A Thesaurus of Predicate-Argument Structure for Japanese Verbs to Deal with Granularity of Verb Meanings

Koichi Takeuchi
Okayama University /
koichi@cl.cs.
okayama-u.ac.jp

Kentaro Inui
Tohoku University /
inui@ecei.
tohoku.ac.jp

Nao Takeuchi
Free Language
Analyst

Atsushi Fujita
Future University Hakodate /
fujita@fun.ac.jp

Abstract

In this paper we propose a framework of verb semantic description in order to organize different granularity of similarity between verbs. Since verb meanings highly depend on their arguments we propose a verb thesaurus on the basis of possible shared meanings with predicate-argument structure. Motivations of this work are to (1) construct a practical lexicon for dealing with alternations, paraphrases and entailment relations between predicates, and (2) provide a basic database for statistical learning system as well as a theoretical lexicon study such as Generative Lexicon and Lexical Conceptual Structure. One of the characteristics of our description is that we assume several granularities of semantic classes to characterize verb meanings. The thesaurus form allows us to provide several granularities of shared meanings; thus, this gives us a further revision for applying more detailed analyses of verb meanings.

1 Introduction

In natural language processing, to deal with similarities/differences between verbs is essential not only for paraphrase but also textual entailment and QA system which are expected to extract more valuable facts from massively large texts such as the Web. For example, in the QA system, assuming that the body text says "He lent her a bicycle", the answer of the question "He gave her a bicycle?" should be "No", however the answer of "She rented the bicycle?" should be "Yes". Thus constructing database of verb similarities/differences enables us to deal with detailed paraphrase/non-paraphrase relations in NLP.

From the view of the current language resource, how the shared/different meanings of "He lent her a bicycle" and "He gave her a bicycle" can be described? The shared meaning of *lend* and *give* in the above sentences is that they are categorized to *Giving Verbs*, as in Levin's English Verb Classes and Alternations (EVCA) (Levin, 1993), while the different meaning will be that *lend* does not imply ownership of the theme, i.e., *a bicycle*. One of the problematic issues with describing shared meaning among verbs is that semantic classes such as *Giving Verbs* should be dependent on the granularity of meanings we assumed. For example, the meaning of *lend* and *give* in the above sentences is not categorized into the same Frame in FrameNet (Baker et al., 1998). The reason for this different categorization can be considered to be that the granularity of the semantic class of *Giving Verbs* is larger than that of the *Giving* Frame in FrameNet[1]. From the view of natural language processing, especially dealing the with propositional meaning of verbs, all of the above classes, i.e., the wider class of *Giving Verbs* containing *lend* and *give* as well as the narrower class of *Giving* Frame containing *give* and *donate*, are needed. Therefore, in this work, in order to describe verb meanings with several granularities of semantic classes, a thesaurus form is adopted for our verb dictionary.

Based on the background, this paper presents a thesaurus of predicate-argument structure for verbs on the basis of a lexical decompositional framework such as Lexical Conceptual Structure (Jackendoff, 1990); thus our

[1]We agree with the concept of Frame and FrameElements in FrameNet but what we propose in this paper is the necessity for granularities of Frames and FrameElements.

Proceedings of the 8th Workshop on Asian Language Resources, pages 1–8,
Beijing, China, 21-22 August 2010.

proposed thesaurus can deal with argument structure level alternations such as causative, transitive/intransitive, stative. Besides, taking a thesaurus form enables us to deal with shared/differenciate meaning of verbs with consistency, e.g., a verb class node of "lend" and "rent" can be described in the detailed layer of the node "give".

We constructed this thesaurus on Japanese verbs and the current status of the verb thesaurus is this: we have analyzed 7,473 verb meanings (4,425 verbs) and organized the semantic classes in a five-layer thesaurus with 71 semantic roles types. Below, we describe background issues, basic design issues, what kind of problems remain, limitations and perspectives of applications.

2 Existing Lexical Resources and Drawbacks

2.1 Lexical Resources in English

From the view of previous lexical databases In English, several well-considered lexical databases are available, e.g., EVCA, Dorr's LCS (Dorr, 1997), FrameNet, WordNet (Fellbaum, 1998), VerbNet (Kipper-Schuler, 2005) and PropBank (Palmer et al., 2005). Besides there is the research project (Pustejovsky and Meyers, 2005) to find general descriptional framework of predicate argument structure by merging several lexical databases such as PropBank, NomBank, TimeBank and PennDiscouse Treebank.

Our approach corresponds partly to each lexical database, (i.e., FrameNet's Frame and FrameElements correspond to our verb class and semantic role labels, and the way to organize verb similarity classes with thesaurus corresponds with WordNet's synset), but is not exactly the same; namely, there is no lexical database describing several granularities of semantic classes between verbs with arguments. Of course, since the above English lexical databases have links with each other, it is possible to produce a verb dictionary with several granularities of semantic classes with arguments. However, the basic categories of classifying verbs would be little different due to the different background theory of each English lexical database; it must be not easy to add another level of semantic granularity with keeping consistency for all the lexical databases; thus, thesaurus form is needed to be a core form for describing verb meanings[2].

2.2 Lexical Resources in Japanese

In previous studies, several Japanese lexicons were published: IPAL (IPA, 1986) focuses on morpho-syntactic classes but IPAL is small[3]. EDR (Jap, 1995) consists of a large-scale lexicon and corpus (See Section 3.4). EDR is a well-considered and wide coverage dictionary focusing on translation between Japanese and English, but EDR's semantic classes were not designed with linguistically-motivated lexical relations between verbs, e.g., alternations, causative, transitive, and detransitive relations between verbs. We believe these relations must be key for dealing with paraphrase in NLP.

Recently Japanese FrameNet (Ohara et al., 2006) and Japanese WordNet (Bond et al., 2008) are proposed. Japanese FrameNet currently published only less than 100 verbs[4]. Besides Japanese WordNet contains 87000 words and 46000 synsets, however, there are three major difficulty of dealing with paraphrase relations between verbs: (1) there is no argument information; (2) existing many similar synsets force us to solve fine disambiguation between verbs when we map a verb in a sentence to WordNet; (3) the basic verbs of Japanese (i.e., highly ambiguous verbs) are wrongly assigned to unrelated synsets because they are constructed by translation from English to Japanese.

[2] As Kipper (Kipper-Schuler, 2005) showed in their examples mapping between VerbNet and WordNet verb senses, most of the mappings are many-to-many relations; this indicates that some two verbs grouped in a same semantic type in VerbNet can be categorized into different synsets in WordNet. Since WordNet does not have argument structure nor syntactic information, we cannot purchase what is the different features for between the synsets.

[3] It contains 861 verbs and 136 adjectives.

[4] We are supplying our database to Japanese FrameNet project.

3 Thesaurus of Predicate-Argument Structure

The proposed thesaurus of predicate-argument structure can deal with several levels of verb classes on the basis of granularity of defined verb meaning. In the thesaurus we incorporate LCS-based semantic description for each verb class that can provide several argument structure such as construction grammar (Goldberg, 1995). This must be high advantage to describe the different factors from the view of not only syntactic functions but also internal semantic relations. Thus this characteristics of the proposed thesaurus can be powerful framework for calculating similarity and difference between verb senses. In the following sections we explain the total design of thesaurus and the details.

3.1 Design of Thesaurus

The proposed thesaurus consists of hierarchy of verb classes we assumed. A verb class, which is a conceptual class, has verbs with a shared meaning. A parent verb class includes concepts of subordinate verb class; thus a subordinate verb class is a concretization of the parent verb class. A verb class has a semantic description that is a kind of semantic skeleton inspired from lexical conceptual structure (Jackendoff, 1990; Kageyama, 1996; Dorr, 1997). Thus a semantic description in a verb class describes core semantic relations between arguments and shadow arguments of a shared meaning of the verb class. Since verb can be polysemous, each verb sense is designated with example sentences. Verb senses with a shared meaning are assigned to a verb class. Every example sentence is analyzed into their arguments and semantic role types; and then their arguments are linked to variables in semantic description of verb class. This indicates that one semantic description in a verb class can provide several argument structure on the basis of syntactic structure. This architecture is related to construction grammar.

Here we explain this structure using verbs such as *rent, lend, give, hire, borrow, lease*. We assume that each verb sense we focus on here is designated by example sentences, e.g., "Mother gives a book to her child", "Kazuko rents a bicycle from her friend", and "Taro lend a car to his friend". As Figure 1 shows that all of the above verb senses are involved in the verb class *Moving of One's Possession* [5]. The semantic description, which expresses core meaning of the verb class *Moving of One's Possession* is

```
([Agent] CAUSE)
BECOME [Theme] BE AT [Goal].
```

Where the brackets [] denote variables that can be filled with arguments in example sentences. Likewise parentheses () denote occasional factor. "Agent" and "Theme" are semantic role labels that can be annotated to all example sentences. Figure 1 shows that the children of the verb class *Moving of One's Possession* are the two verb classes *Moving of One's Possession/Renting* and *Moving of One's Possession/Lending*. In the *Renting* class, *rent, hire and borrow* are there, while in the *Lending* class, *lend and lease* exist. Both of the semantic descriptions in the children verb classes are more detailed ones than the parent's description.

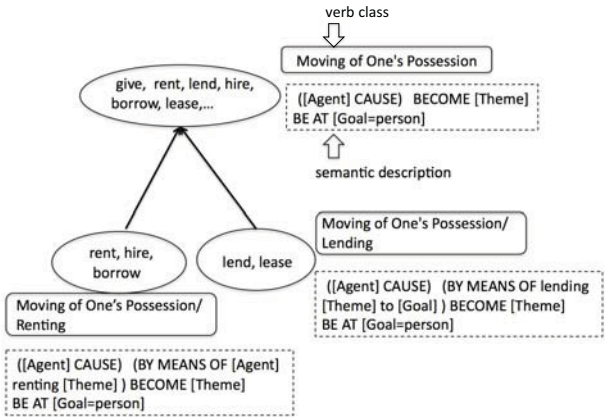

Figure 1: Example of verb classes and their semantic descriptions in parent-children.

A semantic description in the *Renting* class, i.e.,

```
([Agent] CAUSE)
```

[5] The name of a verb class consists of hierarchy of thesaurus; and Figure 1 shows abbreviated verb class name. Full length of the verb class name is *Change of State/Change of Position (Physical)/Moving of One's Possession*.

```
(BY MEANS OF [Agent]  renting [Theme])
BECOME [Theme]  BE AT [Agent],
```

describes semantic relations between "Agent" and "Theme". Since semantic role labels are annotated to all of the example sentences, the variables in the semantic description can be linked to practical arguments in example sentences via semantic role labels (See Figure 2).

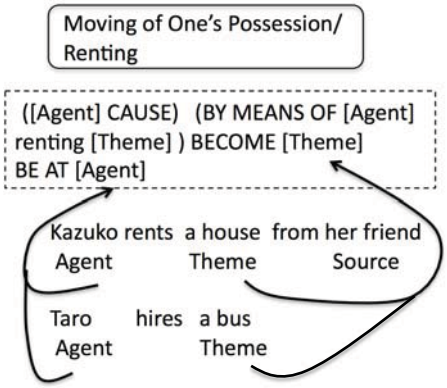

Figure 2: Linking between semantic description and example sentences.

3.2 Construction of Verb Class Hierarchy

To organize hierarchical semantic verb class, we take a top down and a bottom up approaches. As for a bottom up approach, we use verb senses defined by a dictionary as the most fine-grained meaning; and then we group verbs that can be considered to share some meaning. As for a dictionary, we use the Lexeed database (Fujita et al., 2006), which consists of more than 20,000 verbs with explanations of word sense and example sentences.

As a top down approach, we take three semantic classes: *State*, *Change of State*, and *Activity* as top level semantic classes of the thesaurus according to Vendler's aspectual analysis (Vendler, 1967) (See Figure 4). This is because the above three classes can be useful for dealing with the propositional, especially, resultative aspect of verbs. For example "He threw a ball" can be an *Activity* and have no special result; but "He broke the door" can be a *Change of State* and then we can imagine a result, i.e., *broken door*. When other verb senses can express the same results, e.g., "He destroyed the door," we would like to regard them as having the same meaning.

We define verb classes in intermediate hierarchy by grouping verb sense on the basis of aspectual category (i.e., action, state, change of state), argument type (i.e., physical, mental, information), and more detailed aspects depending on aspectual category. For example, *walk the country*, *travel all over Europe* and *get up the stairs* can be considered to be in the *Move on Path* class.

Verb class is essential for dealing with verb meanings as synsets in WordNet. Even if we had given an incorrect class name, the thesaurus will work well if the whole hierarchy keeps is-a relation, namely, the hierarchy does not contain any multiple inheritance.

The most fine-grained verb class before individual verb sense is a little wider than alternations. Currently, for the fine-grained verb class, we are organizing what kind of differentiated classes can be assumed (e.g., manner, background, presupposition, and etc.).

3.3 Semantic Role Labels

The aim of describing arguments of a target verb sense is (1) to link the same role arguments in a related verb sense and (2) to provide disambiguated information for mapping a surface expression to a verb sense. The Lexeed database provides a representative sentence for each word sense. The sentence is simple, without adjunctive elements such as unessential time, location or method. Thus, a sentence is broken down into subject and object, and semantic role labels are annotated to them (Figure 3).

ex.:	nihon-ga	shigen-wo	yunyuu-suru
trans.:	Japan (NOM)	resouces (ACC)	import
AS:	Agent	Theme	

Figure 3: An example of semantic role label.

Of course, only one representative sentence would miss some essential arguments; also, we

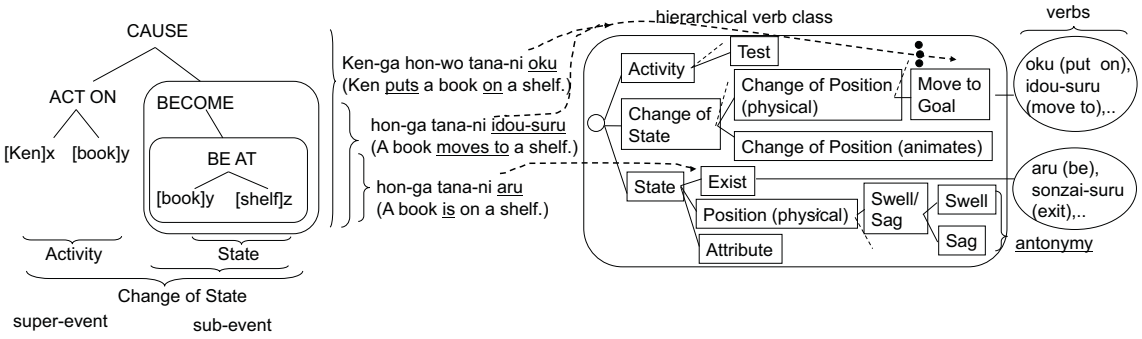

Figure 4: Thesaurus and corresponding lexical decomposition.

do not know how many arguments are enough. This can be solved by adding examples[6]; however, we consider the semantic role labels of each representative sentence in a verb class as an example of assumed argument structure to a verb class. That is to say, we regard a verb class as a concept of event and suppose it to be a fixed argument frame for each verb class. The argument frame is described as compositional relations.

The principal function of the semantic role label name is to link arguments in a verb class. One exception is the *Agent* label. This can be a marker discriminating transitive and intransitive verbs. Since the semantic class of the thesaurus focuses on *Change of State*, transitive alternation cases such as "The president expands the business" and " The business expands" can be categorized into the same verb class. Then, these two examples are differentiated by the *Agent* label.

3.4 Compositional Semantic Description

As described in Section 3.1, we incorporate compositional semantic structure to each verb class to describe syntactically motivated lexical semantic relations and entailment meanings that will expand the thesaurus. The benefit of compositional style is to link entailed meanings by means of compositional manner. As an example of entailment, Figure 5 shows that a verb class *Move to Goal* entails *Theme* to be *Goal*, and this corresponds to a verb class *Exist*.

[6]We are currently constructing an SRL annotated corpus.

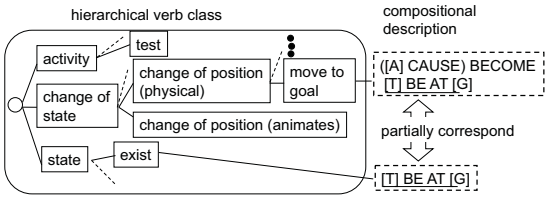

Figure 5: Compositional semantic description.

In this verb thesaurus, being different from previous LCS studies, we try to ensure the compositional semantic description as much as possible by means of linking each sub-event structure to both a semantic class and example sentences. Therefore, we believe that our verb thesaurus can provide a basic example data base for LCS study.

3.5 Intrinsic Evaluation on Coverage

We did manual evaluation that how the proposed verb thesaurus covers verb meanings in news articles. The results on Japanese new corpus show that the coverage of verbs is 84.32% (1825/2195) in 1000 sentences randomly sampled from Japanese news articles[7]. Besides we take 200 sentences and check whether the verb meanings in the sentences can correspond to verb meaning in our thesaurus. The result shows that our thesaurus meaning covers 99.5% (199 verb meanings/200 verb meanings) of 200

[7]Mainichi news article in 2003.

verbs[8].

4 Discussions

4.1 Comparison with Existing Resources

Table 1 and Table 2 show a comparison of statistical characteristics with existing resources. In the tables, WN and Exp denote the number of word meanings and example sentences, respectively. Also, SRL denotes the number corresponding to semantic role label.

Looking at number of concepts, our Thesaurus has 709 types of concepts (verb classes) which is similar to FrameNet and more than VerbNet. This seems to be natural because FrameNet is also language resource constructed on argument structure. Thanks to our thesaurus format, if we need more fine grained concepts, we can expand our thesaurus by adding concepts as new nodes at the lowest layer in the hierarchy. While at the number of SRL, FrameNet has much more types than our thesaurus, and in the other resources VerbNet and EDR the number of SRL is less than our thesaurus. This comes from the different design issue of semantic role labels. In FrameNet they try to differentiate argument types on the basis of the assumed concept, i.e., Frame. In contrast with FrameNet we try to merge the same type of meaning in arguments. VerbNet and EDR also defined abstracted SRL; The difference between their resources and our thesaurus is that our SRLs are defined taking into account what kind of roles in the core concept i.e., verb class; while SRLs in VerbNet and EDR are not dependent on verb's class.

Table 2 shows that our thesaurus does not have large number in registered words and examples comparing to EDR and JWordNet. As we stated in Section 3.5, the coverage of our verb class to newspaper articles are high, but we try to add examples by constructing annotated Japanese corpus of SRL and verb class.

[8]This evaluation is done by one person. Of course we need to check this by several persons and take inter-annotator agreement.

Table 1: Comparing to English resources

	FrameNet	WordNet	VerbNet
Concepts	825 (Frame) (Ver 1.3) 2007	N/A (Synset) (Ver 3.0) 2006	237 (class) (Ver 2.2) 2006
Words	6100	155287	3819
WM	N/A	117659	5257
Exp	13500	48349	N/A
SRL	746	N/A	23
POS	V,N,A,Ad	V,N,A,Ad	V
Lang	E,O	E,O	E

Table 2: Comparing to Japanese resources

	EDR	JWordNet	Our Thesaurus
Concepts	430000 (class) (Ver 3.0) 2003	N/A (Ver 0.92) 2009	709 2008
Words	410000	92241	4425
WM	270000	56741	7473
Exp	200000	48276	7473
SRL	28	N/A	71
POS	all	V,N,A,Ad	V
Lang	EJ	E,J	J

4.2 Limitations of Developed Thesaurus

One of the difficulties of annotating the semantic class of word sense is that a word sense can be considered as several semantic classes. The proposed verb thesaurus can deal with multiple semantic classes for a verb sense by adding them into several nodes in the thesaurus. However, this does not seem to be the correct approach. For example, what kind of *Change of State* semantic class can be considered in the following sentence?

a. *He took on a passenger*.

Assuming that *passenger* is *Theme*, *Move to goal* could be possible when we regard the vehicle[9] as *Goal*. In another semantic class, *Change State of Container* could be possible when we regard the vehicle as a container. Currently, all of the verb senses are linked to only one semantic class that can be considered as the most related semantic class.

[9]*Vehicle* does not appear in the surface expression but *vehicle* can exist. We currently describe the shadow argument in the compositional description, but it would be hard to prove the existence of a shadow argument.

From the user side, i.e., dealing with the propositional meaning of the sentence (**a.**), various meanings should be estimated. Consider the following sentence:

b. *Thus, we were packed.*

As the semantic class of the sentence (**a.**) *Change State of Container* could better explain why they are packed in the sentence (**b.**)

The other related issue is how we describe the scope, e.g.,

c. *He is hospitalized.*

If we take the meaning as a simple manner, *Move to Goal* can be a semantic class. This can be correct from the view of annotation, but we can guess *he cannot work* or *he will have a tough time* as following events. FrameNet seems to be able to deal with this by means of a special type of linking between Frames.

Consequently, we think the above issues of semantic class should depend on the application side's demands. Since we do not know all of the requirements of NLP applications currently, then it must be sufficient to provide an expandable descriptional framework of linguistically motivated lexical semantics.

4.3 Remaining and Conceivable Ways of Extension

One of the aims of the proposed dictionary is to identify the sentences that have the same meanings among different expressions. One of the challenging paraphrase relations is that the sentences expressed from the different view points. Given the buying and selling in Figure 6, a human can understand that both sentences denote almost the same event from different points of view. This indicates that the sentences made by humans usually contain the point of view of the speaker. This is similar to a camera, and we need to normalize the expressions as to their original meaning.

We consider that NLP application researchers need to relate these expressions. Logically, if we know "buy" and "sell" have shared meanings of *giving and taking things*, we can describe their

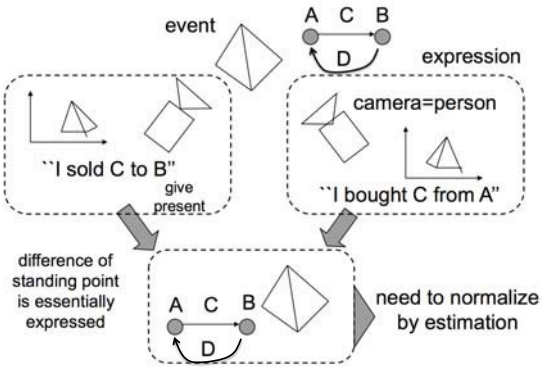

Figure 6: Requirement of normalization to deal with different expressions from the different views.

relations with "or" in logical form. Therefore, finding and describing these verb relations will be essential for dealing with propositional meanings of a sentence.

For further view of application, event matching to find a similar situation in Web documents is supposed to be a practical and useful application. Assuming that a user is confronted with the fact that wireless LAN in the user's PC does not work, and the user wants to search for documents that provide a solution, the problem is that expressions of situations must be different from the views of individual writers, e.g., "wireless LAN did not work" or "wireless LAN was disconnected". How can we find the same meaning in these expressions, and how can we extract the answers by finding the same situation from FAQ documents? To solve this, a lexical database describing verb relations between "go wrong" and "disconnect" must be the base for estimating how the expressions can be similar. Therefore, constructing a lexicon can be worthwhile for developing NLP applications.

5 Conclusion

In this paper, we presented a framework of a verb dictionary in order to describe shared meaning as well as to differentiate meaning between verbs from the viewpoint of relating eventual expressions of NLP. One of the characteristics is that we describe verb relations on the basis of several

semantic granularities using a thesaurus form with argument structure. Semantic granularity is the basis for how we categorize (or recognize which semantic class relates to a verb meaning). Also, we ensure functions and limitations of semantic classes and argument structure from the viewpoint of dealing with paraphrases. That is, required semantic classes will be highly dependent on applications; thus, the framework of the verb-sense dictionary should have expandability. The proposed verb thesaurus can take several semantic granularities; therefore, we hope the verb thesaurus will be applicable to NLP's task[10].

In future work, we will continue to organize differentiated semantic classes between verbs and develop a system to identify the same event descriptions.

Acknowledgments

We would like to thank NTT Communication Research Laboratory for allowing us to use the Lexeed database, and Professor Koyama for many useful comments.

References

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 86–90.

Bond, Francis, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008. Construction of Japanese WordNet from Multi-lingual WordNet. In *Proceedings of the 14th Annual Meeting of Japanese Natural Language Processing*, pages 853–856.

Dorr, Bonnie J. 1997. Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. *Machine Translation*, 12(4):271–325.

Fellbaum, Chistiane. 1998. *WordNet an Electronic Lexical Database*. MIT Press.

Fujita, Sanae, Takaaki Tanaka, Fransis Bond, and Hiromi Nakaiwa. 2006. An implemented description of japanese: The lexeed dictionary and the hinoki treebank. In *COLING/ACL06 Interactive Presentation Sessions*, pages 65–68.

Goldberg, Adele E. 1995. *Constructions*. The University of Chicago Press.

IPA: Information-Technology Promotion Agency, Japan, 1986. *IPA Lexicon of the Japanese Language for Computers*.

Jackendoff, Ray. 1990. *Semantic Structures*. MIT Press.

Japan Electronic Dictionary Research Institute, Ltd, 1995. *EDR: Electric Dictionary the Second Edition*.

Kageyama, Taro. 1996. *Verb Semantics*. Kurosio Publishers. (In Japanese).

Kipper-Schuler, K. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, PhD Thesis, University of Pennsylvania.

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

Ohara, Kyoko Hirose, Seiko Fujii, Toshio Ohori, Ryoko Suzuki, Hiroaki Saito, and Shun Ishizaki. 2006. Frame-based contrastive lexical semantics and japanese framenet: The case of risk and kakeru. In *Proceeding of the Fourth International Conference on Construction Grammar*. http://jfn.st.hc.keio.ac.jp/ja/publications.html.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Pustejovsky, J. and Martha P.and A. Meyers. 2005. Merging propbank, nombank, timebank, penn discourse treebank and coreference. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 5–12.

Vendler, Zeno. 1967. *Linguistics in Philosophy*. Cornell University Press.

[10]The proposed verb thesaurus is available at: http://cl.cs.okayama-u.ac.jp/rsc/data/. (in Japanese).