Japanese SemCor: A Sense-tagged Corpus of Japanese

Francis Bond,[*,***] **Timothy Baldwin,**[**] **Richard Fothergill**[**] and **Kiyotaka Uchimoto**[***]
[*] Linguistics and Multilingual Studies, Nanyang Technological University, Singapore
[**] Computer Science and Software Engineering, Melbourne University, Australia
[***] National Institute of Information and Communications Technology, Japan
bond@ieee.org, tb@ldwin.net, richard.fothergill@gmail.com, uchimoto@nict.go.jp

Abstract

In this paper we describe the creation of the Japanese SemCor (JSEMCOR) sense-tagged corpus of Japanese. The corpus is a translation of the English SEMCOR, with senses projected across from English. The final corpus consists of 14,169 sentences with 150,555 content words of which 58,265 are sense tagged. The corpus is one of the corpora used to provide sense frequency data for the Japanese Wordnet.

1 Introduction

Wordnets have been shown to have utility across a broad range of applications, largely in combination with sense frequency data and sense-tagged corpora. This paper describes Japanese SemCor (JSEMCOR), a sense-tagged corpus for the Japanese Wordnet (Isahara et al., 2008), based on translation of the English SEMCOR and sense projection.

In order to produce annotated text quickly and cheaply, we adopt the method of **annotation transfer** pioneered in the Italian MULTISEMCOR (Bentivogli and Pianta, 2005). In this approach, a sense-tagged text in one language is translated into the language in question, and the sense annotations from the original corpus are projected onto the new language. The sense projection is based on a wordnet in the target language which is aligned with the wordnet that was used to sense tag the source language text. Bentivogli and Pianta (2005) found that annotation transfer led to sense tagging with a precision of 86% and coverage of 81% (that is 19% of open class words still needed to be annotated), at less cost than annotating from scratch. The main differences in our case are: (1) the target language (Japanese) is linguistically further removed from the source language (English); and (2) to boost coverage, we provide the translators with sense-specific translations of each open class words to optionally include in their translations.

Similarly to MULTISEMCOR, our method takes English SEMCOR and translates it into the target language. In addition to the immediate objective of deriving a sense-tagged corpus of Japanese based on Japanese Wordnet, we also create a trilingual (English–Italian–Japanese) sensebank, with potential applications in other tasks such as translation. Because the Japanese (and Italian) texts are translated, the sense distribution may not be truly representative of native Japanese text. Ultimately, we aim to supplement JSEMCOR with other sense-tagged data, based on native Japanese text.

Finally, in the same way that the English and Italian annotation revealed missing word senses for their respective wordnets, we expect to find and correct such errors in the Japanese Wordnet, which will be fed back to the developers.

In the next section we give a brief description of the base resources used in the creation of JSEMCOR. Next, we describe the creation, size and distribution of JSEMCOR (§3). Finally, we discuss future work (§4) and then conclude.

2 Resources

2.1 The English Wordnet

The wordnet used both to tag the English SEMCOR corpus (§2.3) and as the backbone of the Japanese wordnet (§2.2) is the Princeton WordNet of English (Fellbaum, 1998). SEMCOR is tagged with tags from version 1.6 and the Japanese wordnet aligns with version 3.0. PWN has a rich structure of semantic relations, but we are only using it as a source of sense inventories in this task.

Bond, Francis, Timothy Baldwin, Richard Fothergill and Kiyotaka Uchimoto (2012) Japanese SemCor: A Sense-tagged Corpus of Japanese, In Proceedings of the 6th International Global Wordnet Conference (GWC 2012), Matsue, Japan.

2.2 The Japanese Wordnet

The Japanese Wordnet is a large scale, freely available, semantic dictionary of Japanese. The National Institute of Information and Communications Technology (NICT) started developing the Japanese Wordnet in 2006, as part of its support for Natural Language Processing research in Japan. The first version (0.9) was released in February 2009. In the initial phase Japanese equivalents were added to synsets of the Princeton WordNet. These have been expanded and corrected in subsequent releases. The current release is version 1.1. It contains 57,238 synsets (concepts), 93,834 unique Japanese words and 158,058 senses (synset–word pairs). All synsets have Japanese definitions, and over 45,000 also have examples.

We give an example of an entry in Figure 1.

From the beginning, the Japanese Wordnet project planned to tag text in order to verify its coverage and get distribution information (Isahara et al., 2008), but no tagged text has been released so far.

2.3 SemCor and MultiSemCor

The English SEMCOR corpus is a sense-tagged corpus of English created at Princeton University by the WordNet Project research team (Landes et al., 1998). It was created very early in the WordNet project, and was one of the first sense-tagged corpora produced for any language. The corpus consists of a subset of the Brown Corpus (Francis and Kucera, 1979), and has been part-of-speech tagged and sense tagged. We use the subset of SEMCOR which was translated into Italian as part of MULTISEMCOR 1.1 (Bentivogli and Pianta, 2005).

MULTISEMCOR is an English/Italian parallel corpus created by translating the English SEMCOR corpus into Italian. Texts are aligned at the sentence and word level, and annotated with part of speech, lemma and word sense (PWN 1.6). MULTISEMCOR version 1.1 contains 116 English texts: 14,144 sentences and 261,283 tokens, of which 119,802 tokens are annotated with senses. These are aligned with their corresponding Italian translations. In this paper we only use the English texts, which are freely available.[1] The MULTISEMCOR team reports tag errors for around 2.5% of the English open-class tokens in the English SEMCOR (Bentivogli and Pianta, 2005).

3 Japanese SemCor

The initial data for the translation was created by taking the English SEMCOR data and mapping the senses to version 3.0 using the mappings created by Daude et al. (2003). These senses were then used to look up synsets in the Japanese Wordnet to be presented to the translators.

3.1 Creation

Similar to MULTISEMCOR, sense annotation in JSEMCOR was set up as a translation task, where translators were provided with a SEMCOR sentence in English and asked to generate a Japanese translation using the interface depicted in Figure 2. For each sense-indexed word in the original SEMCOR data, we provided translators with a list of all words contained in the corresponding Japanese Wordnet synset. Clicking on one of these words both appended that lemma to the translation, and recorded clickthrough data for the word, which

[1] The Italian texts are available free for research from the Istituto Trentino Di Cultura (ITC) http://multisemcor.itc.it.

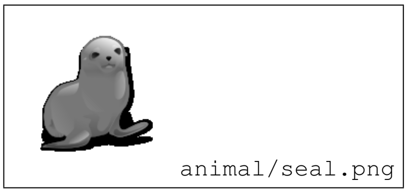
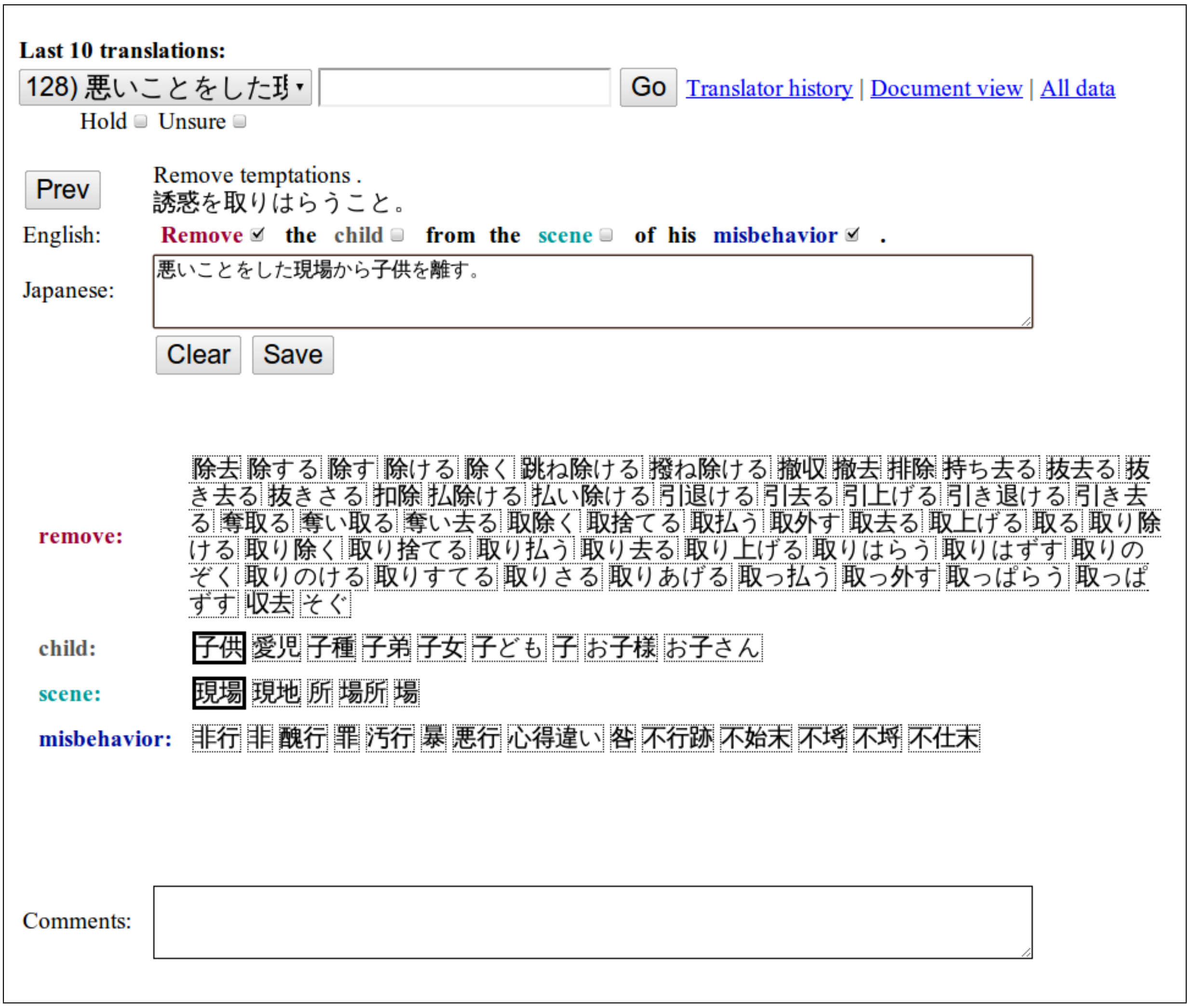
Synset	02076196-n
	
Synonyms	ja 海豹, アザラシ, シール en seal
Def (en)	“any of numerous marine mammals that come on shore to breed; chiefly of cold regions”
Def (ja)	「繁殖のために岸に上がる海洋性哺乳動物の各種；主に寒帯地域に」
Hypernyms	アシカ亜目/pinniped
Hyponyms	？/crabeater_seal ？/eared_seal 海驢/earless_seal

Figure 1: Example Entry for Seal/海豹



Last 10 translations:

128) 悪いことをした現 ▾ | Go | Translator history | Document view | All data

Hold □ Unsure □

Prev

Remove temptations .
誘惑を取りはらうこと。

English: **Remove** ☑ **the child** □ **from the scene** □ **of his misbehavior** ☑ **.**

Japanese: 悪いことをした現場から子供を離す。

Clear Save

remove: 除去 除する 除す 除ける 除く 跳ね除ける 撥ね除ける 撤収 撤去 排除 持ち去る 抜去る 抜き去る 抜きさる 扣除 払除ける 払い除ける 引退ける 引去る 引上げる 引き退ける 引き去る 奪取る 奪い取る 奪い去る 取除く 取捨てる 取払う 取外す 取去る 取上げる 取る 取り除ける 取り除く 取り捨てる 取り払う 取り去る 取り上げる 取りはらう 取りはずす 取りのぞく 取りのける 取りすてる 取りさる 取りあげる 取っ払う 取っ外す 取っぱらう 取っぱずす 収去 そぐ

child: 子供 愛児 子種 子弟 子女 子ども 子 お子様 お子さん

scene: 現場 現地 所 場所 場

misbehavior: 非行 非 醜行 罪 汚行 暴 悪行 心得違い 咎 不行跡 不始末 不埒 不埓 不仕末

Comments:

Figure 2: Screen shot of the annotation interface, for the SEMCOR sentence *Remove the child from the scene of his misbehaviour*

provided the basis for the ultimate sense tags in the translation. Translators also had the option to leave a comment (e.g. if they wanted to note something about the translation lists, or the source English), to mark their lack of confidence in a translation (via the “Unsure” checkbox), or to leave a translation to come back to (via the “Hold” checkbox).

In the example in Figure 2, shown in (1), the translator has used the words 子供 *kodomo* “child” and 現場 *geNba* “scene”, but has not made use of any of the translations for *remove* or *misbehavior* in their final translation (as indicated both by the lack of a highlighted translation and the ticked checkboxes for the respective words). They had the option of adding the new translation of *remove* to the synset, but did not use it here. The single word *misbehavior* was translated as the multi-word expression 悪い こと を した *warui-koto-wo shita* “did something bad”, probably because all of the translations of *misbehavior* sound more criminal than their English equivalents, making them inappropriate for child behaviour.

(1) a. *Remove*$^a_1$ *the child*$^b_1$ *from the scene*$^c_1$ *of his misbehavior*$_1$.

b.

悪い$_?$	こと$_?$	を	した$_?$	子供$^b_1$
warui	koto	wo	shita	kodomo
bad	thing	ACC	done	child

を	現場$^c_1$	から	離す$^a_1$	。
wo	genba	kara	hanasu	.
ACC	scene	from	remove	.

The final result for this sentence is that, of the four sense-tagged words in the original, two words have their sense transferred (^{b,c}), one word could be transferred if the new lemma were added to it (^{a}), and one word gets translated into three, none of which can be easily linked.

At the outset of the translation process, sentences were allocated to translators from a global *sentence* queue, meaning that if two translators were working in tandem, a given translator would often not translate contiguous sentences from a given document, potentially leading to lack of coherence in the translations. While translation coherence was not of primary interest, we switched across to allocating data from a *document* queue about 20% of the way through the translation process, in response to concerns over the resultant consistency in sense annotations within a given document, and requests from the translators. As part of this, we provided support for a "Document view", to allow the translator to look over a document in its entirety, including whatever progress had been made through the translation. We also gave translators the option of viewing the source English and translation for the immediately preceding sentence (*Remove temptations* and 疑惑を払うこと *giwaku-o harau koto*, resp., in Figure 2). They could also view their past 10 translations via the pulldown menu at the top-left of the translation page.

Translators were instructed to use translations provided in the list where possible, in order to maximise sense tagging coverage, except where this led to stilted Japanese (e.g. in translating an English deictic pronoun literally, rather than using a zero pronoun). The purpose of translation lists and the need for the clickthrough data was explained to the translators, although none of the translators were computational linguists, so the significance of sense tagging and the resulting sense-tagged corpus wasn't self-evident to them. Translators were also instructed to:

- use formal "editorial" Japanese, e.g. using the である *dearu* form of the copular, unless the text was clearly written in a colloquial or other style;

- attempt to determine the canonical translation/transliteration of proper names where possible, and failing this, to transliterate, flagging the translation as "Hold" if unsure of the pronunciation; acronyms were to be left as is, unless there was a well-known Japanese rendering of the acronym (e.g. *METI* for 通産省 *tsūsaNshō*);

- refrain from including alternative translations, e.g. in parentheses, in cases of doubt;

Corpus	SemCor	JSemCor
Sentences	12,842	14,169
Words	261,283	382,762
Content Words	119,802	150,555

Table 1: Corpus Size

- be faithful to the English sentence tokenisation (i.e. never translate multiple English sentences into a single Japanese translation), but to translate into multiple sentences in cases where it improved readability (e.g. for particularly long or heavily embedded English sentences);

- reorder the words where necessary to maximise readability in Japanese (esp. for conjunctions of nouns or adjectives);

- include discourse connectives where it improved overall sentence and document readability, irrespective of whether a corresponding sentential adverb (or equivalent) was included in the original English sentence.

3.2 Statistics

In contrast to Bentivogli and Pianta (2005), we have used manual rather than automatic word alignment. However, the alignment requires some post-processing before annotation transfer can occur. In this section, we look at various statistics of the alignment and annotation transfer process.

The word-alignment clickthrough data produced by our translators maps tokens in SemCor to lemmas in Japanese Wordnet, within the context of a translated sentence. In the following, we refer to a translated lemma in context as a **translation lemma**. Each translation lemma must be mapped onto the text of the translated sentence to complete the word alignment.

We perform this mapping automatically by first tokenising the sentence with the morphological analyser MeCab using the IPAdic lexicon and tagset (Kudo et al., 2004) and using the part of speech and lemma information it provides. This results in 382,762 tokens overall and 148,249 open class tokens, giving averages of 27 and 10.5 per sentence respectively.

The segmentation MeCab produces is fine grained relative to both English and to the Japanese Wordnet — in particular splitting compounds into their components — so we map translation lemmas to sequences of tokens. We accept a sequence of tokens as a match for a Princeton WordNet lemma if all parts in the translation match in their canonical word order, optionally allowing the final token to be in its lemmatised form, which is a convenient heuristic for lemmatising Japanese compounds. The numbers are summarized in Table 1.

Of 61,827 translation lemmas available, 7,551 are compounds with respect to IPAdic. Of the rest, 44,813 are single token and 9,463 are not found in the translation: the translation interface allowed free editing of the translation text but did not allow clickthrough word alignments to be undone.

Note that the resulting word alignment is not one-to-one: 1,734 translation lemmas come from more than one source word, though only 190 come from more than one source lemma (and none from more than two). Conversely, 3,252 translation lemmas match more than once in the translated sentence. Also, the alignment coverage is not complete: 51,450 sense tagged tokens in SEMCOR have not been translated, and 90,525 open class tokens in the Japanese sentence translations have no translation lemma mapped to them. Part of speech distributions for unaligned tokens in both languages are shown in Table 2.

After completion of the word alignment, we perform the annotation transfer. For a number of reasons, annotation transfer can result in zero or multiple senses being assigned to a word-aligned translation:

- Due to rearrangement of senses between WordNet versions 1.6 and 3.0, some SEMCOR tokens are annoted with deleted senses and others with more than one sense.

- We introduce additional variation in annotation multiplicity with a many-to-many word alignment.

- The presence in the translation data of user-contributed word translations means that an aligned word is not always in the transferred synset in Japanse Wordnet. In fact, this occurs 13,857 times, suggesting a large number of potential new synset memberships for Japanese Wordnet.

Therefore, of the 61,827 translation lemmas, 131 are assigned more than one sense and 13,771 have none. The remaining 47,925 translation lemmas are assigned a single sense. After taking into account translation lemmas which appear more than once — or not at all — in the target sentence, 46,121 words receive tags from the annotation transfer.

Due to the granularity mismatch between IPAdic and Japanese Wordnet, we take the additional step of mapping Japanese Wordnet lemmas to portions of text without word-aligned translations. The resulting compounds (or single tokens) do not receive a sense tag but are annotated with Japanese Wordnet lemma and part of speech. Where potential matches overlap, precedence is given first to longer matches (e.g., 米国政府 *beiko-kuseifu* "Washington" is chosen over 政府 *seifu* "government") and then to earlier matches (e.g. 近代化 *kiNdaika* "modernisation" is chosen over 化する *ka-suru* "to change" where they intersect in the out-of-vocabulary 近代化する *kiNdaika-suru* "to modernise"). This process produces an additional 61,495 unaligned words. We then include open class MeCab tokens which have still not been assigned a Japanese Wordnet lemma as an additional 34,329 words.

Finally, there are 12,144 monosemous Japanese words (with only a single sense) which were not annotated in the translation process, either because sense transfer fails or because the word is not aligned. Applying these single sense annotations brings the total number of sense annotated words to 58,265.

3.3 Distribution

We use the Kyoto Annotation Format (KAF) to share the corpus (Bosma et al., 2009). This is an emerging standard for wordnet annotation. We only use the two lowest layers (text and term), not including any higher levels such as dependencies or geodata. In order to make the data accessible, we will release it under the same license as the English SEMCOR. JSEMCOR is distributed with the Japanese Wordnet, available from `http://nlpwww.nict.go.jp/wn-ja/`.

A sample KAF record is presented in Figure 3, containing two words with Japanese Wordnet senses (学校 *gakkō* "school" and 戻る *modoru* "return"), IPAdic part-of-speech tags for all tokens, and file and sentence IDs which align with English SEMCOR.

Part of Speech	English Tokens	Japanese Tokens
Verb	13,457	24,698
Noun	9,979	41,394
Adjective	10,337	2,794
Adverb	12,321	5,635

Table 2: Part of speech distribution for tokens without word alignment

4 Discussion and Future Work

We were able to transfer far fewer senses than the MULTISEMCOR (39% vs. 81%). One major reason for this is that the missing terms that this annotation project has found have not yet been added to the Japanese Wordnet. Adding them will raise the coverage by another 9%. Another reason is that we are currently overcounting untagged senses — if a word should be tagged as a multiword expression we count is once as the MWE and once for each of the single terms. However, the greatest reason is the fundamental differences between Japanese and English. There were three major causes that made transfer impossible. The first is that in many cases a word-for-word translation is unnatural — either there is a lexical gap in Japanese so that the English term does not have any translation, or the direct translation has a different connotation.

A major cause of lexical gaps is part-of-speech mismatches. For example, the English Wordnet has these three entries for *French*:[2] French_n_1 "a native or inhabitant of France"; French_n_2 "the Romance language spoken in France" and French_a_1 "of or pertaining to France or the people of France". In Japanese, the first two are productive multiword expressions *furansu-jiN* "France person" and *furansu-go* "France language" and the third is made by adding the postposition *no* "of" to either of these or just to France: *furansu-jiN-no* "French (person) lit: France person of", *furansu-go-no* "French (language) lit: France language of" and *furansu-no* "French (other) lit: France of". Because these postpositional phrases are completely compositional, it seems redundant to list them in the Japanese Wordnet. In addition, to align accurately, we would have to either separate the current adjective synset into three senses: "of or pertaining to the language of France"; "of or pertaining to the people of France" and "of or pertaining to the France" possibly with the third as the hypernym of the first two. A better approach may be to take advantage of the rich structure of the current wordnet and allow alignment between *furansu* "France" and *French* through the pertainym relation (*French_a_1* pertains-to *France_n_1*). However, currently there is no easy way to link *furansu-go* "French_n_1 (Language)" with *French_a_1*. Perhaps the proper solution is to add additional pertainym links: *French_a_1* pertains-to *French_n_1* (language) and *French_a_1* pertains-to *French_n_2* (people). Note that similar differences exist, of course, between English and Italian, but they occur far less often due to greater similarity between the two languages.

We have a rich source of new senses suggested by the translators (13,857 cases) that can be used to extend the cover of the Japanese Wordnet. For example, in Figure 2, *remove* is translated as 離す *hanasu*, even though this word was not one of the synonyms for that synset in the Japanese Wordnet. A preliminary investigation of these found that, in all cases, something had to be added to the wordnet, and in 60% of the cases the suggested translation could be used as is. The remaining cases fall into three groups (similar to those discussed above): loose translations which do not really refer to the same synset; Japanese tokens which should be part of a larger multiword expression; and translations which change the part of speech. In addition, we found some errors in the English sense tagging.

In future work, we intend to investigate techniques for efficiently correcting any remaining errors in the corpus. As much as possible, we would like to fix errors in both English and Japanese, so that we can start to carry out quantitative contrastive semantic analysis.

We would also like to investigate how ambiguities are distributed across different languages. For example, 歯 *ha* "tooth" is used for human teeth and cogwheel teeth in English, Japanese and Italian: all three languages share the same ambiguity. In general, we expect to find less ambiguity shared

[2] In addition there are two more which are not relevant to this discussion.

```
<?xml version="1.0" encoding="utf8"?>
<KAF lang="jpn">
  <kafHeader>
    <fileDesc filename="br-k01"/>
    <linguisticProcessors layer="text">
      <lp timestamp="2011-09-23T11:45:18" version="0.98" name="MeCab"/>
    </linguisticProcessors>
  </kafHeader>
  <text>
    <wf wid="w1.1.1" sent="1" para="1">スコッティ</wf>
    <wf wid="w1.1.2" sent="1" para="1">は</wf>
    <wf wid="w1.1.3" sent="1" para="1">学校</wf>
    <wf wid="w1.1.4" sent="1" para="1">に</wf>
    <wf wid="w1.1.5" sent="1" para="1">戻ら</wf>
    <wf wid="w1.1.6" sent="1" para="1">なかっ</wf>
    <wf wid="w1.1.7" sent="1" para="1">た</wf>
    <wf wid="w1.1.8" sent="1" para="1">。</wf>
  </text>
  <terms>
    <term tid="t1.1.1" lemma="スコッティ" type="open" pos="N.名詞.一般">
      <span>
        <target id="w1.1.1"/>
      </span>
      <component lemma="スコッティ" id="c1.1.1" pos="N.名詞.一般"/>
    </term>
    <term tid="t1.1.3" lemma="学校" type="open" pos="N.n">
      <span>
        <target id="w1.1.3"/>
      </span>
      <component lemma="学校" id="c1.1.3" pos="N.名詞.一般"/>
      <externalReferences>
        <externalRef resource="Wordnet jpn 1.1" reference="jpn-11-学校-n"/>
      </externalReferences>
    </term>
    <term tid="t1.1.5" lemma="戻る" type="open" pos="V.v">
      <span>
        <target id="w1.1.5"/>
      </span>
      <component lemma="戻る" id="c1.1.5" pos="V.動詞.自立"/>
      <externalReferences>
        <externalRef resource="Wordnet jpn 1.1" reference="jpn-11-戻る-v"/>
      </externalReferences>
    </term>
  </terms>
</KAF>
```

Figure 3: Sample KAF record for スコッティ は 学校 に 戻ら なかっ た 。 *Scotty ha gakkō ni modora nakat ta* ., the Japanese translation of English sentence *Scotty did not go back to school*

between languages from very different families (such as Japanese and English/Italian), but there is also extensive borrowing between English and Japanese. With a sense-tagged tritext, we can start to investigate these questions.

5 Conclusion

In this paper we described the creation of the Japanese Semantic Corpus JSEMCOR. The corpus is a translation of the English SEMCOR, with senses projected across from English. The final corpus consists of 14,169 sentences with 150,555 content words of which 58,265 are sense tagged.

References

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering*, 11(3):247–261, sep. Special Issue on Parallel Texts.

Wauter Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. 2009. KAF: a generic semantic annotation format. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon (GL 2009)*, Pisa.

Jordi Daude, Lluis Padro, and German Rigau. 2003. Validation and tuning of Wordnet mapping techniques. In *Proceedings of the International Confer-*

ence on Recent Advances in Natural Language Processing (RANLP'03), Borovets, Bulgaria.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

W. Nelson Francis and Henry Kucera, 1979. *BROWN CORPUS MANUAL*. Brown University, Rhode Island, 3 edition. (`http://khnt.aksis.uib.no/icame/manuals/brown/`).

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 230–237, Barcelona, Spain, July. Association for Computational Linguistics.

Shari Landes, Claudia Leacock, and Christiane Fellbaum. 1998. Building semantic concordances. In Fellbaum (Fellbaum, 1998), chapter 8, pages 199–216.