2003) or the Kyoto Corpus (Kurohashi and Nagao, 2003)) and semantic annotation (e.g. Senseval: Kilgariff and Rosenzweig, 2000; Shirai, 2002), there are almost no corpora that combine both. This makes it difficult to carry out research on the interaction between syntax and semantics.

Projects such as the Penn Propbank are adding structural semantics (i.e. predicate argument structure) to syntactically annotated corpora, but not lexical semantic information (i.e. word senses). Other corpora, such as the English Redwoods Corpus (Oepen et al., 2002), combine both syntactic and structural semantics in a monostratal representation, but still have no lexical semantics.

In this paper we discuss the (lexical) semantic annotation for the Hinoki Corpus, which is part of a larger project in psycho-linguistic and computational linguistics ultimately aimed at language understanding (Bond et al., 2004).

2 Corpus Design

In this section we describe the overall design of the corpus, and is constituent corpora. The basic aim is to combine structural semantic and lexical semantic markup in a single corpus. In order to make the first phase self contained, we started with dictionary definition and example sentences. We are currently adding other genre, to make the langauge description more general, starting with newspaper text.

2.1 Lexeed: A Japanese Basic Lexicon

We use word sense definitions from Lexeed: A Japanese Semantic Lexicon (Kasahara et al., 2004). It was built in 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 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 to use only the 28,000 familiar words. 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. This figure includes the sense annotation and information derived from it that is described in this paper.

Table 1 shows the relation between polysemy and familiarity. The #WS column indicates the average number of word senses that polysemous

Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006, pages 62–69,
Sydney, July 2006.

INDEX	ドライバー *doraiba-*	
POS	noun	Lexical-Type `noun-lex`
FAMILIARITY	6.5 [1–7] ($\geq$ 5)	Frequency 37 Entropy 0.79
SENSE 1 (0.11)	DEFINITION	ねじ$_1$/を/差し入れ$_1$/たり/ 、/抜き取っ$_1$/たり/する/道具$_1$/。 a tool for inserting and removing screws .
	EXAMPLE	彼 は 細い ドライバー で 眼鏡 の ねじ を 締め た 。 he used a small screwdriver to tighten the screws on his glasses.
	HYPERNYM	道具$_1$ *equipment* "tool"
	SEM. CLASS	⟨`942:tool/implement`⟩ (⊂ ⟨`893:equipment`⟩)
	WORDNET	*screwdriver*$_1$
SENSE 2 (0.84)	DEFINITION	自動車$_1$/を/運転$_1$/する/人$_1$/。 Someone who drives a car.
	EXAMPLE	父 は 優良 な ドライバー として 表彰 さ れ た 。 my father was given an award as a good driver.
	HYPERNYM	人$_1$ *hito* "person"
	SEM. CLASS	⟨`292:chauffeur/driver`⟩ (⊂ ⟨`5:person`⟩)
	WORDNET	*driver*$_1$
SENSE 3 (0.05)	DEFINITION	ゴルフ$_1$/で/ 、/遠距離 $_1$/用/の/クラブ$_3$/。 一番/ウッド/。 In golf, a long-distance club. A number one wood.
	EXAMPLE	彼 は ドライバー で ３ ０ ０ ヤード 飛ばし た 。 he hit (it) 30 yards with the driver.
	HYPERNYM	クラブ$_3$ *kurabu* "club"
	SEM. CLASS	⟨`921:leisure equipment`⟩ (⊂ 921)
	WORDNET	*driver*$_5$
	DOMAIN	ゴルフ$_1$ *gorufu* "golf"

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

words have. Lower familiarity words tend to have less ambiguity and 70 % of words with a familiarity of less than 5.5 are monosemous. Most polysemous words have only two or three senses as seen in Table 2.

Fam	#Words	Poly-semous	#WS	#Mono-semous(%)
6.5 -	368	182	4.0	186 (50.5)
6.0 -	4,445	1,902	3.4	2,543 (57.2)
5.5 -	9,814	3,502	2.7	6,312 (64.3)
5.0 -	11,430	3,457	2.5	7,973 (69.8)

Table 1: Familiarity vs Word Sense Ambiguity

2.2 Ontology

We also have an ontology built from the parse results of definitions in Lexeed (Nichols and Bond, 2005). The ontology includes more than 50 thousand relationship between word senses, e.g. synonym, hypernym, abbreviation, etc.

2.3 Goi-Taikei

As part of the ontology verification, all nominal and most verbal word senses in Lexeed were linked to semantic classes in the Japanese thesaurus, Nihongo Goi-Taikei (Ikehara et al., 1997). Common nouns are classified into about 2,700 semantic classes which are organized into a

#WS	#Words
1	18460
2	6212
3	2040
4	799
5	311
6	187
7	99
8	53
9	35
10	15
11	19
12	13
13	13
14	6
15	6
16	3
17	2
18	3
19	1
20	2
$\geq$ 21	19

Table 2: Number of Word Senses

semantic hierarchy.

2.4 Hinoki Treebank

Lexeed definition and example sentences are syntactically and semantically parsed with HPSG and correct results are manually selected (Tanaka et al., 2005). The grammatical coverage over all sentences is 86%. Around 12% of the parsed sentences were rejected by the treebankers due to an incomplete semantic representation. This process had been done independently of word sense annotation.

2.5 Target Corpora

We chose two types of corpus to mark up: a dictionary and two newspapers. Table 3 shows basic statistics of the target corpora.

The dictionary Lexeed, which defined word senses, is also used for a target for sense tagging. Its definition (LXD-DEF) and example (LXD-EX) 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 one example sentence.

Both newspaper corpora where taken from the Mainichi Daily News. One sample (Senseval2) was the text used for the Japanese dictionary task in Senseval-2 (Shirai, 2002), which has some words marked up with word sense tags defined in the Iwanami lexicon (Nishio et al., 1994). The second sample was those sentences used in the Kyoto Corpus (Kyoto), which is marked up with dependency analyses (Kurohashi and Nagao, 2003). We chose these corpora so that we can compare our annotation with existing annotation. Both these corpora were thus already segmented and annotated with parts-of-speech. However, they used different morphological analyzers to the one used in Lexeed, so we had to do some remapping. E.g. in Kyoto the copula is not split from nominal-adjectives, whereas in Lexeed it is: 元気だ *genkida* "lively" vs 元気 だ *genki da*. This could be done automatically after we had written a few rules.

Although the newspapers contain many words other than basic words, only basic words have sense tags. Also, a word unit in the newspapers does not necessarily coincide with the headword in Lexeed since part-of-speech taggers used for annotation are different. We do not adjust the word segmentation and leave it untagged at this stage, even if it is a part of a basic word or consists of multiple basic words. For instance, Lexeed has the compound entry 貨幣価値 *kahei-kachi* "monetary value", however, this word is split into two basic words in the corpora. In this case, both two words 貨幣 *kahei* "money" and 価値 *kachi* "value" are tagged individually.

Corpus	Tokens	Content Words	Basic Words	%Mono-semous
LXD-DEF	691,072	318,181	318,181	31.7
LXD-EX	498,977	221,224	221,224	30.5
Senseval2	888,000	692,069	391,010	39.3
Kyoto	969,558	526,760	472,419	36.3

Table 3: Corpus Statistics

The corpora are not fully balanced, but allow some interesting comparisons. There are effectively three genres: dictionary definitions, which tend to be fragments and are often syntactically highly ambiguous; dictionary example sentences, which tend to be short complete sentences, and are easy to parse; and newspaper text from two different years.

3 Annotation

Each word was annotated by five annotators. We actually used 15 annotators, divided into 3 groups. None were professional linguists or lexicographers. All of them had a score above 60 on a Chinese character based vocabulary test (Amano and Kondo, 1998). We used multiple annotators to measure the confidence of tags and the degree of difficulty in identifying senses.

The target words for sense annotation are the 9,835 headwords having multiple senses in Lexeed (§ 2.1). They have 28,300 senses in all. Monosemous words were not annotated. Annotation was done word by word. Annotators are presented multiple sentences (up to 50) that contain the same target word, and they keep tagging that word until occurrences are done. This enables them to compare various contexts where a target word appears and helps them to keep the annotation consistent.

3.1 Tool

A screen shot of the annotation tool is given in Figure 2. The interface uses frames on a browser, with all information stored in SQL tables. The left hand frame lists the words being annotated. Each word is shown with some context: the surrounding paragraph, and the headword for definition and example sentences. These can be clicked on to get more context. The word being annotated is highlighted in red. For each word, the annotator chooses its senses or one or more of the other tags as clickable buttons. It is also possible to choose one tag as the default for all entries on the screen.

The right hand side frame has the dictionary definitions for the word being tagged in the top frame, and a lower frame with instructions. A single word may be annotated with senses from more than one headword. For example バス is divided into two headwords *basu* "bus" and *basu* "bass", both of which are presented.

As we used a tab-capable browser, it was easy for the annotators to call up more information in different tabs. This proved to be quite popular.

3.2 Markup

Annotators choose the most suitable sense in the given context from the senses that the word have in lexicon. Preferably, they select a single sense for a word, although they can mark up multiple tags if the words have multiple meanings or are truly ambiguous in the contexts.

When they cannot choose a sense in some reasons, they choose one or more of the following special tags.

o *other sense*: an appropriate sense is not found in a lexicon. Relatively novel concepts (e.g. ドライバー *doraibā* "driver" for "software driver") are given this tag.

c *multiword expressions (compound / idiom)*: the target word is a part of a non-compositional compound or idiom.

p *proper noun*: the word is a proper noun.

x *homonym*: an appropriate entry is not found in a lexicon, because a target is different from head words in a lexicon (e.g. only a headword バス *bass* "bus" is present in a lexicon for バス*basu* "bass").

e *analysis error*: the word segmentation or part-of-speech is incorrect due to errors in pre-annotation of the corpus.

3.3 Feedback

One of the things that the annotators found hard was not knowing how well they were doing. As they were creating a gold standard, there was initially no way of knowing how correct they were. We also did not know at the start of the annotation how fast senses could or should be annotated (a test of the tool gave us an initial estimate of around 400 tokens/day).

To answer these questions, and to provide feedback for the annotators, twice a day we calculated and graphed the speed (in words/day) and majority agreement (how often an annotator agrees with the majority of annotators for each token, measured over all words annotated so far). Each annotator could see a graph with their results labelled, and the other annotators made anonymous. The results are grouped into three groups of five annotators. Each group is annotating a different set of words, but we included them all in the feedback. The order within each group is sorted by agreement, as we wished to emphasise the importance of agreement over speed. An example of a graph is given in Figure 3. When this feedback was given, this particular annotator has the second worst agreement score in their subgroup (90.27%) and is reasonably fast (1799 words/day) — they should slow down and think more.

The annotators welcomed this feedback, and complained when our script failed to produce it. There was an enormous variation in speed: the fastest annotator was 4 times as fast as the slowest, with no appreciable difference in agreement. After providing the feedback, the average speed increased considerably, as the slowest annotators agonized less over their decisions. The final average speed was around 1,500 tokens/day, with the fastest annotator still almost twice as fast as the slowest.

4 Inter-Annotator Agreement

We employ inter-annotator agreement as our core measure of annotation consistency, in the same way we did for treebank evaluation (Tanaka et al., 2005). This agreement is calculated as the average of pairwise agreement. Let w_i be a word in a set of content words W and $w_{i,j}$ be the jth occurrence of a word w_i. Average pairwise agreement between the sense tags of $w_{i,j}$ each pair of annotators marked up $a(w_{i,j})$ is:

$$a(w_{i,j}) = \frac{\sum_k ({}_{m_{i,j}(s_{ik})}\mathrm{C}_2)}{{}_{n_{w_{i,j}}}\mathrm{C}_2} \quad (1)$$

where $n_{w_{i,j}} (\geq 2)$ is the number of annotators that tag the word $w_{i,j}$, and $m_{i,j}(s_{ik})$ is the number

Figure 2: Sense Annotation tool (word 暫く *shibaraku* “briefly”)

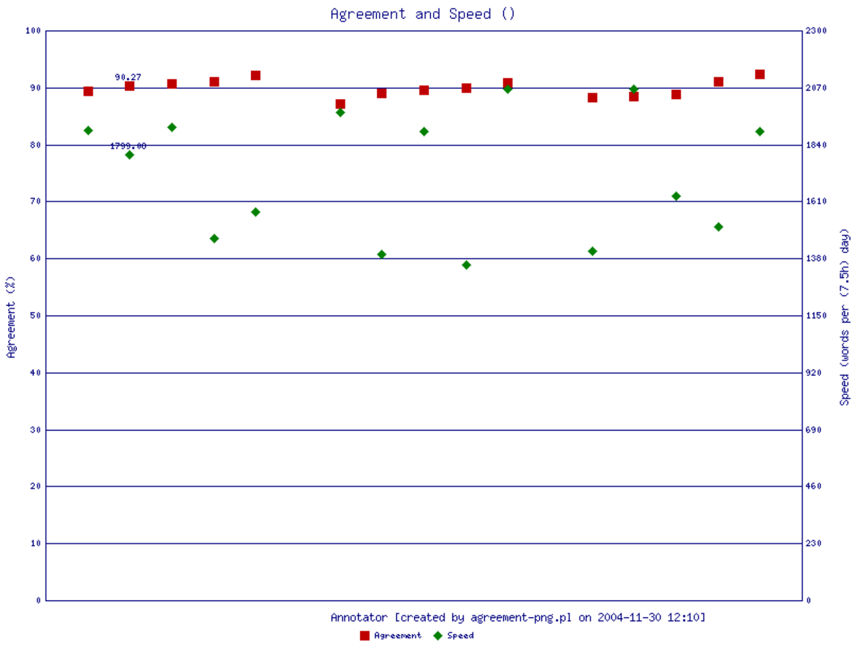

Figure 3: Sample feedback provided to an annotator

of sense tags s_{ik} for the word $w_{i,j}$. Hence, the agreement of the word w_i is the average of $a_{w_{i,j}}$ over all occurrences in a corpus:

$$a(w_i) = \frac{\sum_j a(w_{i,j})}{N_{w_i}} \tag{2}$$

where N_{w_i} is the frequency of the word w_i in a corpus.

Table 4 shows statistics about the annotation results. The average numbers of word senses in the newspapers are lower than the ones in the dictionary and, therefore, the token agreement of the newspapers is higher than those of the dictionary sentences. %Unanimous indicates the ratio of tokens vs types for which all annotators (normally five) choose the same sense. Snyder and Palmer (2004) report 62% of all word types on the English all-words task at SENSEVAL-3 were labelled unanimously. It is hard to directly compare with our task since their corpus has only 2,212 words tagged by two or three annotators.

4.1 Familiarity

As seen in Table 5, the agreement per type does not vary much by familiarity. This was an unexpected result. Even though the average polysemy is high, there are still many highly familiar words with very good agreement.

Fam	Agreement token (type)	#WS	%Monosem
6.5 -	.723 (.846)	7.00	22.6
6.0 -	.780 (.846)	5.82	28.0
5.5 -	.813 (.853)	3.79	42.4
5.0 -	.821 (.850)	3.84	46.2
ALL	.787 (.850)	5.18	34.5

Table 5: Inter-Annotator Agreement (LXD-DEF)

4.2 Part-of-Speech

Table 6 shows the agreement according to part of speech. Nouns and verbal nouns (vn) have the highest agreements, similar to the results for the English all-words task at SENSEVAL-3 (Snyder and Palmer, 2004). In contrast, adjectives have as low agreement as verbs, although the agreement of adjectives was the highest and that of verbs was the lowest in English. This partly reflects differences in the part of speech divisions between Japanese and English. Adjectives in Japanese are much close in behaviour to verbs (e.g. they can head sentences) and includes many words that are translated as verbs in English.

4.3 Entropy

Entropy is directly related to the difficulty in identifing senses as shown in Table 7.

POS	Agreement (type)	#WS	%Monosemous
n	.803 (.851)	2.86	62.9
v	.772 (.844)	3.65	34.0
vn	.849 (.865)	2.54	61.0
adj	.770 (.810)	3.58	48.3
adv	.648 (.833)	3.08	46.4
others	.615 (.789)	3.19	50.8

Table 6: POS vs Inter-Annotator Agreement (LXD-DEF)

Entropy	Agreement (type)	#Words	#WS
2 -	.672	84	14.2
1 -	.758	1096	4.38
0.5 -	.809	1627	2.88
0.05 -	.891	495	3.19
0 -	.890	13778	2.56

Table 7: Entropy vs Agreement

4.4 Sense Lumping

Low agreement words have some senses that are difficult to distinguish from each other: these senses often have the same hypernyms. For example, the agreement rate of 草花 *kusabana* "grass/flower" in LXD-DEF is only 33.7 %. It has three senses whose semantic class is similar: *kusabana*$_1$ "flower that blooms in grass", *kusabana*$_2$ "grass that has flowers" and *souka*$_1$ "grass and flowers" (hypernyms flower$_1$, grass$_1$ and flower$_1$ & grass$_1$ respectively).

In order to investigate the effect of semantic similarity on agreement, we lumped similar word senses based on hypernym and semantic class. We use hypernyms from the ontology (§ 2.1) and semantic classes in Goi-Taikei (§ 2.3), to regard the word senses that have the same hypernyms or belong to the same semantic classes as the same senses.

Table 8 shows the distribution after sense lumping. Table 9 shows the agreement with lumped senses. Note that this was done with an automatically derived ontology that has not been fully hand corrected.

As is expected, the overall agreement increased, from 0.787 to 0.829 using the ontology, and to 0.835 using the coarse-grained Goi-Taikei semantic classes. For many applications, we expect that this level of disambiguation is all that is required.

4.5 Special Tags

Table 10 shows the ratio of special tags and multiple tags to all tags. These results show

Corpus	Annotated Tokens	#WS	Agreement token (type)	%Unanimous token (type)	Kappa
LXD-DEF	199,268	5.18	.787 (.850)	62.8 (41.1)	0.58
LXD-EX	126,966	5.00	.820 (.871)	69.1 (53.2)	0.65
Senseval2	223,983	4.07	.832 (.833)	73.9 (45.8)	0.52
Kyoto	268,597	3.93	.833 (.828)	71.5 (46.1)	0.50

Table 4: Basic Annotation Statistics

Corpus	%Other Sense	%MWE	%Homonym	%Proper Noun	%Error	%Multiple Tags
LXD-DEF	4.2	1.5	0.084	0.046	0.92	11.9
LXD-EX	2.3	0.44	0.035	0.0018	0.43	11.6
Senseval2	9.3	5.6	4.1	8.7	5.7	7.9
Kyoto	9.8	7.9	3.3	9.0	5.5	9.3

Table 10: Special Tags and Multiple Tags

Fam	Agreement token (type)	#WS	%Monosem
6.5 -	.772 (.863)	6.37	25.6
6.0 -	.830 (.868)	5.16	31.5
5.5 -	.836 (.872)	3.50	45.6
5.0 -	.863 (.866)	3.76	58.7
ALL	.829 (.869)	4.72	39.1

Lumping together Hypernyms
(4,380 senses compressed into 1,900 senses)

Fam	Agreement token (type)	#WS	%Monosem
6.5 -	.775 (.890)	6.05	26.8
6.0 -	.835 (.891)	4.94	36.4
5.5 -	.855 (.894)	3.29	50.6
5.0 -	.852 (.888)	3.46	49.9
ALL	.835 (.891)	4.48	41.7

Lumping together Semantic Classes
(8,691 senses compressed into 4,030 senses)

Table 8: Sense Lumping Results (LXD-DEF)

(LXD-DEF)	Agreement token (type)	#WS	%Monosem
no lumping	.698 (.816)	8.81	0.0
lumping	.811 (.910)	8.24	20.0

Hypernum Lumping

(LXD-DEF)	Agreement token (type)	#WS	%Monosem
no lumping	.751 (.814)	7.09	0.0
lumping	.840 (.925)	5.99	21.9

Semantic Class Lumping

Table 9: Lumped Sense Agreement (LXD-DEF)

the differences in corpus characteristics between dictionary and newspaper. The higher ratios of Other Sense and Homonym at newspapers indicate that the words whose surface form is in a dictionary are frequently used for the different meanings in real text, e.g. 銀 *gin* “silver” is used for the abbrebiation of 銀行 *ginkou* “bank”. %Multiple Tags is the percentage of tokens for which at least one annotator marked multiple tags.

5 Discussion

5.1 Comparison with Senseval-2 corpus

The Senseval-2 Japanese dictionary task annotation used senses from a different dictionary (Shirai, 2002). In the evaluation, 100 test words were selected from three groups with different entropy bands (Kurohashi and Shirai, 2001). D_a is the highest entropy group, which contains the most hard to tag words, and D_c is the lowest entropy group.

We compare our results with theirs in Table 11. The Senseval-2 agreement figures are slightly higher than our overall. However, it is impossible to make a direct comparison as the numbers of annotators (two or three annotators in Senseval vs more than 5 annotators in our work) and the sense inventories are different.

5.2 Problems

Two main problems came up when building the corpora: word segmentation and sense segmentation. Multiword expressions like compounds and idioms are tied closely to both problems.

The word segmentation is the problem of how to determine an unit expressing a meaning. At the present stage, it is based on headword in Lexeed, in particular, only compounds in Lexeed are recognized, we do not discriminate non-decomposable compounds with decomposable ones. However, if the headword unit in the dictionary is inconsistent, word sense tagging inherits this problem. For examples, 一部 *ichibu* has two main usage: one + classifier and a part of something. Lexeed has an entry including both two senses. However, the former is split into two

POS	D_a		D_b		D_c		Total	
	Hinoki	Senseval	Hinoki	Senseval	Hinoki	Senseval	Hinoki	Senseval
noun	.768	.809	.784	.786	.848	.957	.806	.859
	14.4	13.1	5.0	4.1	3.1	3.8	5.9	5.1
verb	.660	.699	.722	.896	.738	.867	.723	.867
	16.7	21.8	10.3	9.3	5.2	5.9	9.6	10.9
total	.710	.754	.760	.841	.831	.939	.768	.863
	15.6	18.8	7.0	6.2	4.2	4.9	7.6	7.9

Table 11: Comparison of Agreement for the Senseval-2 Lexical Sample Task Corpus (upper row: agreement, lower row: the number of word senses)

words by our morphological analyser in the same way as other numeral + classifier.

The second problem is how to mark off metaphorical meaning from literal meanings. Currently, this also depends on the Lexeed definition and it is not necessarily consistent either. Some words in institutional idioms (Sag et al., 2002) have the idiom sense in the lexicon while most words do not. For instance, 尻尾 *shippo* “tail of animal”) has a sense for the reading “weak point” in an idiom 尻尾を掴む *shippo-o tsukamu* “lit. to grasp the tail, idiom. to find one’s weak point”, while 汗 *ase* “sweat” does not have a sense for the applicable meaning in the idiom 汗を流す *ase-o nagasu* “lit. to sweat, idiom, to work hard”.

6 Conclusions

We built a corpus of over three million words which has lexical semantic information. We are currently using it to build a model for word sense disambiguation.

Acknowledgement

We thank the 15 sense annotators and 3 treebankers (Takayuki Kuribayashi, Tomoko Hirata and Koji Yamashita).

References

Anne Abeillé, editor. 2003. *Treebanks: Building and Using Parsed Corpora*. Kluwer Academic Publishers.

Shigeaki Amano and Tadahisa Kondo. 1998. Estimation of mental lexicon size with word familiarity database. In *International Conference on Spoken Language Processing*, volume 5, pages 2119–2122.

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

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)*, pages 554–559. Hainan Island.

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, 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).

Adam Kilgariff and Joseph Rosenzweig. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities*, 34(1–2):15–48. Special Issue on SENSEVAL.

Sadao Kurohashi and Makoto Nagao. 2003. Building a Japanese parsed corpus — while improving the parsing system. In Abeillé (2003), chapter 14, pages 249–260.

Sadao Kurohashi and Kiyoaki Shirai. 2001. SENSEVAL-2 Japanese task. SIG NLC 2001-10, IEICE. (in Japanese).

Eric Nichols and Francis Bond. 2005. Acquiring ontologies using deep and shallow processing. In *11th Annual Meeting of the Association for Natural Language Processing*, pages 494–498. Takamatsu.

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

Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christoper D. Manning, Dan Flickinger, and Thorsten Brant. 2002. The LinGO redwoods treebank: Motivation and preliminary applications. In *19th International Conference on Computational Linguistics: COLING-2002*, pages 1253–7. Taipei, Taiwan.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbuk, editor, *Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing-2002*, pages 1–15. Springer-Verlag, Hiedelberg/Berlin.

Kiyoaki Shirai. 2002. Construction of a word sense tagged corpus for SENSEVAL-2 Japanese dictionary task. In *Third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 605–608.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of Senseval-3*, pages 41–44. ACL, Barcelona.

Takaaki Tanaka, Francis Bond, Stephan Oepen, and Sanae Fujita. 2005. High precision treebanking – blazing useful trees using POS information. In *ACL-2005*, pages 330–337.

Ann Taylor, Mitchel Marcus, and Beatrice Santorini. 2003. The Penn treebank: an overview. In Abeillé (2003), chapter 1, pages 5–22.