A Japanese Particle Corpus Built by Example-Based Annotation

Hiroki Hanaoka[†], Hideki Mima[‡], Jun'ichi Tsujii[†§¶]

[†]Department of Computer Science, University of Tokyo, [‡]School of Engineering, University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan
[§]School of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[¶]National Centre for Text Mining, 131 Princess Street, Manchester, M1 7DN, UK
hkhana@is.s.u-tokyo.ac.jp, mima@t-adm.t.u-tokyo.ac.jp, tsujii@is.s.u-tokyo.ac.jp

Abstract

This paper is a report on an on-going project of creating a new corpus focusing on Japanese particles. The corpus will provide deeper syntactic/semantic information than the existing resources. The initial target particle is *to* which occurs 22,006 times in 38,400 sentences of the existing corpus: the Kyoto Text Corpus. In this annotation task, an "example-based" methodology is adopted for the corpus annotation, which is different from the traditional annotation style. This approach provides the annotators with an example sentence rather than a linguistic category label. By avoiding linguistic technical terms, it is expected that any native speakers, with no special knowledge on linguistic analysis, can be an annotator without long training, and hence it can reduce the annotation cost. So far, 10,475 occurrences have been already annotated, with an inter-annotator agreement of 0.66 calculated by Cohen's kappa. The initial disagreement analyses and future directions are discussed in the paper.

1. Introduction

As well as other languages, many Japanese resources have been created. The Kyoto Text Corpus (Kurohashi and Nagao, 1997), which is here abbreviated as KTC, is one of the largest Japanese corpora annotated with parts-of-speech and dependency[1] information[2]. However, the dependency information is too coarse to fully describe Japanese linguistic phenomena that are important in advanced NLP applications. It is difficult to distinguish different types of syntactic relations, for example a complement from an adjunct, and, in certain cases, it is impossible to determine whether or not a syntactic/semantic relation between two words exists. Consider the following sentence:

- *kanozyo-ga okotte isu-de tobira-o kowasita.*
  (she-NOM) (get angry) (chair-INST) (door-ACC) (broke)
  She got angry and broke the door with a chair.

Taking the dependency shown in Figure 1, the noun phrase *kanozyo* (she) may or may not be the subject of the predicate *kowasita* (broke); there might be an omitted subject of *kowasita*, which is different from *kanozyo*. It is more natural to consider that *kanozyo* (she) is the subject for *kowasita* (broke) in this example, but it cannot be decided only using the KTC annotations.

As a source of deeper information, the NAIST Text Corpus (Iida et al., 2007) is available, which is called NTC in this paper. This corpus is annotated with predicate-argument relations and coreference information on the same texts as KTC. In the current version of NTC, the predicate-argument relations are annotated for three major argument types: *ga* (nominative), *o* (accusative) and *ni* (dative). However, even with this corpus, some phenomena still cannot be properly distinguished. Taking again the example in Figure 1, although *kanozyo* (she) is annotated as a subject of *kowasita* (broke) in NTC, *isu* (chair) can be interpreted either as a locative ("on a chair") or an instrumental ("with a chair") modifier of *kowasita* (broke), because neither KTC nor NTC discriminates different usages of the case marker *de*.

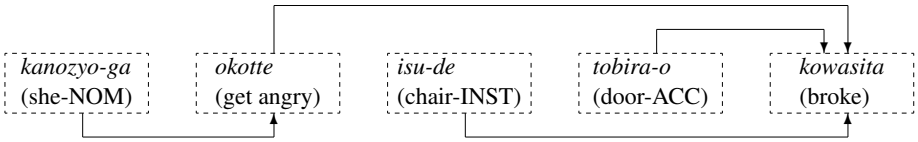
Such phenomena concerning particles, a subset of Japanese function words, are crucial for advanced NLP, because they have diverse functions (The National Institute for Japanese Language, 1951) and are used quite frequently. Therefore, it would be useful to build a corpus of Japanese particles with annotations of their syntactic/semantic functions in each occurrence. Especially, it is expected to improve such applications as machine translation or semantic role labeling, for which deep information about function words are useful.

In our corpus, particle usage is categorized and the most appropriate category is manually assigned to each occurrence in the texts of KTC. In the initial version of our corpus, the particle *to* is focused on because it is one of the most frequent particles while it has various functions. Although it is typically used as a comitative case marker, it can behave as a complementizer, or make coordinate or subordinate conjunction structures. As is the case of *de*, even given the same dependency structure, *to* can still be ambiguous. For example, *to* in Figure 2a) is a complementizer, but *to* in Figure 2b) is a subordinate conjunction. Note that they have similar structures as shown in Figure 3. The problem is exacerbated by the fact that in KTC *to* is in many cases tagged as a case particle, even when it actually has another function[3]. Since the annotations in KTC and NTC are insufficient to explain such phenomena, even though the ini-

[1]The dependencies are restricted so that each constituent only depends on another one.

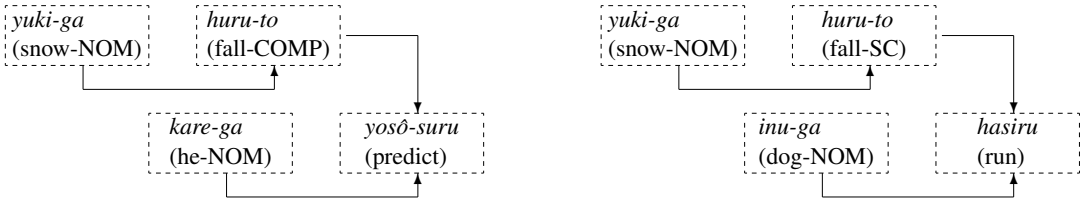
[2]In addition, predicate-argument relations are also present in KTC; however, they are only annotated in a subset (13.4%) of KTC sentences.

[3]99.9% of occurrences are annotated as a case particle. Even eliminating the sentences annotated with predicate-argument relations, more than 86% have only a case particle tag.



(She got angry and broke the door with a chair.)

Figure 1: A dependency structure of the sentence which is difficult to analyze from the KTC and NTC annotations



a) (He predicts that it will snow.)

b) (If it snows, dogs will run.)

Figure 2: An example of a) complementizer, b) conjunction in the same dependency structure

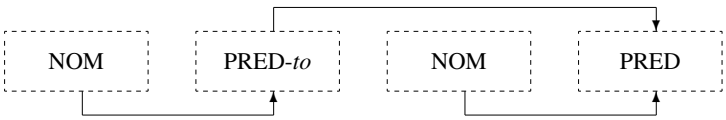


Figure 3: The dependency structure of examples in Figure 2a) and 2b)

tial version of our corpus only covers *to*, the authors believe it will nonetheless be a valuable resource for the research field of syntax/semantics.

2. Annotation methodology

In our annotation task, annotators select the most proper usage category for each occurrence of *to*, from a given set of categories. We adopt an "example-based" methodology for the corpus annotation, which is different from the traditional annotation style. In this methodology, each category is exemplified by a sentence, and annotators select an example rather than a category label. Consider the situation to classify the occurrence of *to* in the sentence:

watasi-ga kare-to asobu.
- (I-NOM) (him-COM) (play)
  I play with him.

into three categories: comitative case marker, complementizer and subordinate conjunction. Usually, an annotation tool presents these three categories as the candidates of annotation labels, and an annotator selects a label "comitative case marker" from the set of candidates. In contrast, our system presents a set of example sentences instead of category labels. For example, the following sentences are shown to the annotator as example sentences for the three categories:

watasi-ga kare-to akusyu-suru.
- (I-NOM) (him-COM) (shake hands)
  I shake hands with him.

yuki-ga huru-to kare-ga yosô-suru.
- (snow-NOM) (fall-COMP) (he-NOM) (predict)
  He predicts that it snows.

yuki-ga huru-to inu-ga hasiru.
- (snow-NOM) (fall-SC) (dog-NOM) (run)
  If it snows dogs will run.

By avoiding category labels, which are technical terms in general and hence require some expertise on Japanese linguistics to understand, it is expected that any native speakers can be an annotator without long training.

In order to accelerate the annotation, a few plausible categories are automatically suggested to annotators by a rule-based method using KTC information. In fact, this suggestion is not only for acceleration but also for implicitly providing KTC information to an annotator although KTC information is not directly offered to an annotator because it uses technical terms, which can cancel the advantages of our example-based methodology.

Since this suggestion may affect the annotation results positively and negatively, it should be implemented with discretion. We currently use a set of simple pattern-matching rules for the suggestion. The rules are based on the syntactic annotations given in KTC. In order to reduce the suggestion error, we do not discriminate difficult cases, for which the syntactic patterns do no give enough clue to pick up a single usage of a particle. For such cases, we simply suggest all matched categories. For example, it is generally easy to judge whether one occurrence be a comitative case marker or a subordinate conjunction because in the former case the previous constituent is nominal while in the latter case it is verbal. In contrast, it is not easy to distinguish complementizer from subordinate conjunction because for both usage the particle appears between two VPs.

Currently, the number of categories for *to* is 13. The category list is shown in Table 1. As it shows, the categories are hierarchically classified. The super-categories are designed based on the syntactic property of the usages, and

category		example	
coordination	nominal	*watasi-ga ringo-to momo-o taberu* (I-NOM) (apple-CONJ) (peach-ACC) (eat) I eat apples and peaches.	(1)
	predicate	*hanasu-to kiku-o dôzi-ni zissen-sita.* (speaking-CONJ) (listening-ACC) (at the same time) (put into practice) He put into practice speaking and listening at the same time.	(2)
nominal subcategorizer	complement	*watasi-ga kare-to akusyu-suru.* (I-NOM) (him-COM) (shake hands) I shake hands with him	(3)
	adjunct	*yama-to tumareta momo-o taberu.* (like a mountain) (piled up) (peach-ACC) (eat) He eats peaches piled up like a mountain.	(4)
	ellipsis	*ringo-to kodomo.* (apple-COMP) (child) The child says "apple."	(5)
predicative subcategorizer	complementizer	*ringo-ga oisî-to kotaeru.* (apple-NOM) (tasty-COMP) (reply) He replies that apples taste good.	(6)
	subordinate	*yuki-ga huru-to inu-ga hasiru.* (snow-NOM) (fall-SC) (dog-NOM) (run) If it snows dogs will run.	(7)
	adjunct	*sigoto-ga owatta-to yorokobu.* (work-NOM) (finish-PP) (rejoice) He feels happy because the work finished.	(8)
	ellipsis	*oisî-to kodomo.* (tasty-COMP) (child) The child says "it tastes good."	(9)
ending	ellipsis	*yatto owatta-to.* (finally) (finish-END) Finally finished, you mean.	(10)
	inversion	*kare-wa omotta. oisî-to.* (he-NOM) (thought) (tasty-END) He thought that it tasted good.	(11)
idiomatic	onomatopoeia	*wanwan-to inu-ga hoeru.* (bowbow) (dog-NOM) (bark) The dog barks.	(12)
	beginning	*to iuno-wa* (-COMPS) (saying-NOM) that is because	(13)

Table 1: Category list of *to* in the initial version of our corpus

the sub-categories are divided mainly based on the semantic aspects. By such a hierarchical categorization, it is expected that a plausible super-category can be selected with high precision by a rule-based suggestion mechanism based on syntactic information of KTC, and hence, annotators can focus on only a couple of sub-categories. However, such a design can arise a pseudo disagreement problem. For example, the group (5) and the group (9) are quite similar categories, which would be difficult to create a clear guideline of classification. Therefore, as for the category design, there is much room for consideration.

Because the category design is immature, there is a possibility that these categories cannot cover all of the linguistic phenomena, or the design is inadequate to let an annotator select the correct one category, i.e., ambiguous cases. Therefore, as a tentative solution, annotators are permitted to select zero or several categories for one instance. Those selections will be uniformly weighted in evaluating the agreement score shown below.

3. Annotation statistics

In the 38,400 sentences of KTC, there are 22,006 occurrences of the particle *to*. In the initial version of our corpus, we annotated *to* and its topicalized forms *towa* and *tomo*; other forms such as *tono* and *toka* are not annotated yet, whose typical functions are adnominal marker and conjunctive, respectively. As a result, 20,422 occurrences of *to*, including *towa* and *tomo*, should be annotated. The breakdown is shown in Table 2.

The number of annotators is two. One annotator has annotated over eight months with little training while the other annotator has annotated over six months with re-annotation

target	#occurrences
to	19,453
towa	570
tomo	399
total	20,422

Table 2: Number of annotation targets

data	#occurrences
annotator training	660
IAA evaluation	1,474
already annotated	10,475
total	20,422

Table 3: Current state of annotation

training: she re-annotated 660 occurrences of the particle with referring to the other annotator's result for one week. Among the total occurrences, 10,475 (51.29%) have been already annotated including the data for annotator training and inter-annotator agreement (IAA) evaluation (see Table 3). IAA statistics between the two annotators are calculated using Cohen's and Fleiss' kappa (Cohen, 1960; Fleiss, 1971):

$$\begin{aligned} Cohen's\ \kappa &= 0.660 \\ Fleiss'\ \kappa &= 0.655 \end{aligned}$$

This IAA result implies that the agreement was practically significant, though we consider there is still room for improvement.

4. Disagreement analysis

Table 4 shows the agreement matrix of *to* annotations by the annotators. Each cell integer stands for the number of *to* occurrences annotated as a corresponding category. Although the cell value can have fractional portion because the annotator selections will be uniformly weighted, those are rounded off for visibility in the table. It should be noted that the value in the sixth column cell (complementizer of predicative subcategorizer) of the third row (complement of nominal subcategorizer) is quite high: 166. This means that there are many instances difficult to distinguish between nominal and predicative. It is interesting that the value of the symmetric cell is very small: only 7. This means that the two annotators do not have an agreed criteria to discriminate nominal from predicative. It suggests that it may be difficult to completely preclude the annotator training and documented annotation guideline even with the example-based methodology.

We analyzed first 30 cases from the disagreements in the evaluation set. These 30 cases could be divided into three sets. The first set, including 11 disagreements, seems to be caused by the ambiguity among different categories exemplified by the difficulty in nominal/predicative discrimination described above. For example, in the following sentence *to* should be considered as a complementizer:

- *aite-ga* *kodomo-to* *yudan-sita.*
  (opponent-NOM) (child-COMP) (was careless)
  He was careless since (he thought that) the opponent was a child.

However, an annotator selected group (3) for nominal subcategorizer usage, while the other selected the correct category, i.e., group (6). This disagreement was presumably caused by the fact that the annotator might have misunderstood the difference between (3) and (6), e.g., she might have focused only on the word adjacent to *to*. Then, she would select group (3) because in the above example, although *aite-ga kodomo* ("the opponent was a child") is a sentential clause, *kodomo* (child) alone is a nominal. It might achieve better accuracy if we created more thorough guidelines, but it would cancel the advantages of the example-based methodology. We believe it is enough, and probably important, to carefully design examples so as to clarify the difference among categories, for example, by using the same words except at the position that one should focus on. However, it would make the category design unduly redundant or complicated. For designing examples, it could be quite useful to use "negative" examples[4]. By adding a typical error sentence as a negative example, the disagreements will be reduced without losing the example-based philosophy.

The principal cause of the second disagreement set, which includes 7 cases, seems that the category list was insufficient. For the following example:

- *momo-ga* *hyakuen-to* *yasui.*
  (peach-NOM) (one-hundred yen) (inexpensive)
  The peach is inexpensive; only one-hundred yen.

the very category does not exist in our design. Although *to* in this example is similar to both group (4) and (6), it is difficult to conclude which is more appropriate. A naive solution is to add a new category, but it may augment other disagreements by increasing similar categories. As a matter of course, there may be a case where a new category is needed, but in some situations, we can "decide" the more proper category. For instance, we can assume that the above example belongs to group (4), if the syntactic/semantic difference from (4) is unimportant. This can be achieved by creating a guideline, but it is also possible by adding an example for a category since there can be multiple examples for one category.

The remained 12 cases seem to be mainly caused by the intra-annotator inconsistency although further discussion is necessary. One of the annotators was hardly given an explanation about the design of the category set. She has only seen the annotation results by the other annotator for the training section. The lack of instruction may have caused the inconsistency in her annotation. These disagreements also imply the difficulty of eliminating the training.

5. Future directions

There are two main purposes in our work: to establish an efficient annotation methodology and to create a useful linguistic resource. As a first attempt, we proposed

[4]This idea was presented by an anonymous reviewer.

			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(*)
coordination	nominal	(1)	148	2	14	0	1	24	0	1	0	0	0	0	0	1
	predicate	(2)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
nominal subcategorizer	complement	(3)	6	0	194	6	0	166	0	0	0	0	0	0	0	0
	adjunct	(4)	1	0	10	14	0	18	2	1	0	0	0	0	0	0
	ellipsis	(5)	0	0	0	0	0	0	0	0	1	0	0	0	0	0
predicative subcategorizer	complementizer	(6)	0	0	7	0	0	623	13	4	4	0	0	0	1	0
	subordinate	(7)	0	0	0	2	0	17	103	2	0	0	0	0	0	0
	adjunct	(8)	0	0	0	0	0	5	3	0	1	1	0	0	0	0
	ellipsis	(9)	0	0	0	0	0	0	0	0	8	0	0	0	0	0
ending	ellipsis	(10)	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	inversion	(11)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
idiomatic	onomatopoeia	(12)	0	0	7	0	0	7	0	0	0	0	0	10	0	0
	beginning	(13)	0	0	0	1	0	0	0	0	0	0	0	0	1	0
other		(*)	0	0	0	0	0	0	0	1	0	0	0	0	0	0

Table 4: Agreement matrix of *to* annotations

an example-based approach. However, since experiments and evaluation for the methodology is still insufficient, we need to perform further experiments and to compare it with a traditional style annotation, conducted with an enough instruction of the category design and thorough explicit guidelines.

For a new linguistic resource, we focused on Japanese particles. There are two orthogonal future directions for the corpus. One is to annotate *to* with deeper information such as semantics. In order to obtain an even more useful resource, it may be necessary to further refine the categorization. The other direction is to annotate another frequent particle such as *mo*, which is mainly used for topicalization, but has some functions: a type of case marker, coordinate conjunctive, and, obviously, topicalization. In parallel, we need to apply our corpus to a practical application and evaluate the statistics. In addition to such direct usage as machine translation or semantic role labeling, it can be used for the corpus-oriented grammar development (Miyao, 2006). Since our corpus is currently targeting the same texts as KTC and NTC, we can obtain detailed syntactic/semantic analyses, and hence a fine-grained grammar, by combining the three corpora.

6. Conclusion

We reported an on-going project of creating a new corpus focusing on Japanese particles. The initial version of the corpus only focuses on *to*, and so far, about 50% of the occurrences have been annotated by the example-based approach. By providing an initial disagreement analysis, problems in example-based approach and their solutions were roughly indicated. As a next step, we need quantitative and qualitative evaluation of our approach by comparing it to the traditional category annotation.

7. Acknowledgements

This work was partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan). This work was also sponsored by the Center for Knowledge Structuring at the University of Tokyo.

8. References

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76(5):378–382.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations. In *ACL Workshop 'Linguistic Annotation Workshop'*, pages 132–139.

Sadao Kurohashi and Makoto Nagao. 1997. Building a Japanese Parsed Corpus while Improving the Parsing System. In *Proceedings of the NLPRS-97*, pages 451–456.

Yusuke Miyao. 2006. *From Linguistic Theory to Syntactic Analysis: Corpus-Oriented Grammar Development and Feature Forest Model*. Ph.D. thesis, The University of Tokyo.

The National Institute for Japanese Language. 1951. *Gendai-go no Zyosi, Zyodôsi: Yôhô to Ziturei (trans. The Modern Japanese Particles and Auxiliary Verbs: Usage and Example)*. Syuei Press, May.