Lang Resources & Evaluation (2008) 42:243–251
DOI 10.1007/s10579-008-9062-z

ERRATUM

The Hinoki syntactic and semantic treebank of Japanese

Francis Bond · Sanae Fujita · Takaaki Tanaka

Published online: 22 February 2008

Abstract In this paper we describe the current state of a new Japanese lexical resource: the Hinoki treebank. The treebank is built from dictionary definitions, examples and news text, and uses an HPSG based Japanese grammar to encode both syntactic and semantic information. It is combined with an ontology based on the definition sentences to give a detailed sense level description of the most familiar 28,000 words of Japanese.

Keywords Japanese · Treebank · Sensebank · HPSG · Ontology

1 Introduction

In this paper we describe the current state of the Hinoki project (Bond et al. 2004a; Tanaka et al. 2006), an empirical investigation into the structure and meaning of

Erratum to: Lang Resources & Evaluation DOI 10.1007/s10579-007-9036-6.

The online version of the original article can be found under doi: 10.1007/s10579-007-9036-6.

F. Bond · S. Fujita · T. Tanaka
NTT Communication Science Laboratories, Nippon Telegraph
and Telephone Corporation, Atsugi-shi, Kyoto, Japan

S. Fujita
e-mail: sanae@cslab.kecl.ntt.co.jp

T. Tanaka
Research and Development Center, Nippon Telegraph
and Telephone West Corporation, Osaka, Japan
e-mail: taka.tanaka@rdc.west.ntt.co.jp

F. Bond (✉)
Computational Linguistics Group, NICT, Kyoto 619-0225, Japan
e-mail: bond@ieee.org

Japanese. We have tagged a treebank and sensebank over a corpus of over a million words, and used them to refine a grammar and ontology. We are now extending the corpus to different genre and training NLP systems using the corpus. The ultimate goal of our research is natural language understanding—we aim to take text and parse it into a useful semantic representation.

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. 2005). Statistical techniques have also been shown to be useful for word sense disambiguation (Stevenson 2003). However, to date, there have been almost no combinations of lexical semantic (word 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 do not provide much traction, 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, although, the superordinate concepts capital (⊃ *stocks*) and move upward (⊃ *sky rocket*) frequently do appear together. This lack should motivate the use of similarity and/or class based approaches but there has been little success in this area to date.

We hypothesize that there are two major reasons for the lack of progress. The first reason is that there are few resources that combine syntactic and semantic annotation, including both structural semantics (predicate-argument structure) and lexical semantics (word senses), 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 necessary or how to obtain it. For example, classes from both WordNet and Goi-Taikei have been shown to be useful in a variety of tasks, but their granularity is very different, and it is an open question as to how finely senses need to be divided.

Our solution to these problems has three phases. In the first phase, we built a treebank based on the Japanese semantic database Lexeed (Kasahara et al. 2004) and constructed a thesaurus from it (Bond et al. 2004b). In the second phase, we have tagged the definition sentences with senses (Tanaka et al. 2006) and are using the lexical semantic information and the thesaurus to build a model that combines syntactic and semantic information. In phase three, we will look at ways of combining the lexical and structural semantics and extending our lexicon and ontology to less familiar words.

We are now finishing phase two: each definition and example sentence has been parsed, and the most appropriate analysis selected. Each content word in the sentences has been marked with the appropriate Lexeed sense. The syntactic model is embodied in a grammar, while the semantic model is linked by an ontology. We are now testing the use of similarity and/or semantic class based back-offs for parsing and generation with both symbolic grammars and statistical models (Fujita et al. 2007; Tanaka et al. 2007).

2 The Lexeed semantic database of Japanese

The Lexeed semantic database of Japanese consists of all Japanese words with a familiarity greater than or equal to five on a seven point scale (Kasahara et al. 2004), henceforth *basic words*. This gives 28,000 words in all, with 46,000 different senses. Definition sentences for these sentences were rewritten to use only the 28,000 familiar words (and some function words). The defining vocabulary is only 16,900 different words (60% of the entire vocabulary). A simplified example entry for the word ドライバー *doraibā* "driver" is given in Fig. 1, with English glosses. Lexeed itself consists of just the definitions, familiarity and part of speech, all underlined features are added by the Hinoki project.

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.

2.1 Target corpora

We chose two types of corpus to mark up: a dictionary and two sets of newspaper text. Table 1 shows the basic statistics of the target corpora.

Lexeed's 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. The sentences are short, around 10 words on average and relatively self-contained. The example sentences (LXD−EX) are relatively easy to

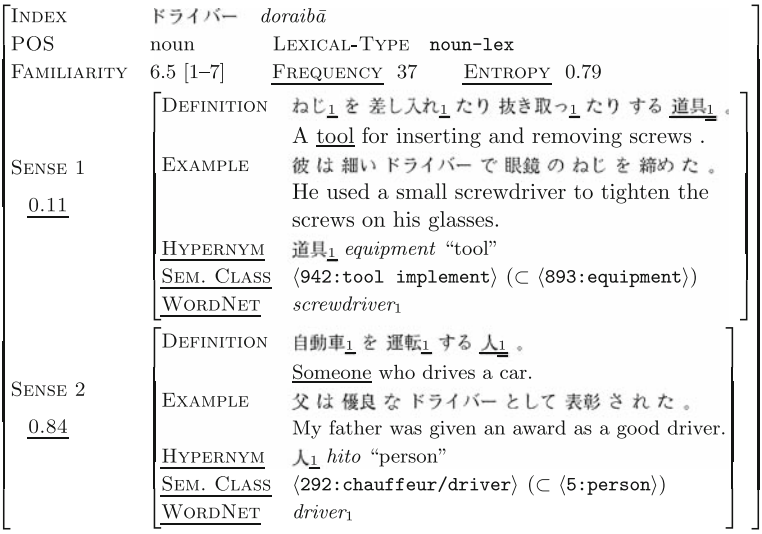

Fig. 1 First two senses for the word ドライバー *doraibā* "driver"

Table 1 Corpus statistics

Corpus	Sentences	Words	Content words	Basic words	% Monosemous
LXD−DEF	75,000	691,072	318,181	318,181	31.7
LXD−EX	45,000	498,977	221,224	221,224	30.5
Senseval2	36,000	888,000	692,069	391,010	39.3
Kyoto	38,000	969,558	526,760	472,419	36.3

parse. The definition sentences (LXD−DEF) contain many coordinate structures and are relatively hard to parse.

Both newspaper corpora were 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 the Senseval sense annotation). 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 already segmented and part-of-speech annotated.

This collection of corpora is not fully balanced, but allows 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. Tagging multiple genres allows us to measure the portability of our NLP tools and models across different text types.

3 The Hinoki treebank

The basic approach to the syntactic annotation is *grammar based corpus annotation*. 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 (following Oepen et al. 2004). The system selects features that distinguish between different parses, and the annotator selects or rejects the features until only one parse is left. The average number of decisions for each sentence is proportional to its length (around $\log_2$ of the number of parses). In general, even a sentence with 5,000 parses requires around 12 decisions (Tanaka et al. 2005).

We use a Japanese grammar (JACY) based on a monostratal theory of grammar (Head Driven Phrase Structure Grammar: HPSG, Pollard and Sag 1994), so that we can simultaneously annotate syntactic and structural semantic structure without overburdening the annotator. The native HPSG representation is a *sign* that integrates various levels of representation—syntactic, semantic, pragmatic and more—all accessible in the same structure. The JACY grammar is an HPSG-based grammar of Japanese (Siegel 2000). We extended JACY by manually adding the Lexeed defining vocabulary, and some new rules and lexical-types (Bond et al. 2004a).

The treebank records the complete syntacto-semantic analysis provided by the HPSG grammar, along with an annotator's choice of the most appropriate parse.

$$\langle h_0, x_1 \{ h_0 : proposition\_m(h_1) \\ h_1 : hito\_n(x_1) \text{“person”}, \quad h_2 : udef\_q(x_1, h_1, h_6), \\ h_3 : jidosha\_n(x_2) \text{“car”}, \quad h_4 : udef\_q(x_2, h_3, h_7), \\ h_5 : unten\_s(e_1, x_1, x_2) \text{“drive”} \} \rangle$$

Fig. 2 MRS view of 自動車 を 運転 する 人。“A person who drives a car”

From this record, all kinds of information can be extracted at various levels of granularity. For example, the semantics are stored in the sign in the form of Minimal Recursion Semantics (Copestake et al. 2005). A simplified example of this structural semantic representation (for the definition of ドライバー$_2$ *doraibā* “driver”) is given in Fig. 2.

In the Hinoki annotation, we have deliberately chosen not to annotate sentences for which we do not have a complete analysis. This allows us to immediately identify where the grammar coverage is incomplete. If an application can use partial results, then the PET parser (Callmeier 2000) can still return the fragments of an incomplete analysis.

Because the disambiguating choices made by the annotators are recorded, it is possible to efficiently 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.

We had 5,000 sentences from the definition sentence corpus annotated by 3 speakers of Japanese with a high score in a Japanese proficiency test but no linguistic training (Tanaka et al. 2005). The average annotation speed was 50 sentences an hour.

We measured inter-annotator agreement as follows: the proportion of sentences for which two annotators selected the exact same parse (65.4%), the proportion for which both chose parses, but there was no agreement, 18.2% of sentences, the proportion for which both annotators found no suitable analysis, 12.4% of sentences. For 4.0% of sentences, one annotator found no suitable parses, but one selected one or more.

The grammatical coverage over all sentences in the dictionary domain (definitions and example sentences) is now 86%. Around 12% of sentences with a spanning parse were rejected by the treebankers, because the semantics were incorrect. We therefore have a complete analysis for 76% of the sentences. The total size of the treebank is currently 53,600 definition sentences and 36,000 example sentences: 89,600 sentences in total. We are currently parsing and annotating the newspaper text.

4 The Hinoki sensebank

In this section we discuss the (lexical) semantic annotation for the Hinoki project (Tanaka et al. 2006). Each word was annotated by five annotators (15 annotators, divided into 3 groups). They were all native speakers of Japanese with a high score in a Japanese proficiency test but no linguistic training. 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 basic words having multiple senses in Lexeed (Sect. 2). 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 keep the annotation consistent.

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. Annotators can also choose not to assign a sense for the following reasons: lexicon missing sense; non-compositional idiom sub part; proper name; analysis error.

An example of a sense-tagged sentence is given in (1). Each open class word has been tagged with its sense: the senses are shown disambiguated by their hypernyms in the gloss.

(1)	ゴルフ$_{\underline{1}}$	で	、	遠距離$_{\underline{1}}$	用	の	クラブ$_{\underline{3}}$
	gorufu	*de*	,	*choukyri*	*you*	*no*	*kurabu*
	golf$_{competition}$	in	,	long-distance$_{distance}$	for	of	club$_{group}$

“In golf, a club for long-distances”

We provided feedback for the annotators by twice a day calculating and graphing 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 own results labelled, and the other annotators made anonymous. This feedback was popular; after it was introduced 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 almost twice as fast as the slowest.

We employ average pair-wise inter-annotator agreement as our core measure of annotation consistency, in the same way as we did for treebank evaluation. Table 2 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.

Table 2 Basic annotation statistics

Corpus	Annotated tokens	#WS	Agreement token (type)	%Unanimous token (type)	Kappa
LXD−DEF	199,268	5.18	0.787 (0.850)	62.8 (41.1)	0.58
LXD−EX	126,966	5.00	0.820 (0.871)	69.1 (53.2)	0.65
Senseval2	223,983	4.07	0.832 (0.833)	73.9 (45.8)	0.52
Kyoto	268,597	3.93	0.833 (0.828)	71.5 (46.1)	0.50

Table 3 POS vs. inter-annotator agreement (LXD−DEF)

POS	n	vn	v	adj	adv	others
Agreement (Token)	0.803	0.849	0.772	0.770	0.648	0.615
Agreement (Type)	0.851	0.865	0.844	0.810	0.833	0.789
# Word senses	2.86	2.54	3.65	3.58	3.08	3.19
% Monosemous	62.9	61.0	34.0	48.3	46.4	50.8

%Unanimous indicates the ratio of tokens vs. types for which all annotators (normally five) chose 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.

Table 3 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, in Japanese, 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 include many words that are translated as verbs in English.

5 Hinoki ontology

We constructed an ontology from the parse results of definitions in Lexeed (Bond et al. 2004b). The ontology includes more than 50,000 relationships between word senses, e.g. synonym, hypernym, abbreviation, etc.

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 semantic representation (MRS) of the first ranked parse. In cases where JACY fails to return a parse, we use a dependency parser instead (Nichols et al. 2005). The highest scoping real predicate is generally the hypernym. For example, for *doraibā*$_2$ the hypernym is 人 *hito* "person" and for *doraibā*$_3$ the hypernym is クラブ *kurabu* "club". We also extract other relationships, such as synonym and domain. Because the words are sense tagged, we can specialize the relations to relations between senses, rather than just words: ⟨hypernym : doraibā$_3$, *kurabu*$_3$⟩. The relationships extracted for ドライバー *doraibā* "driver" are shown in Fig. 1.

One application of the synonym/hypernym relations is linking the lexicon to other lexical resources. We use a hierarchical match to link to the (Ikehara et al. 1997) and WordNet (Fellbaum 1998). Although looking up the translation adds noise, the additional filter of the relationship triple effectively filters it out again (Bond et al. 2004b). These links are shown in Fig. 1.

6 Discussion and further work

Similar annotation efforts in other languages include the Penn Propbank (Palmer et al. 2005) for English and Chinese, which has added structural semantics and some lexical semantics (predicate argument structure and role labels) to syntactically annotated corpora, but not full lexical semantic information (i.e. word senses). The most similar project to ours is OntoNotes (Hovy et al. 2006). It combines syntactic annotation (treebank) structural semantics (propbank), lexical semantics (word senses) and an ontology, along with co-reference annotation, for both English and Chinese. The main difference (apart from the target languages) is in the static dynamic design: in the Hinoki project we expect to improve our grammar and ontology and update accordingly.

The Hinoki data is currently being used to provide data for a range of experiments, including training a parse ranking model and a word sense disambiguation (WSD) system; acquisition of deep lexical types using super tagging; annotation of lexical conceptual structure for Japanese verbs at the sense level; and calculation of sentence similarity using lexical and structural semantics. Using sense information improves the parse-ranking accuracy by as much as 5.6% compared to using purely syntactic features (Fujita et al. 2007). Similarly using the parse results improves the sense disambiguation (Tanaka et al. 2007).

In further work, we are improving (i) feature engineering for the parsing and disambiguation models, ultimately leading to a combined model; (ii) the coverage of the grammar, so that we can parse more sentences to a correct parse; and (iii) the knowledge acquisition, in particular learning other information from the parsed defining sentences, such as lexical-types, meronyms, and antonyms.

7 Conclusion

In this paper we have described the current state of the Hinoki treebank. We have further showed how it is being used to develop a language-independent system for acquiring thesauruses from machine-readable dictionaries. With the improved grammar and ontology, we will use the knowledge learned 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.

References

Bond, F., Fujita, S., Hashimoto, C., Kasahara, K., Nariyama, S., Nichols, E., Ohtani, A., Tanaka, T., & Amano, S. (2004a). The Hinoki treebank: A treebank for text understanding. In *Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)* (pp. 554–559). Hainan Island.

Bond, F., Nichols, E., Fujita, S., & Tanaka, T. (2004b). Acquiring an ontology for a fundamental vocabulary. In *20th International Conference on Computational Linguistics: COLING-2004* (pp. 1319–1325). Geneva.

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

Copestake, A., Flickinger, D., Pollard, C., & Sag, I. A. (2005). Minimal recursion semantics. An introduction. *Research on Language and Computation, 3*(4), 281–332.

Fellbaum, C. (Ed.) (1998). *WordNet: An electronic lexical database*. MIT Press.

Fujita, S., Bond, F., Oepen, S., & Tanaka, T. (2007). Exploiting semantic information for HPSG parse selection. In *ACL 2007 Workshop on Deep Linguistic Processing*, Prague (pp. 25–32).

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, New York City, USA (pp. 57–60).

Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Ooyama, Y., & Hayashi, Y. (1997). *Goi-Taikei – A Japanese lexicon* (5 volumes/CDROM). Tokyo: Iwanami Shoten.

Kasahara, K., Sato, H., Bond, F., Tanaka, T., Fujita, S., Kanasugi, T., & Amano, S. (2004). Construction of a Japanese semantic lexicon: Lexeed. In *IPSG SIG: 2004-NLC-159*, Tokyo (pp. 75–82). (in Japanese).

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In E. Hinrichs & D. Roth (Eds.), *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 423–430).

Kurohashi, S., & Nagao, M. (2003). Building a Japanese parsed corpus – While improving the parsing system. In A. Abeillé (Ed.), *Treebanks: Building and using parsed corpora* (Chap. 14, pp. 249–260). Kluwer Academic Publishers.

Nichols, E., Bond F., & Flickinger, D. (2005). Robust ontology acquisition from machine-readable dictionaries. In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-2005*, Edinburgh (pp. 1111–1116).

Oepen, S., Flickinger, D., Toutanova, K., & Manning, C. D. (2004). LinGO redwoods: A rich and dynamic treebank for HPSG. *Research on Language and Computation, 2*(4), 575–596.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics, 31*(1), 71–105.

Pollard, C., & Sag, I. A. (1994). *Head driven phrase structure grammar*. Chicago: University of Chicago Press.

Shirai, K. (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)* (pp. 605–608).

Siegel, M. (2000) HPSG analysis of Japanese. In W. Wahlster (Ed.), *Verbmobil: Foundations of speech-to-speech translation* (pp. 265–280). Berlin, Germany: Springer.

Snyder, B., & Palmer, M. (2004). The English all-words task. In *Proceedings of Senseval-3* (pp. 41–44). Barcelona.

Stevenson, M. (2003). *Word sense disambiguation*. CSLI Publications.

Tanaka, T., Bond, F., Baldwin, T., Fujita, S., & Hashimoto, C. (2007). Word sense disambiguation incorporating lexical and structural semantic information. In *The 2007 Joint Meeting of the Conference on Empirical Methods on Natural Language Processing (EMNLP) and the Conference on Natural Language Learning (CONLL)*. Prague.

Tanaka, T., Bond, F., & Fujita, S. (2006). The Hinoki sensebank – A Large-scale word sense tagged corpus of Japanese. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006* (pp. 62–69). Sydney.

Tanaka, T., Bond, F., Oepen, S., & Fujita, S. (2005). High precision treebanking – Blazing useful trees using POS information. In *ACL-2005* (pp. 330–337).

Toutanova, K., Manning, C. D., Flickinger, D., & Oepen, S. (2005). Stochastic HPSG parse disambiguation using the redwoods corpus. *Research on Language and Computation, 3*(1), 83–105.