Semantic annotation of a Japanese speech corpus

John Fry
Linguistics Dept. & CSLI
Stanford University
Stanford CA 94305-2150 USA
fry@csli.stanford.edu

Francis Bond*
Machine Translation Research Group
NTT Communication Science Laboratories
2-4 Hikari-dai, Kyoto 619-0237 JAPAN
bond@cslab.kecl.ntt.co.jp

Abstract

This paper describes the semantic annotations we are performing on the **CallHome Japanese** corpus of spontaneous, unscripted telephone conversations (LDC, 1996). Our annotations include (i) semantic classes for all nouns and verbs; (ii) verb senses for all main verbs; and (iii) relations between main verbs and their complements in the same utterance. Our semantic tagset is taken from NTT's **Goi-Taikei** semantic lexicon and ontology (Ikehara et al., 1997). A pilot study demonstrates that the verb sense tagging can be efficiently performed by native Japanese speakers using computer-generated HTML forms, and that good inter-annotator reliability can be obtained in the right conditions.

1 Introduction

Semantic annotations have proved valuable for a variety of NLP tasks, including parsing, word sense disambiguation, coreference resolution, summarization, and information retrieval and extraction. The most challenging domain for all these tasks is *spontaneous spoken language*, which tends to be more terse, less grammatical, less structured, and more ambiguous than planned or written text. For this reason, the annotation of spoken language corpora with accurate, high-quality linguistic tags has become a topic of great interest recently (Dybkjær et al., 1998; Ide, 1998; Core et al., 1999).

The target of our semantic annotations is the CallHome Japanese (CHJ) corpus (LDC, 1996). The CHJ corpus consists of digitized speech data and text transcriptions of 120 spontaneous, unscripted telephone conversations in Japanese. Each transcript is encoded in EUC-format Japanese characters and covers a contiguous 5 or 10 minute segment taken from a recorded conversation lasting up to 30 minutes. To illustrate, a brief fragment (the first three utterances) of a CHJ transcript is given in Figure 1 (an English gloss appears below the fragment). Each utterance in a transcript is analyzed into individual morphemes, with transcriber comments in brackets. The speaker (A or B) and the start and end times of each utterance (i.e. speaker turn) are also provided in the transcripts. The 120 conversations in the CHJ corpus contain a total of about 340,000 word/morpheme tokens, 12,000 unique word/morpheme types, and 39,000 speaker turns.

The CHJ corpus was originally created for research in large vocabulary speech recognition. However, we hope to make the corpus useful for other types of NLP research (by ourselves and others) by supplementing it with a variety of linguistic annotations. When finished, we plan to make the annotated CHJ corpus available to the research community through the LDC.

We are annotating the CHJ corpus with a variety of syntactic, semantic, and acoustic/prosodic tags. In this paper we focus on our semantic annotations, which include the following:

- **Semantic classes** for all verbs and nouns.
- **Verb senses** for each main verb.
- **Predicate-argument relations** between main verbs and their explicitly mentioned verb complements, labeled with thematic roles.

By way of example, Table 1 shows the end time boundary t_e (i.e. duration), POS, pronunciation, canonical form (including verb sense

* Visiting CSLI, Stanford University (1999-2000).

```
120.20 123.35 A: 嘘、 @私[[あたし,col]] 職 が なくて さ 日本 に 帰った ら。
123.26 123.70 B: うん。
124.28 128.50 A: トニー の お 誕生日 九月 二日 だ から @何か[[なんか,col]] 買って あげ
よう と 思った けど、 @私[[あたし,col]] さ 無職 の 人間 だ {laugh} 。
```

Translation:

A:	嘘、	私	職	が	なくて	さ	日本	に	帰った	ら。
	Uso,	atashi	shoku	-ga	nakute	sa	nihon	-ni	kaetta	-ra.
	Lie,	I	employment	NOM	not-have	y'know	Japan	DAT	returned	if.

'No way. Having no employment, y'know, maybe I should return to Japan?'

B: うん。
Un.

'Uh-huh.'

A:	トニー	の	お	誕生日	九月	二日	だ	から	何か	買って	あげ	よう	と
	Tonii	-no	o-	tanjoubi	kugatsu	futsuka	da	kara	nanka	katte	age-	you	-to
	Tony	GEN	HON	birthday	Sept.	2nd	is	because	something	buy	give	VOL	COMP

思った	けど、	私	さ	無職	の	人間	だ。
omotta	kedo,	atashi	sa	mushoku	-no	ningen	da.
think	but,	I	y'know	jobless	GEN	person	am.

'Tony's birthday is Sept. 2 so I want to buy him something, but I'm unemployed.'

Figure 1: Fragment from CHJ transcript 0696

ID	Word	t_e (s)	POS	Phonetic	Canonical	Class	Arguments
003	あたし	0.810	pro	atasi	私	c0008	
004	職	1.060	noun	syoku	職	c1939	
005	が	1.180	part	ga			
006	なくて	1.580	v-neg,te	nakute	無い(3)	v0003	N1:003,N2:004
007	さ	1.840	part	sa			
008	日本	2.134	prop	nihoN	日本	c0385,c0463,p0030	
009	に	2.250	part	ni			
010	帰った	2.610	v-r5	kaeqta	帰る(2)	v0014	N1:003,N5:008
011	ら	2.940	cond2	ra			

Table 1: Sample annotations from the first utterance of 0696

number), semantic class, and argument indexes for most of the first utterance from Figure 1.

Our current plan is to provide our annotations in the simple tabular text format shown in Table 1, rather than in one of the of the numerous annotation frameworks currently in contention.[1] If a reliable and widely-accepted XML encoding framework emerges before we release our annotations, then we will consider adopting that scheme. However, our primary aim is to provide simple, accurate, low-level annotations (upon which other, higher-level annotations might be based) so that language researchers can use the corpus more flexibly and with greater confidence.

Some of the annotations which we have already completed, or nearly completed, but will *not* discuss in this paper include the following:

- **Phonetic transcriptions**, in Roman characters (*kunreisiki* transliteration), of all 120 conversations.
- **POS tags** using the LDC's existing inventory of 60 syntactic and morphological tags for Japanese.

[1] For example, 53 annotation frameworks are listed at `http://morph.ldc.upenn.edu/annotation/`. For speech corpora, notable contributions include the Corpus Encoding Standard (CES, `http://www.cs.vassar.edu/CES/`), MATE (Dybkjær et al., 1998), and the 'annotation graph' approach of Bird and Liberman (1998).

- Raw **acoustic data** from the ESPS/waves+ speech processing software, including f_0 (fundamental frequency) and power (root-mean-square amplitude) measurements at 10ms intervals.
- The **duration** of each word, based on semi-automatic word segmentation of the speech data.

The remainder of this paper is organized as follows. In Section 2 we describe NTT's Goi-Taikei semantic dictionary, which is the source of our semantic tagset. Section 3 describes our tagging methodology and our pilot study of the verb sense annotation task. Finally, Section 4 describes our browser-based annotation application.

2 Goi-Taikei

As a base for our tags, we are using the Goi-Taikei (**GT**) Japanese lexicon (Ikehara et al., 1997), a 400,000-word lexicon and ontology developed by NTT for machine translation (MT) applications.

We decided that **GT** is an appropriate resource for our semantic annotation task for three reasons. First, semantic information from **GT** has already proved valuable in a variety of NLP applications in Japan, including parsing, morphological analysis, text-to-speech, proofreading, and MT (Ikehara et al., 1994; Shirai et al., 1995; Akiba et al., 1995; Oku, 1996; Nakaiwa and Seki, 1999; Baldwin et al., 1999; Baldwin and Tanaka, 1999; Yokoyama and Ochiai, 1999). Secondly, the **GT** lexicon and ontology, at 400,000 words, is significantly larger than earlier dictionaries, such as the 260,000-word EDR Dictionary (EDR, 1996) and the 2,000-word IPAL lexicon (IPA, 1987; IPA, 1990; IPA, 1996). **GT** also contains detailed valency information for 16,000 predicate senses, which makes it more suited to our task than the Kadokawa thesaurus (Hamanishi and Ono, 1990). Finally, **GT** is available in book and CD-ROM format at a price (around US $750) that is several times lower than EDR.

GT consists of three main components: (i) an ontology, (ii) a semantic word dictionary, and (iii) a semantic structure dictionary which includes subcategorization frames for verbs and adjectives.

2.1 Ontology

GT's ontology classifies concepts to use in expressing relationships between words. The meanings of common nouns are given in terms of a semantic hierarchy of 2,710 nodes. Most of the top four levels of the semantic hierarchy are shown in Figure 2, with two examples of deeper nodes. Each node represents a semantic class. Edges in the hierarchy represent **is-a** or **has-a** relationships, so that the child of a semantic class related by an **is-a** relation is subsumed by it. For example, `nation` **is-a** `organization`. In addition to the 2,710 classes (12-level tree structure) for common nouns, there are 200 classes (9-level tree structure) for proper nouns and 108 classes (5-level tree structure) for predicates.

2.2 Semantic Word Dictionary

The **GT** semantic dictionary includes 100,000 common nouns, 200,000 proper nouns, 70,000 technical terms and 30,000 other words: 400,000 words in all.

Figure 3 shows a simplified example of one record of the Japanese semantic word dictionary. Each record specifies an index form, pronunciation, canonical form, syntactic information and a set of semantic classes. The syntactic information includes the part of speech, inflectional class, detailed parts of speech, conjunctive conditions and so on. Each word can have up to five common noun semantic classes and ten proper noun semantic classes. The numbering system gives common-noun semantic classes a prefix of `c`, proper-noun classes a prefix of `p`, and predicate classes a prefix of `v`. In Figure 3, for example, the word 日本 *nihon* "Japan" belongs to the common-noun classes `c0385 nation` ($\subset$ `organization`) and `c0463 territory` ($\subset$ `place`), and to proper-noun class `p0030 country` ($\subset$ `place name`). More examples of semantic classes for the nouns and verbs from the annotated CHJ fragment in Table 1 are listed in that table under the column labeled 'Class'.

2.3 Semantic Structure Dictionary

The basic structure of a clause comes from the relationship between the main verb and nouns. **GT**'s structure transfer dictionary, designed for MT applications, provides this basic clause structure. **GT** provides 10,000 patterns in its common structure transfer dictionary and

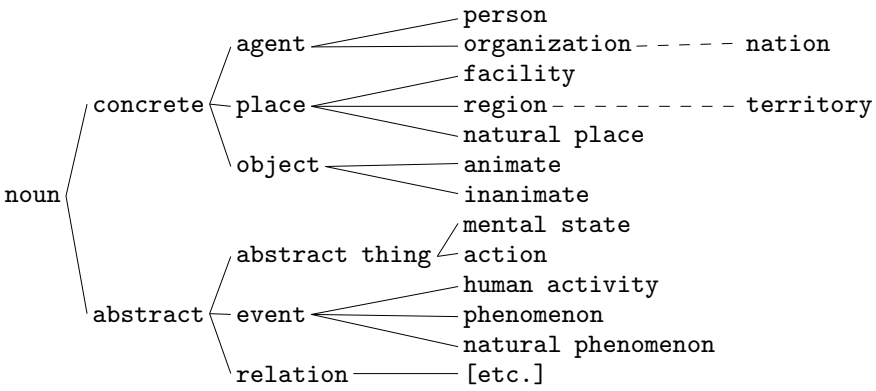

Figure 2: **GT**'s Semantic Hierarchy (top 4 levels)

Index Form	日本
Pronunciation	/nihon/
Canonical Form	日本
Part of Speech	noun
Semantic Classes	common noun: nation (c0385), territory (c0463) proper noun: country (p0030)

Figure 3: Japanese Lexical Entry for noun 日本 *nihon* "Japan"

5,000 patterns in its idiomatic structure transfer dictionary. The common structure transfer dictionary contains an average of 2.3 patterns for each verb.

Figure 4 gives an example from the common structure transfer dictionary. Each predicate is associated with one or more arguments labeled N1, N2, ... Each case-slot contains information such as grammatical function, case-marker, case-role, semantic restrictions on the filler and default order (not all features are shown in the example). The arguments correspond between Japanese and English, thus giving the backbone of the transfer.

2.3.1 Case Roles

Case-elements in the valency dictionary are associated with particular case roles (also known as thematic roles, θ-roles, or deep cases). The current set of case roles is given in Table 2, along with the most commonly associated case markers[2] in Japanese, and prepositions or grammatical functions (gf) in English. The annotated CHJ fragment in Table 1 shows some specific case-role fillers under the column labeled 'Arguments'. For example, word 008, 日本 *nihon* "Japan", serves as the goal argument (N5) of the verb 帰った *kaetta* "returned (home)" in that utterance.

There seems to be no consensus among linguists on what the best set of case roles is, or even whether case roles should be replaced by more abstract primitives or more concrete participant-roles. In any case, case roles have in practice proved extremely useful for NLP and are used in most MT systems (Bond and Shirai, 1997).

3 Tagging methodology

We are annotating the CHJ corpus with (i) semantic classes for all nouns and verbs; (ii) verb senses for all main verbs; and (iii) predicate-argument relations between main verbs and their complements in the same utterance. Our tagging of verb senses and predicate-argument relations relies on the browser-based annotation application described in Section 4. The predicate-argument tags, based on **GT**'s semantic structure dictionary (Section 2.3), provide a basic dependency structure for each utterance.

[2] Japanese case markers (also known as 'particles') are postpositions, and are similar to English prepositions in many ways.

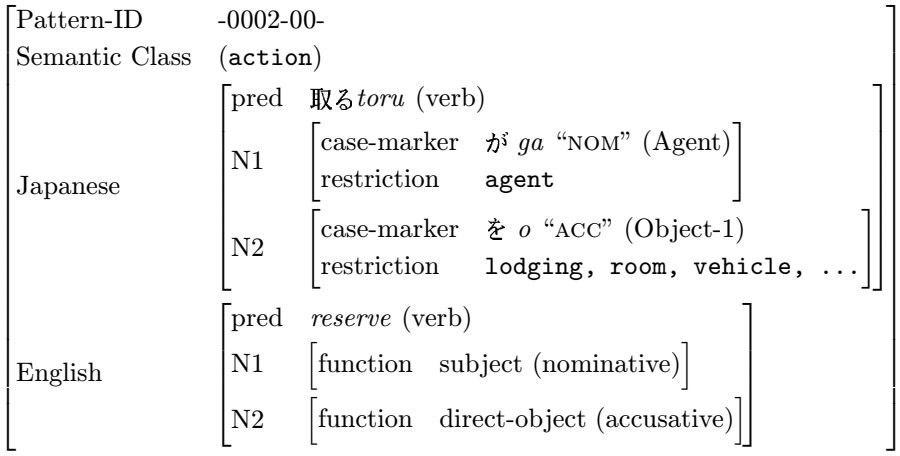

Figure 4: Part of the common structure transfer dictionary for one sense of 取る *toru* "take"

Label	Name	Case-marker	English	Label	Name	Case-marker	English
N1	Agent	ga (kara, towa)	Subj (gf)	N8	Locative	ni, o, de, e, kara	in/at/on
N2	Object-1	o (nituite)	Obj (gf)	N9	Comitative	to	with
N3	Object-2	ni (...)	I-Obj (gf)	N10	Quotative	to	
N4	Source	kara, yori	from	N11	Material	kara, yori, de	with, from
N5	Goal	ni, e, made	to (until)	N12	Cause	kara, yori, de	for
N6	Purpose	ni	for	N13	Instrument	de	with
N7	Result	ni, to	as	N14	Means	de	by

Table 2: Case-roles

Although spontaneous utterances like those in the CHJ corpus are often fragmentary and ungrammatical, rendering full syntactic parsing impractical, the basic relations between predicates and their arguments still hold.

3.1 Tagging semantic classes

We assign **GT** semantic classes to individual CHJ nouns and verbs by automatic table lookup on their **GT** canonical form. In both **GT** and the CHJ lexicon, the canonical forms of words are generally in Chinese characters (*kanji*). For example, the annotated CHJ fragment in Table 1 shows some canonical forms under the column labeled 'Canonical'. For the verbs in Table 1, the canonical (dictionary) form includes the **GT** verb sense number.

In the majority of cases, the **GT** canonical form of a verb or noun is identical to the canonical form which appears in the CHJ lexicon. The small percentage of cases where the canonical forms differ are corrected by hand. For the approximately 700 nouns in the CHJ corpus that are not covered in the **GT** lexicon (mainly personal names), we assign the closest available **GT** class(es) by hand.

We are marking each noun and verb in the corpus with *all* of its **GT** semantic classes, even those which might be inappropriate for the word in its particular utterance context. For example, in the annotated CHJ fragment in Table 1, the noun 日本 *nihon* "Japan" is marked with all three of its **GT** classes: `c0385 nation` (⊂ `organization`), `c0463 territory` (⊂ `place`), and `p0030 country` (⊂ `place name`). Naturally, it would preferable to exclude those semantic classes which are inappropriate for a given noun or verb in its particular context of use in the CHJ corpus. However, this would require human coders to classify hundreds of thousands of word tokens based on the perceived context in the conversation. In addition, it would be hard to obtain high inter-annotator reliability, given the context-dependent nature of the task and the amount of overlap in the semantic classes.

3.2 Tagging verb senses and arguments

We are providing human-tagged verb sense and verb argument annotations for each main verb in the corpus. Auxiliary verbs and other forms of verb morphology are ignored, except in cases like the passive and causative in which the valence of the main verb is altered. In those cases, special passive or causative senses are provided.

Our plan is to annotate the **GT** verb senses and argument indexes according to the majority judgments of three native-speaker student assistants. The students will make the annotations by clicking on menu choices in a web application that we generate automatically from the **GT** dictionary files and CHJ transcript files (Section 4).

3.3 A pilot study

In preparation for the verb sense tagging project, we conducted a pilot study in which we asked five native speakers to select **GT** verb senses and identify intrasentential arguments for all 110 main verbs in one five-minute CHJ transcript.

Our initial results showed that pairwise inter-annotator agreement on verb senses was 0.68. When chance agreement is taken into account via the kappa statistic, the result, $\kappa = 0.63$, shows that annotator agreement was not reliable (Carletta, 1996). However, we discovered that this result was largely attributable to the annotators' selection of the category "none (of the above)," which the five judges picked with highly variable frequency ($\{7, 11, 22, 25, 26\}, s = 8.6$). For 51 of the 110 verb tokens (46%), "none" was selected by one or more judges, and agreement was low (0.48, $\kappa = 0.42$). When the "none" answers were disregarded, pairwise agreement on those verbs rose considerably (to 0.67, $\kappa = 0.64$). For the remaining 59 verb tokens (54%), "none" was never chosen and pairwise agreement was very reliable (0.84, $\kappa = 0.82$). For the second task, identifying the intrasentential arguments to verbs, pairwise agreement was also very high: 0.89 among annotators who chose the same verb sense.

We then examined more closely those cases in which the low agreement was attributable to inconsistent use of the category "none". We found that most of these cases involved very common, 'light' verbs such as する *suru* "do" and なる *naru* "become". As it turned out, **GT** was lacking some common colloquial senses for these verbs. For example, the verb する *suru* "do" was often used as in utterance (1).

(1) 後 2週間 した ら
ato ni-shuu-kan shita ra
after 2-weeks-long did if
'in about two weeks'

This sense of する *suru* "do" does not appear in the **GT** lexicon. In written Japanese *suru* would not normally be used in this way; rather, a more specialized verb such as 経つ *tatsu* "pass" would be preferred.

In sum, the results of our pilot study lead us to conclude that we are likely to obtain reliable inter-annotator agreement on the verb sense task, provided the following steps are taken:

- The highly general, spoken-language senses of certain common 'light' verbs need to be added to **GT**.
- The conditions under which the coders are to use the category "none (of the above)" need to be carefully delineated. In particular, "none" should only be selected when there is a particular, standard, well-defined sense of the verb in question that clearly should be listed in the lexicon but is not.

We are encouraged in this regard by the results of Kilgarriff and Rosenzweig (1999) who were able, using careful experimental methodology, to achieve replicable agreement on English word senses (albeit by lexicographers, not students) of 95% in the SENSEVAL project. Our experience echoes the observation of Kilgarriff (1998) that annotators should be given the opportunity to offer feedback to the lexicographers, including information such as inadequate or missing verb senses.

4 The annotation application

We developed a browser-based semantic annotation application specifically for this project. The application is implemented using HTML forms within a web browser (Figure 5). We wrote Perl scripts to generate the annotation forms automatically, using the CHJ transcripts, CHJ lexicon, and **GT** database as input.

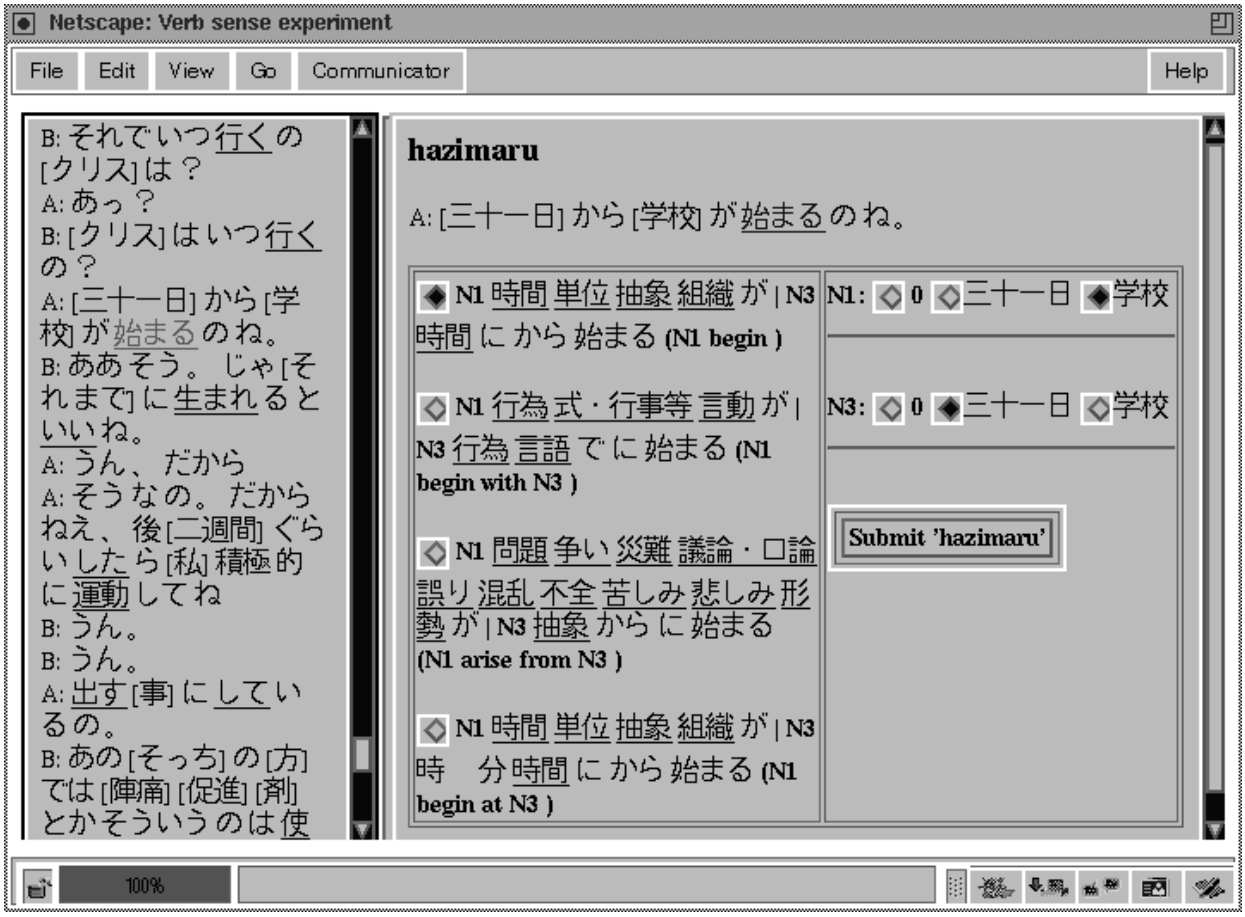

Figure 5: Screen shot of verb sense diambiguation application

The left frame of the application displays the transcript of a CHJ conversation. NPs in the transcript are enclosed in square brackets, and main verbs are underlined and hyperlinked. Clicking on a main verb in the transcript (left frame) brings up a verb sense menu for that verb in the right frame.

For example, in Figure 5, the verb sense menu for a token of the verb 始まる *hazimaru* "begin" (fourth utterance in the left frame) has been brought up in the right frame. The four **GT** senses of *hazimaru* are listed as menu choices. Each sense is assigned a unique subcategorization frame, including the case roles (N1, N2, etc.; cf. Table 2). The subcategorized-for semantic categories are also underlined and hyperlinked to a diagram of the complete **GT** ontology (cf. Figure 2), so that the coders can see examples of each category and how that category fits into the broader semantic framework. Finally, an English gloss of each verb sense (from the **GT** transfer component) is given at the end of each subcategorization frame.

Once a coder selects the correct verb sense for the verb token in question (in the case of Figure 5, it is the first sense listed), the coder then selects that verb's NP complements (case-role fillers), if any, from within the same utterance. For example, in Figure 5, the coder has selected the first verb sense for 始まる *hazimaru* "begin", which subcategorizes for the NP arguments N1 (subject) and N3 (start time). Separate menu forms are displayed for both N1 and N3, with all NPs in the utterance listed as possible fillers. In this case the coder selected 学校 *gakkou* "school" as the subject and 三十一日 *sanjuuichinichi* "the 31st" as the start time. If no NP in the utterance fills a given role, the zero option is selected (this is the default choice). Finally, the coder clicks '`Submit`' and then moves on to the next verb in the transcript (left frame).

Acknowledgments

This project is supported by U.S. National Science Foundation grant BCS-0002646, and by a dissertation grant in Japanese studies to the first author from Stanford University's Institute for International Studies.

References

Y. Akiba, M. Ishii, H. Almuallim, and S. Kaneda. 1995. Learning English verb selection rules from hand-made rules and translation examples. *Sixth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-95*, pp. 206–220, Leuven, July.

T. Baldwin and H. Tanaka. 1999. Argument status in Japanese verb sense disambiguation. *TMI-99*, pp. 196–206, Chester, UK, August.

T. Baldwin, F. Bond, and B. Hutchinson. 1999. A valency dictionary architecture for machine translation. *TMI-99*, pp. 207–217, Chester, UK, August.

S. Bird and M. Liberman. 1998. Towards a formal framework for linguistic annotations. *ICSLP'98*, Sydney.

F. Bond and S. Shirai. 1997. Practical and efficient organization of a large valency dictionary. *NLPRS-97 Workshop on Multilingual Information Processing*, Phuket. (handout).

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

M. Core, M. Ishizaki, J. Moore, C. Nakatani, N. Reithinger, D. Traum, and S. Tutiya. 1999. Report of the third workshop of the Discourse Resource Initiative. Technical Report CC-TR-99-1, Chiba Corpus Project, Chiba U.

L. Dybkjær, N. Bernson, H. Dybkjær, D. McKelvie, and A. Mengel. 1998. The MATE markup framework. MATE Deliverable D1.2, LE Telematics Project LE4–8370, Denmark. `http://mate.nis.sdu.dk/`.

EDR. 1996. EDR Electronic Dictionary Version 1.5. Japan Electronic Dictionary Research Institute, Ltd. `http://www.iijnet.or.jp/edr/`.

M. Hamanishi and S. Ono, editors. 1990. *Ruigo Kokugo Jiten [Japanese Synonym Dictionary]*. Kadokawa Shoten, Tokyo.

N. Ide. 1998. Encoding linguistic corpora. *Proceedings of the Sixth Workshop on Very Large Corpora*, pp. 9–17, Montreal.

S. Ikehara, S. Shirai, and K. Ogura. 1994. Criteria for evaluating the linguistic quality of Japanese to English machine translations. *Journal of Japanese Society for Artificial Intelligence*, 9(4):569–579. (In Japanese).

S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. 1997. *Goi-Taikei — A Japanese Lexicon*. Iwanami Shoten, Tokyo. 5 volumes. `http://www.kecl.ntt.co.jp/icl/mtg/resources/GoiTaikei`.

IPA. 1987. *IPAL Lexicon of Basic Verbs*. Software Technology Center, Information-Technology Promotion Agency (IPA), Tokyo. (In Japanese).

IPA. 1990. *IPAL Lexicon of Basic Adjectives*. IPA, Tokyo. (In Japanese).

IPA. 1996. *IPAL Lexicon of Basic Nouns*. IPA, Tokyo. (In Japanese).

A. Kilgarriff and J. Rosenzweig. 1999. SENSEVAL: Report and results. *NLPRS'99*, pp. 362–367, Beijing, November.

A. Kilgarriff. 1998. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language*, 12(4):453–472.

LDC. 1996. Callhome Japanese corpus. Linguistic Data Consortium, University of Pennsylvania. `http://www.ldc.upenn.edu`.

H. Nakaiwa and K. Seki. 1999. Automatic addition of verbal semantic attributes to a Japanese-to-English valency transfer dictionary. *TMI-99*, pp. 185–195, Chester, UK.

M. Oku. 1996. Analyzing Japanese double-subject construction having an adjective predicate. *COLING-96*, pp. 865–870.

S. Shirai, S. Ikehara, A. Yokoo, and H. Inoue. 1995. The quantity of valency pattern pairs required for Japanese to English machine translation and their compilation. *NLPRS-95*, pp. 443–448, Seoul.

S. Yokoyama and T. Ochiai. 1999. Aimai-na sūryōshi-o fukumu meishiku-no kaisekihō [a method for analysing noun phrases with ambiguous quantifiers]. *ANLP-99*, pp. 550–553. (In Japanese).