The Japanese FrameNet Software Tools

Hiroaki Saito, Shunta Kuboya, Takaaki Sone, Hayato Tagami, Kyoko Ohara
Keio University
3-14-1 Hiyoshi, Yokohama, 223-8522, Japan
E-mail: hxs@ics.keio.ac.jp

Abstract

This paper describes an ongoing project "Japanese FrameNet (JFN)", a corpus-based lexicon of Japanese in the FrameNet style. This paper focuses on the set of software tools tailored for the JFN annotation process. As the first step in the annotation, annotators select target sentences from the JFN corpus using the JFN kwic search tool, where they can specify cooccurring words and/or the part of speech of collocates. Our search tool is capable of displaying the parsed tree of a target sentence and its neigbouring sentences. The JFN corpus mainly consists of balanced and copyright-free "Japanese Corpus" which is being built as a national project. After the sentence to be annotated is chosen, the annotator labels syntactic and semantic tags to the appropriate phrases in the sentence. This work is performed on an annotation platform called JFNDesktop, in which the functions of labeling assist and consistency checking of annotations are available. Preliminary evaluation of our platform shows such functions accelerate the annotation process.

1. Introduction

The goal of the Japanese FrameNet (hereafter JFN) project is to create a prototype of a corpus-based lexicon of Japanese in the FrameNet style (Ohara et al., 2004, http://jfn.st.hc.keio.ac.jp/; cf. Baker, 2006; Fontenelle (ed.), 2003, http://framenet.icsi.berkeley.edu/). The resulting database will contain valence descriptions of Japanese words and a collection of annotated sentences. This paper reports on our progress in the last three years, focusing on the set of software tools tailored for the JFN process. The software tools have been designed and implemented to ensure JFN as a part of multilingual lexical resource.

2. The JFN Process

JFN contains a set of annotated sentences of lexical units which are analyzed on the basis of the semantic frames that they evoke. Each of the annotated sentences is labeled with semantic tags called Frame Elements (FEs). Here we describe how the process of building JFN proceeds (Figure 1).

The JFN process involves the following three steps. First, the annotator selects appropriate sentences to be annotated from the JFN corpus, using the JFN kwic tool (Section 3). Selected sentences are imported to the JFN database on the server machine. Then, the annotator annotates the imported sentences using the JFNDesktop (Section 4). Finally, the result of the annotation can be viewed through a Web browser and queried through FrameSQL (Sato 2008). Figure 2 shows a screenshot of Web browsing of JFN annotation.

Initially, we used two kinds of corpora to build a prototype of JFN, namely, 10-years of newspaper articles and novels. We strongly felt, however, these texts are not balanced enough for JFN annotation. In September 2006, a national project called "Japanese Corpus" was launched to build a huge balanced and copyright-free corpus. We have joined the project and have added some portions of the data to the JFN corpus. Currently available data includes government white papers and various books (http://www.tokuteicorpus.jp/). The national project also provides a morphological dictionary called UniDic (http://www.tokuteicorpus.jo/dist/). This resource is incorporated into the lemmatization table of the lexical database of the JFN database.

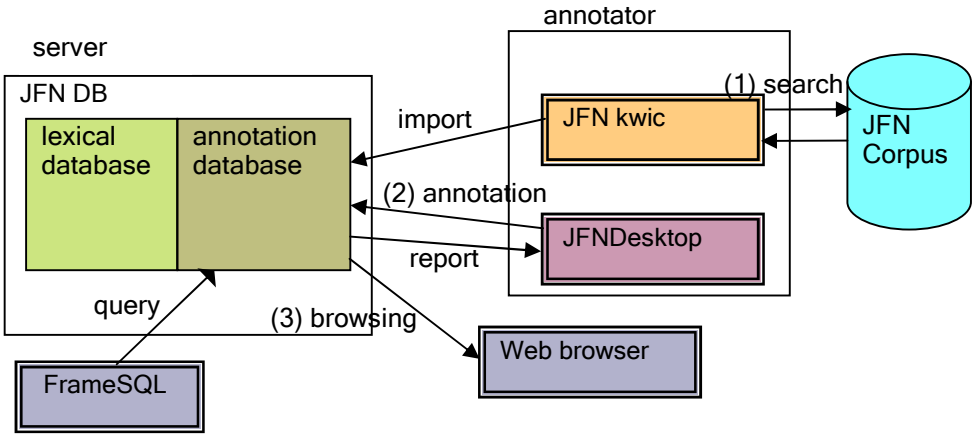

Figure 1: The JFN software tools and the process of annotation

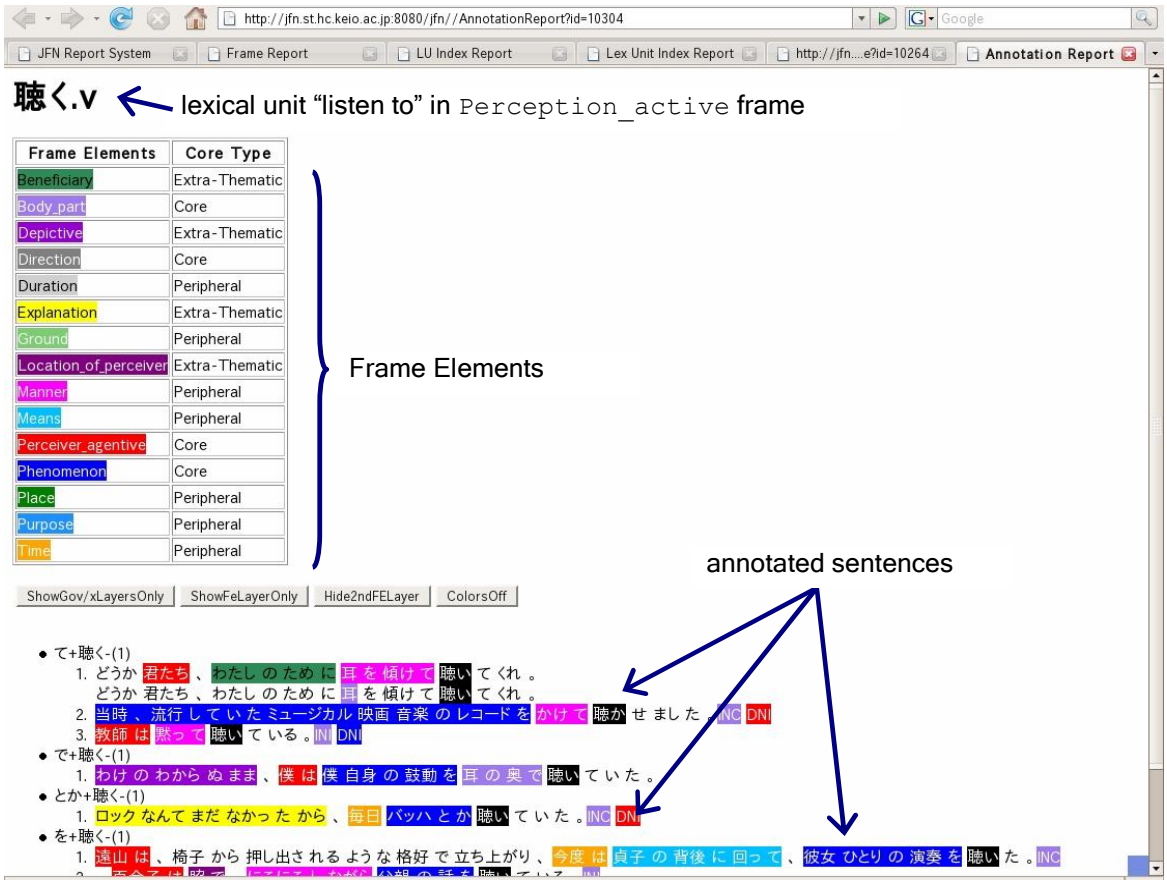

Figure 2: A screenshot of JFN web report

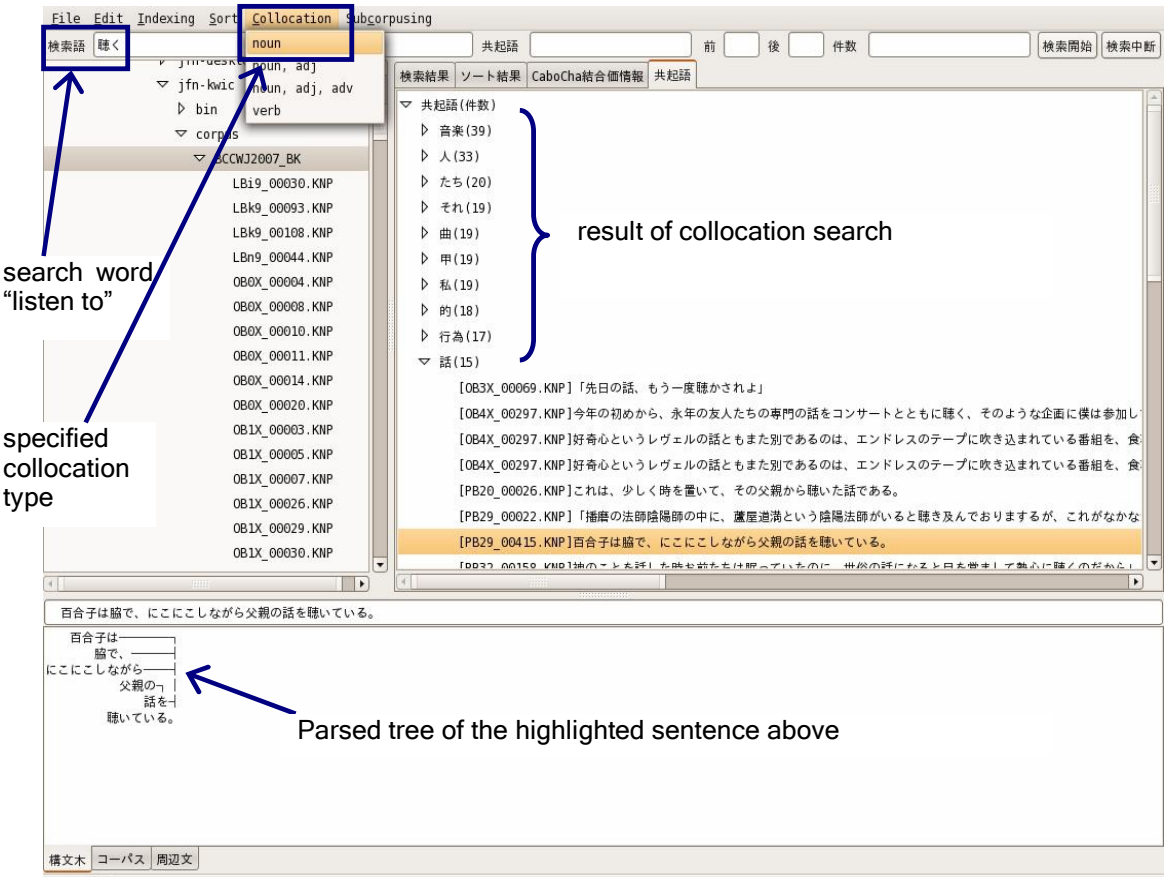

Figure 3: A screenshot of JFN kwic

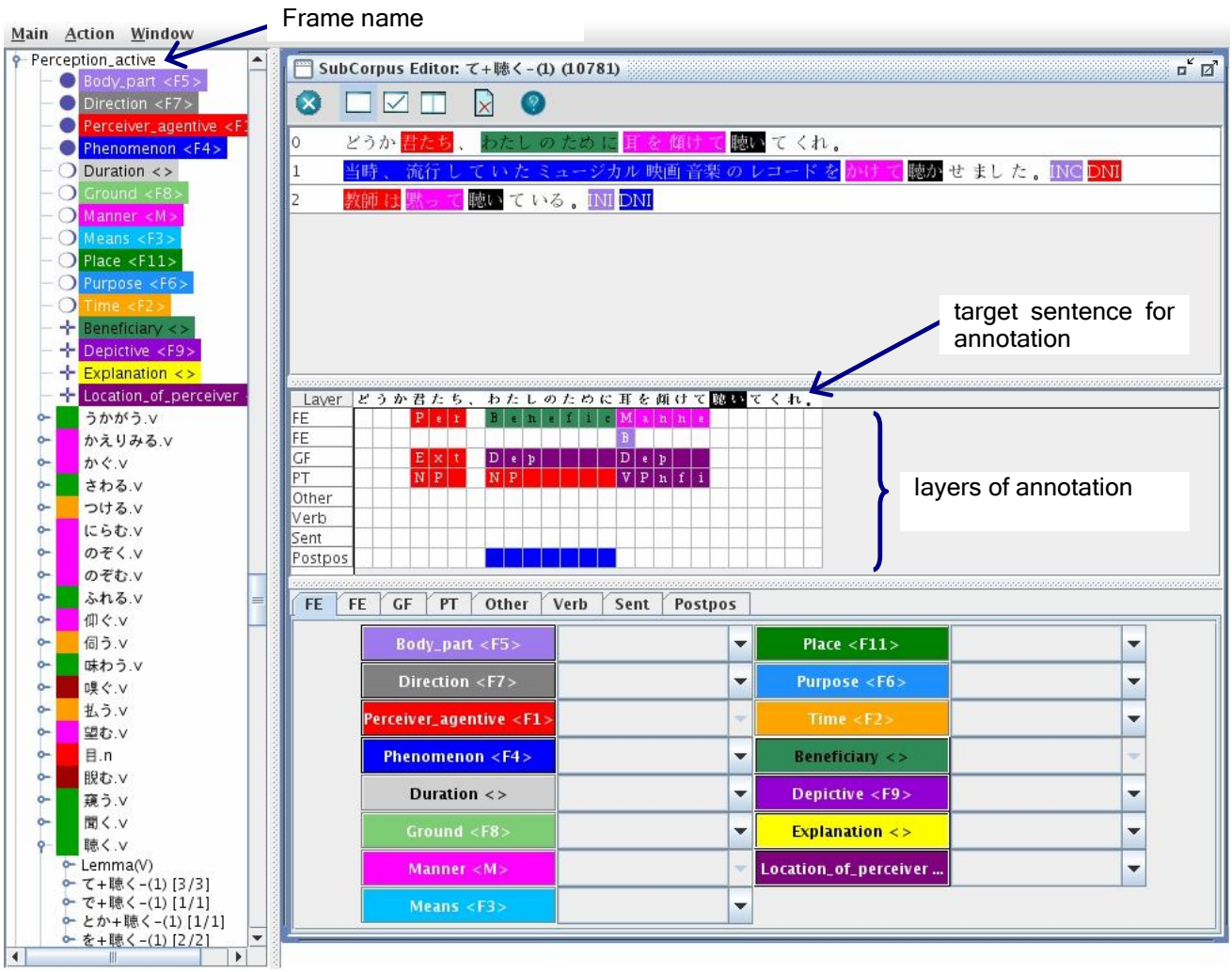

Figure 4: A screenshot of JFNDesktop

3. JFN kwic

A search tool called JFN kwic has been developed and improved in terms of functions. The tool searches for both the root form and conjugated forms of verbs, adjectives, and adjectival nouns as well as nouns. A key feature of the tool is that it can be used with sentences parsed by a dependency structure analyzer called CaboCha (http://cl.aist-nara.ac.jp/). CaboCha performs morphological analysis as well as syntactic parsing of any Japanese sentence. Figure 3 shows a snapshot of the parsed tree mode. Also, it has the function to specify in advance the number of search results and stop the search in the middle of the process when there are too many matched sentences.

A recently-added feature of JFN kwic is that the annotator can specify the context of the keyword. The context can be a specific word or a part of speech of possible collocates. In Figure 3, the annotator is searching for the verb "聴く (listen to)", which co-occurs with any noun. The search result is shown in the right window of Figure 3, where collocated nouns in the JFN corpus are listed in the order of appearance count. Current implementation allows the collocation search of noun+adj, noun+adj+adv, and verbs as well as nouns.

4. Tools for Annotation

Once the target sentences are determined, the annotator labels them through the platform called JFNDesktop, which is the Japanese version of FNDesktop in FrameNet.

As Figure 4 shows, JFNDesktop provides an annotation environment similar to that of FNDesktop. Many changes have been made, however, to handle Japanese, a non-European language. For example, in Japanese, morpheme boundaries are not evident on the surface because no space is placed in between morphemes. Thus, JFNDesktop internally holds space in between every character and allows any boundaries to be specified.

As a feature unique to JFNDesktop, annotators can consult lexical resources from there while annotating. One of the useful resources is IPAL (IPALexicon of the Japanese Language for computers) which holds syntactic and semantic descriptions of 861 fundamental verbs, 136 adjectives and 1081 nouns.

There are several layers to be annotated currently: FE (frame elements), GF (grammatical functions, e.g. Ext, Obj), PT (phrase types, e.g. NP, VPnoninfinite), PostPos (Japanese postpositions), etc. Among these layers, the last layer, namely, the PostPos layer, is unique to JFN.

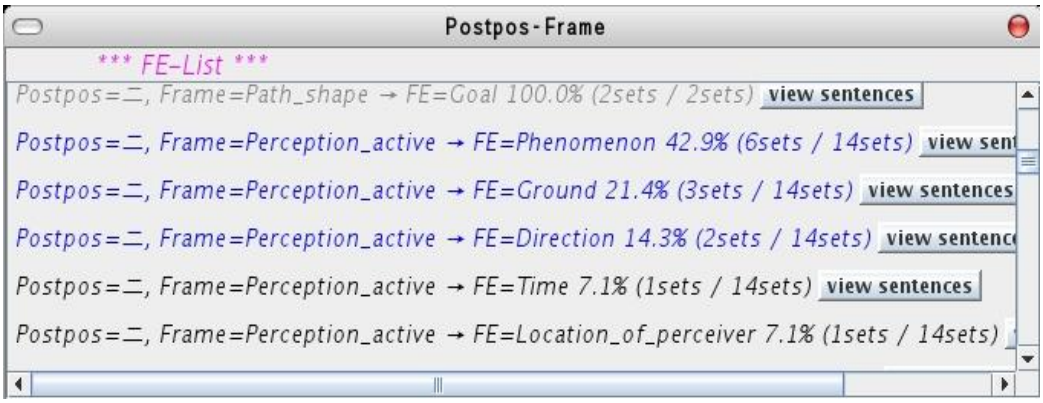

Figure 5: Viewing statistics of annotated tags

Although research on automatic labeling of FEs has been done by our collaborators, its accuracy is not high enough. Thus, the JFN annotators label the FE layer by hand. Once the FE layer is tagged, some of the other layers can be automatically labeled in JFNDesktop. The system fills the layers according to pre-defined rules, for example, "if case = 'wo' (a Japanese postposition indicating objects) then (PT = NP) and (GF = OBJ)". Annotators can overwrite the labels, if necessary.

When perusing annotated data, we have noticed some correlations across the layers. In the `Path-shape` frame, for instance, the phrase ending with the postposition 'ni' is likely to be tagged with GOAL in the FE layer. Thus, we have implemented a feature to view the statistics of annotated tags (Figure 5). Many erroneus labels were detected thanks to this feature. This feature is also useful for consistency checking of accumulated annotations.

5. Conclusion

We have reported on the progress of JFN, focusing on the software tools which facilitate manual annotation. We are improving our system to help annotators further. Our next plan is to make the statistics viewing function active, so that some probabilistic model incrementally learns the annotation result and suggests appropriate tags to the annotator.

Acknowledgements

This research is partly funded by the Grant-in-Aid for Scientific Research in Priority Areas "Japanese Corpus", Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

Baker, Collin. (2006). "Frame Semantics in Operation: The FrameNet Lexicon as an Implementation of Frame Semantics." In The Fourth International Conference on Construction Grammar Plenary Lectures. pp.34-43.

Fillmore, Charles J. (1987). "A private history of the concept 'frame'." Dirven, Rene and Radden, Gunter. (Eds). Concepts of Case. Gunter Narr Verlag, Tubingen. pp.28-36.

Fontenelle, Thierry. (Ed.). (2003). Special Issue: FrameNet and Frame Semantics. International Journal of Lexicography. Vol.16, Special Issue 3, Oxford, Oxford University Press.

Ohara, Kyoko H., Fujii, S., Ohori, T., Suzuki, R., Saito, H., Ishizaki, S. (2004) The Japanese FrameNet Project: An Introduction. LREC 2004. The Fourth international conference on Language Resources and Evaluation. Proceedings of the Satellite Workshop "Building Lexical Resources from Semantically Annotated Corpora", pp.9-11.

Sato, Hiroaki. (2008) New Functions of FrameSQL for Multilingual FrameNet. LREC 2008.