Pacific Association for Computational Linguistics

THE JAPANESE FRAMENET PROJECT: A Preliminary Report

KYOKO HIROSE OHARA, SEIKO FUJII[1], HIROAKI SAITO[2],

SHUN ISHIZAKI[3], TOSHIO OHORI[1], RYOKO SUZUKI

Keio University, 4-1-1 Hiyoshi, Kohoku-ku, Yokohama City, Kanagawa Pref. 223-8521, JAPAN

This report presents an overview of the Japanese FrameNet (JFN) research project, which started in July 2002.[4] The goal of JFN is to create a corpus-based lexicon of Japanese described in terms of frame semantics.

While JFN aims at building a Japanese lexicon in collaboration with the Berkeley FrameNet Project, an important question being asked by JFN is whether Japanese words can be described in FrameNet style, i.e., along the same lines as English words, employing the same frame semantic approach. This point is illustrated in this paper with an example of preliminary analysis of Japanese communication verbs.

Finally, it is argued that while JFN can be described as a lexicographic project, such an effort will be of great use to NLP applications such as machine translation and to learners of Japanese.

Key words: lexical semantics, corpus linguistics, frame semantics, Japanese

1. INTRODUCTION

This report presents an overview of the Japanese FrameNet (JFN) research project, which started in July 2002. The goal of JFN is to create a corpus-based lexicon of Japanese described in terms of frame semantics (Fillmore 1982).

JFN can be described as a counterpart to the English-based FrameNet project, an ongoing project undertaken at the International Computer Science Institute in Berkeley, California (Ruppenhofer, Baker, Fillmore 2002). The key features of Berkeley FrameNet are: (a) a commitment to corpus evidence for semantic and syntactic generalizations; (b) the representation of the valences of the target words using frame semantics for the semantic portion. The resulting database will contain: (a) descriptions of the semantic frames with their frame elements (FE) which underlie the meanings of the words described, and (b) the valence representation of words and phrases, each accompanied by (c) a collection of annotated corpus attestations (Baker, Fillmore, Lowe 1998).

In recent years, attempts to create lexical entries for languages other than English using the frame semantic approach have been undertaken. Besides JFN, German FrameNet and Spanish FrameNet are currently under development (Boas 2002). JFN aims at building a Japanese lexicon in collaboration with the projects.

JFN is headquartered on Hiyoshi Campus of Keio University and includes researchers from Keio University and University of Tokyo. So far, a corpus has been chosen and a pilot study is being undertaken to analyze communication and motion verbs in Japanese. Also, a corpus tool for extracting data from corpora has been implemented.

---

[1] University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-0041, Japan.

[2] Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama City, Kanagawa Pref. 223-8522, Japan.

[3] Keio University, 5322 Endo, Fujisawa City, Kanagawa Pref. 252-8520, Japan.

[4] This work is being supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan and Keio University.

The rest of the paper is structured as follows. Section 2 describes the goals of JFN. Section 3 illustrates the corpus and computational tools used in the project. Section 4 gives an example of preliminary analysis of Japanese communication verbs.

2. PROJECT GOALS

The ultimate goal of JFN is to produce a FrameNet-style database of Japanese words. The resulting database will thus contain valence descriptions of Japanese words and a collection of annotated corpus attestations. In producing this database we will explore whether Japanese words can be described along the same lines as English words, employing the same frame semantic approach.

JFN is currently concentrating on analyzing basic verbs in Japanese. More specifically, it focuses on verbs and uses of verbs that have not been described in detail in existing Japanese dictionaries, whose lexical descriptions tend not to be based on corpus attestations.

3. THE CORPUS AND THE TOOL FOR ANALYZING JAPANESE

3.1 The JFN Corpus

Currently the JFN corpus contains approximately 1 million sentences, taken from Kyoto University Annotated Text Corpus (hereafter Kyoto University Corpus) and the Mainichi newspaper (CD-Mainichi Newspaper 1995). Kyoto University Corpus contains morphologically and syntactically annotated data for 40,000 sentences (about 1.6 million words). An example of morphological annotation of Kyoto University Corpus is shown:

#S-ID:950909001-001

*	0	2D				
彼	kare ‘he’	*	noun	noun	*	*
は	wa ‘topic marker’	*	particle	adverbial particle	*	*
*	1	2D				
東京	tookyo ‘Tokyo’	*	noun	proper noun	*	*
に	ni ‘GOAL’	*	particle	case particle	*	*
*	2	-1D				
行った	itta ‘went’	行く	verb	*	子音動詞カ行促音便形	
	タ形					

EOS

Sentence boundaries are shown by a start ID and an EOS label. Each row represents a morpheme: shown from left to right are its surface form, phonetic form, root form (‘*’ for words that do not conjugate), part of speech, detailed part of speech, conjugation type and conjugated form. An asterisk (‘*’) in the first column represents a phrasal delimiter and the following integer indicates the phrase number. The number and the letter following the phrase number specify one of the three kinds of relations between phrases: ‘D’ stands for dependency, ‘P’ for parallel relations, and ‘A’ for adposition.

3.2 The KWIC Search Tool

A KWIC search tool has been developed in JFN. Using the morphological annotations just

described, the tool searches for both the root form and conjugated forms of a keyword at the same time. Another key feature of this KWIC search tool is the fact that it can be used with a dependency structure analyzer called CaboCha. CaboCha, developed at Nara Institute of Science and Technology, performs morphological analysis as well as syntactic parsing of any Japanese sentence. Although CaboCha sometimes parses colloquial sentences incorrectly, using our KWIC search tool together with CaboCha enables us to add any text to our corpus.

Currently there are three display modes in our KWIC search tool: 'Parse Tree Mode', 'Morphological Analysis Mode', and 'Context Display Mode'. Figure 1 shows a snapshot of Parse Tree Mode:

FIGURE 1. A Screenshot of Parse Tree Mode

The entire screen consists of a Search Input Window to inputs a keyword to be searched, a File Window to specify file(s) in which a keyword is searched, a KWIC Window displaying all the sentences containing the keyword and allowing the user to highlight any sentence by clicking on it, a Sentence Window showing the highlighted sentence, and a Parse Tree Window which displays a tree of the highlighted sentence.

This KWIC search tool is written in Ruby script language, and runs on Linux and Solaris operating systems as well as on various Windows platforms. JFN plans to make it publicly available.

4. A PRELIMINARY ANALYSIS OF JAPANESE COMMUNICATION VERBS

This section briefly reports on a preliminary analysis of several basic verbs in the semantic domain of communication in Japanese, with an eye to establishing frames and FEs for JFN, based on relevant frames and FEs established for English FN. Examined at the initial stage have been such frames as Statement, Conversation, Communication, Contacting, etc. An important question to keep in mind here is to what extent the existing English-based frames, FEs, and their descriptions can be applied to the Japanese cases. In the efforts to establish Japanese-based frames and FEs, the key issues faced in the communication domain are: (i) how to identify and

capture multiple senses and uses associated with a single form; and (ii) how to deal with recognized differences in senses and conditions of use among verbs related in meaning.

Basic communication verbs in Japanese include: *yuu* ('say'), *hanasu* ('speak' or 'talk'), *syaberu* ('chat' or 'chatter'), *noberu* ('state'), *kataru* ('tell' or 'narrate'), etc. These verbs have different distributions, as noted by Shibata, Kunihiro, Nagashima, Yamada, & Asano (1979). For example, *yuu* can be used in such expressions as (1a) expressing a verbal act which does not involve the addressee, but *hanasu* cannot be used in the same way (shown in 1b).

(1) a. *Hitori-goto* *o* *yuu*.
One-person-talk (monologue) ACC say
'Mumble to oneself.'
b. * *Hitori-goto o* *hanasu*.
speak/talk

On the other hand, *hanasu* can be used in such expressions as (2a) 'talking about one's experience' (cf. Figure 1 in Section 3), but not *yuu* as in (2b):

(2) a. *Keiko-tyan wa tennyuugo itukakan wa*
Keiko TOP transfer-in-after five-days TOP
zisin-taiken o hanasitagaranakatta.
earthquake-experience ACC speak/talk-want-NEG-PAST
'Keiko did not want to talk about her experiences with the earthquake for five days after being transferred into (the program).'
b. * *zisin-taiken o* *yuu*
say

The uses illustrated in (1) and (2) above take a direct object, marked with the accusative marker *o*. In addition, both *yuu* and *hanasu* can take a complement marked with the quotative particle *to*, as illustrated in (3) and (4):

(3) *Dansi-seito ga [Tokyo ni iku] to itte ...*
Male student NOM Tokyo to go QUO(tative) say
'The male student said that (I would) go to Tokyo.'

(4) *Hasimoto tizi wa [kokuseki zyookoo no hituyoosei wa*
Hashimoto governor TOP [nationality article GEN necessity TOP
kanzinai] to hanasite-iru.
feel-NEG] QUO speak-ASP
'Governor Hashimoto says that he does not feel the necessity for an article regarding nationality.'

(5) [*Tie o dase] to yuwaretemo (yuu + (r)are+ te-mo)*
Ideas ACC squeeze QUO say-passive-concessive
I am told, "Come up with some ideas."
* [*Tie o dase] to hanasaretemo (hanasu+ (r)are+ te-mo)*

Here again, these verbs exhibit different conditions of use. As shown in (5), the quoted message with the imperative form can be the complement of *yuu*, but not of *hanasu*.

To deal with the phenomena exemplified above, the preliminary analysis has been conducted in two approaches: capturing different uses and senses by setting up frames necessary for Japanese in addition to frames already established in English FN; and capturing different uses

and senses by refining FEs, in particular, by identifying sub-categories of the FE called MESSAGE.

In the first approach, *yuu* and *hanasu* can be associated with both the STATEMENT and CONVERSATION frames, just as English verbs are associated in the English FN. To capture the phenomena illustrated in (1), however, it appears useful to break the STATEMENT frame into STATEMENT-1 (Verbal Act) and STATEMENT-2 (Verbal Transfer). This distinction, together with other frames newly introduced for Japanese, has been proposed and is currently being tested with the corpus data.

In the second approach pertaining to FE, MESSAGE is one of the key FEs for these Japanese verbs, exactly as with the English counterparts. Both the *o*-marked direct object (as in 1 and 2) and the *to*-marked complement (as in 3, 4, 5) all represent the MESSAGE FE. For Japanese, however, further distinctions for this FE have been hypothesized. The four-way distinction, which is currently being tested with the corpus data, is summarized as follows:

MESSAGE:	Message-Report-Form
	Message-Report-Content
	Message-Description-Form
	Message-Description-Content

Using these frames and FEs, examples (1) through (5) can be annotated:

STATEMENT-1 (Verbal Act)

(1) a. [INI SPEAKER] [*Hitori-goto o* Message-Description-Form] *yuu*.
'Mumble to oneself.'

STATEMENT-2 (Verbal Transfer)

(2) a. [*Keiko-tyan wa* SPEAKER] *tennyuugo itukakan wa*
[zisin-taiken o Message-Description-Content] *hanasitagaranakatta*.
'Keiko did not want to talk about her experiences with the earthquake for five days after being transferred into (the program).'

STATEMENT-2 (Verbal Transfer)

(3) [*Dansi-seito ga* SPEAKER] *[Tokyo ni iku to* Message-Report-Content/Form] *itte*
'The male student said that (I would) go to Tokyo.'

STATEMENT-2 (Verbal Transfer)

(4) [*Hasimoto tizi wa* SPEAKER] [*kokuseki zyookoo no hituyoosei wa kanzinai to* Message-Report-Content] *hanasite-iru*.
'Governor Hashimoto says that he does not feel the necessity for an article regarding nationality.'

STATEMENT-2 (Verbal Transfer)

(5) [*Tie o dase to* Message-Report-Form] *yuwaretemo*
I am told, "Come up with some ideas."

Only a few examples of variations with the two verbs *yuu* and *hanasu* and their annotations have been shown in the present section, but further important variations can be found with various communication verbs. Using target sentences extracted from our corpus, these variations are being sorted and annotated, examining the two approaches explained above and testing the significance of the newly-proposed frames and FEs for Japanese communication verbs.

5. CONCLUSION AND OUTLOOK

This paper has outlined the overview, computational environments, and a preliminary analysis of JFN. In its second year, JFN plans to continue analyzing communication and motion verbs in Japanese, while at the same time expanding our corpus. Also, following the pilot study conducted in the first year, JFN will start semantic annotation of sentences extracted from our corpus.

JFN is primarily concerned with building a Japanese lexicon based on the frame semantic approach and thus can be described as a lexicographic project. We believe, however, that such an effort will be of great use to NLP applications such as machine translation and to learners of Japanese.

ACKNOWLEDGMENTS

We wish to thank Takuya Kosugi, Kiyoko Uchiyama, and Eric Long for their valuable discussions and comments regarding this paper.

REFERENCES

Baker, Collin F., Fillmore, Charles J., and Lowe, John B. (1998): The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*, Montreal, Canada.

Boas, Hans C. (2002): Bilingual FrameNet Dictionaries for Machine Translation. In *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas, Spain. Vol. IV: 1364-1371.

CaboCha

http://cl.aist-nara.ac.jp/~taku-ku/software/cobocha

CD-Mainichi Newspaper 1995.

Fillmore, Charles J. (1982): Frame semantics; in Linguistics in the Morning Calm pp. 111-137, Hanshin Publishing Co., Seoul, South Korea.

FrameNet Website

http://www.icsi.berkeley.edu/~framenet/

JUMAN

http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman-e.html

Kyoto University Annotated Text Corpus

http://www.kc.t.u-tokyo.ac.jp/nl-resource/corpus-e.html

Ruppenhofer, Josef, Collin F. Baker and Charles J. Fillmore (2002): Collocational Information in the FrameNet Database. In Braasch, Anna and Claus Povlsen (eds.), *Proceedings of the Tenth Euralex International Congress*. Copenhagen, Denmark. Vol. I: 359-369.

Shibata, Takeshi, Tetsuya Kunihiro, Yoshiro Nagashima, Susumu Yamada, Yuriko Asano (1979): *Kotoba no imi* (Meanings of words) Vol. 2. Heibonsha, Tokyo, Japan.