# Reusing an ontology to generate numeral classifiers

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

**Kyonghee Paik**
Center for the Study of Language and Information
Stanford University, CA 94305-2150, USA
kpaik@usa.net

## Abstract

In this paper, we present a solution to the problem of generating Japanese numeral classifiers using semantic classes from an ontology. Most nouns must take a numeral classifier when they are quantified in languages such as Chinese, Japanese, Korean, Malay and Thai. In order to select an appropriate classifier, we propose an algorithm which associates classifiers with semantic classes and uses inheritance to list only those classifiers which have to be listed. It generates sortal classifiers with an accuracy of 81%. We reuse the ontology provided by Goi-Taikei — a Japanese lexicon, and show that it is a reasonable choice for this task, requiring information to be entered for less than 6% of individual nouns.

## 1 Introduction

In this paper we consider two questions. The first is: how to generate numeral classifiers such as *piece* in *2 pieces of paper*? To do this we use a semantic hierarchy originally developed for a different task. The second is: how far can such a hierarchy be reused?

In English, uncountable nouns cannot be directly modified by numerals, instead the noun must be embedded in a noun phrase headed by a classifier. Knowing when to do this is a language specific property. For example, French *deux renseignement* must be translated as *two pieces of information* in English.[1] In many languages, including most South-East Asian languages, Chinese, Japanese and Korean, the majority of nouns are uncountable and must be quantified by numeral classifier combinations. These languages typically have many different classifiers. There has been some work on the analysis of numeral classifiers in natural language processing, particularly for Japanese (Asahioka et al., 1990; Kamei and Muraki, 1995; Bond et al., 1996; Bond et al., 1998; Yokoyama and Ochiai, 1999), but very little on their generation. We could only find one paper on generating classifiers in Thai (Sornlertlamvanich et al., 1994). One immediate application for the generation of classifiers is machine translation, and we shall take examples from there, but it is in fact needed for the generation of any quantified noun phrase with an uncountable head noun.

The second question we address is: how far can an ontology be reused for a different task to the one it was originally designed for. There are several large ontologies now in use (WordNet (Fellbaum, 1998); Goi-Taikei (Ikehara et al., 1997); Mikrokosmos (Nirenburg, 1989)) and it is impractical to rebuild one for every application. However, there is no guarantee that an ontology built for one task will be useful for another.

The paper is structured as follows. In Section 2, we discuss the properties of numeral classifiers in more detail and suggest an improved algorithm for generating them. Section 3 introduces the ontology we have chosen, the Goi-Taikei ontology (Ikehara et al., 1997). Then we show how to use the ontology to generate classifiers in Section 4. Finally, we discuss how well it performs in Section 5.

## 2 Generating Numeral Classifiers

In this section we introduce the properties of numeral classifiers, focusing on Japanese, then give an algorithm to generate classifiers. Japanese was chosen because of the wealth of published data on Japanese classifiers and the availability of a large lexicon with semantic classes marked.

### 2.1 What are Numeral Classifiers

Japanese is a language where most nouns can not be directly modified by numerals. Instead, nouns are modified by a numeral-classifier combination as shown in (1).[2]

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

[1] Numeral-classifier combinations are shown in **bold**, the noun phrases they quantify are underlined.

[2] We use the following abbreviations: NOM = nominative; ACC = accusative; ADN = adnominal; CL = classifier; ARGSTR

(1) **2-tsū**-no denshimēru
2-CL-ADN email
**2 pieces** of email
2 emails

In Japanese, numeral classifiers are a subclass of nouns. The main property distinguishing them from prototypical nouns is that they cannot stand alone. Typically they postfix to numerals, forming a quantifier phrase. Japanese also allows them to combine with the quantifier *sū* "some" or the interrogative *nani* "what" (2). We will call all such combinations of a numeral/quantifier/interrogative with a numeral classifier a numeral-classifier combination.

(2) a. *2-hiki* "2 animals" (Numeral)
b. *sū-hiki* "some animals" (Quantifier)
c. *nan-biki* "how many animals" (Interrogative)

Classifiers have different properties depending on their use. There are five major types: **sortal** which classify the kind of the noun phrase they quantify (such as *-tsu* "piece"); **event** which are used to quantify events (such as *-kai* "time"); **mensural** which are used to measure the amount of some property (such as *senchi* "-cm"), **group** which refer to a collection of members (such as *-mure* "group"); and **taxonomic** which force the noun phrase to be interpreted as a generic kind (such as *-shu* "kind").

We propose the following basic structure for sortal classifiers (3). The lexical structure we adopt is an extension of Pustejovsky's (1995) generative lexicon, with the addition of an explicit quantification relationship (Bond and Paik, 1997).

(3) *classifier*
$$\begin{bmatrix} \text{ARGSTR} & \begin{bmatrix} \text{ARG1} & \texttt{x:numeral+} \\ \text{D-ARG1} & \texttt{y: ?} \end{bmatrix} \\ \text{QUANT} & \texttt{quantifies(x,y)} \end{bmatrix}$$

There are two variables in the argument structure: the numeral, quantifier or interrogative (represented by `numeral+`), and the noun phrase being classified. Because the noun phrase being classified can be omitted in context, it is a default argument, one which participates in the logical expressions in the qualia, but is not necessarily expressed syntactically.

= argument structure; ARG = argument; D-ARG = default argument, QUANT = quantification.

Sortal classifiers differ from each other in the restrictions they place on the quantified variable `y`. For example the classifier *-nin* adds the restriction `y:human`. That is, it can only be used to classify human referents.

Japanese has two number systems: a Sino-Japanese one based on Chinese for example, *ichi* "one",*ni* "two",*san* "three", etc., and an alternative native-Japanese system, for example, *hitotsu* "one" *futatsu* "two",*mitsu* "three", etc. In Japanese the native system only exists for the numbers from one to ten. Most classifiers combine with the Chinese forms, however, different classifiers select Sino-Japanese for some numerals, for example, *ni-hiki* "two-cl", and most classifiers undergo some form of sound change (such as *-hiki* to *-biki* in (2)). We will not be concerned with these morphological changes, we refer interested readers to Backhouse (1993, 118–122) for more discussion.

Numeral classifiers characteristically premodify the noun phrases they quantify, linked by an adnominal case marker, as in (4); or appear 'floating' as adverbial phrases, typically to before the verb: (5). The choice between pre-nominal and floating quantifiers is largely driven by discourse related considerations (Downing, 1996). In this paper we concentrate on the semantic contribution of the quantifiers, and ignore the discourse effects.

(4) **2-tsū**-no tegami-o yonda
2-CL-ADN letter-ACC read
I read two letters

(5) tegami-o **2-tsū** yonda
letter-ACC 2-CL read
I read two letters

Quantifier phrases can also function as noun phrases on their own, with anaphoric or deictic reference, when what is being quantified is recoverable from the context. For example (7) is acceptable if the letters have already been referred to, or are clearly visible.

(6) [some background with letters salient]

(7) **2-tsū**-o yonda (Japanese)
2-CL-ACC read
I read two letters

In the pre-nominal construction the relation between the target noun phrase and quantifier is explicit. For numeral-classifier combinations the

quantification can be of the object denoted by the noun phrase itself as in (8); or of a sub-part of it as in (9) (see Bond and Paik (1997) for a fuller discussion).

(8) **3-tsū**-no tegami
3-CL-ADN letter
3 letters

(9) **3-mai**-no tegami
3-CL-ADN letter
a 3 page letter

### 2.2 An Algorithm to Generate Numeral Classifiers

The only published algorithm to generate classifiers is that of Sornlertlamvanich et al. (1994). They propose to generate classifiers in Thai as follows: First create a lexicon with default classifiers listed for as many nouns as possible. This was done by automatically extracting noun classifier pairs from a sense-tagged corpus, and taking the classifier that appeared most often with each sense of a noun.[3] Then, the most frequent classifier is listed for each semantic class. Generation is then simple: if a noun has a default classifier in the lexicon, then use it, otherwise use the default classifier associated with its semantic class.

Unfortunately, no detailed results were given as to the size of the concept hierarchy, the number of nodes in it or the number of nouns for which classifiers were found. As the generation procedure was not implemented, there was no overall accuracy given for the system.

As a default, Sornlertlamvanich et al.'s algorithm is useful. However, it does not cover several exceptional cases, so we have refined it further. The extended algorithm is shown in Figure 1.

Firstly, we have made explicit what to do when a noun is a member of more than one semantic class or of no semantic class. In the lexicon we used, nouns are, on average, members of 2 semantic classes. However, the semantic classes are ordered so that the most typical use comes first. For example, *usagi* "rabbit" is marked as both `animal` and `meat`, with `animal` coming first (Figure 3). In this case, we would take the classifier associated with the first semantic class. However, in the case of *usagi* it is not counted with the default classifier for animals *-hiki*, but with that for birds *-wa*, this must be listed as an exception.

Secondly, we have added a method for generating classifiers that quantify coordinate noun phrases. These commonly appear in appositive noun phrases such as *ABC-to XYC-no 2-sha* "the two companies, ABC and XYZ".

---

1. For a simple noun phrase
   (a) If the head noun has a default classifier in the lexicon:
   use the noun's default classifier
   (b) Else if it exists, use the default classifier of the head noun's first listed semantic class (the class's default classifier)
   (c) Else use the **residual** classifier *-tsu*
2. For a coordinate noun phrase
   generate the classifier for each noun phrase
   use the most frequent classifier

Figure 1: Algorithm to generate numeral classifiers

---

In addition, we investigate to what degree we could use inheritance to remove redundancy from the lexicon. If a noun's default classifier is the same as the default classifier for its semantic class, then there is no need to list it in the lexicon. This makes the lexicon smaller and it is easier to add new entries. Any display of the lexical item (such as for maintenance or if the lexicon is used as a human aid), should automatically generate the classifier from the semantic class. Alternatively (and equivalently), in a lexicon with multiple inheritance and defaults, the class's default classifier can be added as a defeasible constraint on all members of the semantic class.

## 3 The Goi-Taikei Ontology

We used the ontology provided by Goi-Taikei — A Japanese Lexicon (Ikehara et al., 1997). We choose it because of its rich ontology, its extensive use in many other NLP applications, its wide coverage of Japanese, and the fact that it is being extended to other numeral classifier languages, such as Malay.

The ontology has several hierarchies of concepts:

[3] In fact, Thai also has a great many group classifiers, much like *herd*, *flock* and *pack* in English. Therefore each noun has two classifiers, a sortal classifier and a group classifier listed. Japanese does not, so we will not discuss the generation of group classifiers here.

with both `is-a` and `has-a` relationships. 2,710 semantic classes (12-level tree structure) for common nouns, 200 classes (9-level tree structure) for proper nouns and 108 classes for predicates. We show the top three levels of the common noun ontology in Figure 2. Words can be assigned to semantic classes anywhere in the hierarchy. Not all semantic classes have words assigned to them.

The semantic classes are used in the Japanese word semantic dictionary to classify nouns, verbs and adjectives. The dictionary includes 100,000 common nouns, 70,000 technical terms, 200,000 proper nouns and 30,000 other words: 400,000 words in all. The semantic classes are also used as selectional restrictions on the arguments of predicates in a separate predicate dictionary, with around 17,000 entries.

Figure 3 shows an example of one record of the Japanese semantic word dictionary, with the addition of the new DEFAULT CLASSIFIER field (underlined for emphasis).

Each record has an index form, pronunciation, a canonical form, part-of-speech and semantic classes. Each word can have up to five common noun classes and ten proper noun classes. In the case of *usagi* "rabbit", there are two common noun classes and no proper noun classes.

## 4 Mapping Classifiers to the Ontology

In this section we investigate how far the semantic classes can be used to predict default classifiers for nouns. Because most sortal classifiers select for some kind of semantic class, we thought that nouns grouped together under the same semantic class should share the same classifier.

We associated classifiers with semantic classes by hand. This took around two weeks. We found that, while some classes were covered by a single classifier, around 20% required more than one. For example, `1056:song` is counted only by *-kyoku* "tune", and `989:water vehicle` by only by *-seki* "ship", but the class [961:weapon] had members counted by *-hon* "long thin", *-chō* "knife", *-furi* "swords", *-ki* "machines" and more.

We show the most frequent numeral classifiers in Table 1. We ended up with 47 classifiers used as semantic classes' default classifiers. This is in line with the fact that most speakers of Japanese know and use between 30 and 80 sortal classifiers (Downing, 1996). Of course, we expect to add more classifiers at the noun level.

801 semantic classes turned out not to have classifiers. This included classes with no words associated with them, and those that only contained nouns with referents so abstract we considered them to be uncountable, such as greed, lethargy, etc.

We used the default classifiers assigned to the semantic classes to generate defeasible defaults for the noun entries in the common and technical term dictionaries (172,506 words in all). We did this in order to look at the distribution of classifiers over words in the lexicon. In the actual generation this would be done dynamically, after the semantic classes have been disambiguated. The distributions of classifiers were similar to those of the semantic classes, although there was a higher proportion counted with the residual classifier *-tsu*, and the classifier for machines *-dai*. This may be an artifact of the 70,000 word technical term dictionary. As further research, we would like to calculate the distribution of classifiers in some text, although we expect it to depend greatly on the genre.

The mapping we created is not complete because some of the semantic classes have nouns which do not share the same classifiers. We have to add more specific defaults at the noun level. As well as more specific sortal classifiers, there are cases where a group classifier may be more appropriate. For example, among the nouns counted with *-nin* there are entries such as couple, twins and so on which are often counted with *-kumi* "pair".

In addition, the choice of classifier can depend on factors other than just semantic class, for example, *hito* "people" can be counted by either *-nin* or *-mei*, the only difference being that *-mei* is more polite.

It was difficult to assign default classifiers to the semantic classes that referred to events. These classes mainly include deverbal nouns (e.g. *konomi* "liking") and nominal verbs (e.g., *benkyō* "study"). These can stand for both the action or the result of the action: e.g. *kenkyū* "a study/research". In these cases, every application we considered would distinguish between event and sortal classification in the input, so it was only necessary to choose a classifier for the result of the action.

## 5 Evaluation and Discussion

The algorithm was tested on a 3700 sentence machine translation test set of Japanese with English translations, although we only used the Japanese.[4]

[4]The test set is available at `www.kecl.ntt.co.jp/icl/mtg/resources`.

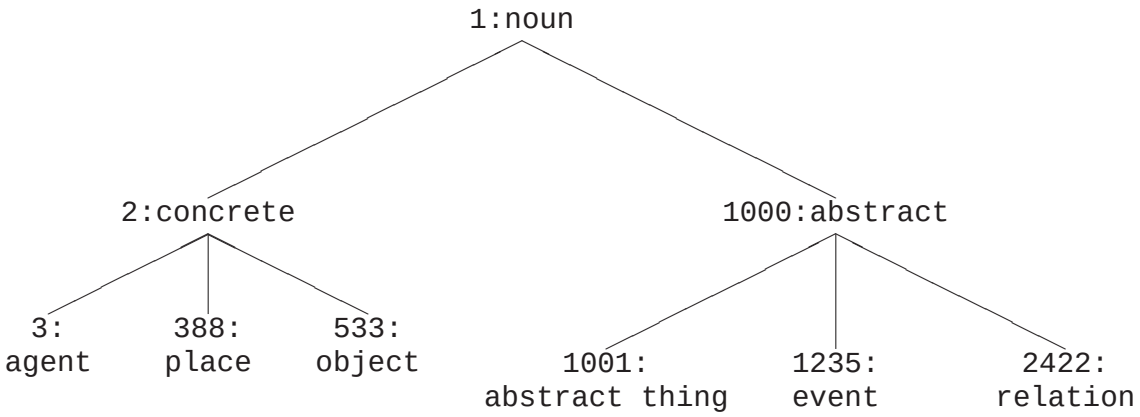


Figure 2: Top three levels of the Goi-Taikei Common Noun Ontology

record
- INDEX FORM: ウサギ(*usagi*)
- PRONUNCIATION: うさぎ/usagi/
- CANONICAL FORM: 兎(*usagi*)
- PART OF SPEECH: noun
- DEFAULT CLASSIFIER: 羽(*-wa*)
- SEMANTIC CLASSES: COMMON NOUN 537:beast, 843:meat/egg

Figure 3: Japanese Lexical Entry for *rabbit* "usagi"

We only considered sentences with a noun phrase modified by a sortal classifier. Noun phrases modified by group classifiers, such as *-soku* "pair" were not evaluated, as we reasoned that the presence of such a classifier would be marked in the input to the generator. We also did not consider the anaphoric use of numeral classifiers. Although there were many anaphoric examples, resolving them requires robust anaphor resolution, which is a separate problem. We estimate that we would achieve the same accuracy with the anaphoric examples if their referents were known, unfortunately the test set did not always include the full context, so we could not identify the referents and test this. A typical example of anaphoric use is (10).

(10) shukka-ga ruiseki-de **500-hon**-wo
shipment-NOM cumulative 500-CL-ACC
toppa-shita
reached

Cumulative shipments reached 500 ?barrels/rolls/logs/...

In total, there were 90 noun phrases modified by a sortal classifier. Our test of the algorithm was done by hand, as we have no Japanese generator. We assumed as input only the fact that a classifier was required, and the semantic classes of the head noun given in the lexicon. Using only the default classifiers predicted by the semantic class, we were able to generate 73 (81%) correctly. A classifier was only judged to be correct if it was exactly the same as that in the original test set. This was almost double the base line of generating the most common classifier (*-nin*) for all noun phrases, which would have achieved 41%. The results, with a breakdown of the errors, are summarized in Table 2.

In this small sample, 6 out of 90 (6.7%) of noun phrases needed to have the default classifier marked for the noun. In fact, there were only 4 different nouns, as two were repeated. We therefore estimate that fewer than 6% of nouns will need to have their own default classifier marked. Had the default classifier for these nouns been marked in the lexicon, our accuracy would have been 88%, the maximum achievable for our method.

| CLASSIFIER | | Referents classified | Semantic Class (2,710) | | | Noun (172,506) | |
|---|---|---|---|---|---|---|---|
| | | | No. | % | Example | No. | % |
| None | | Uncountable referents | 794 | 29.3 | `3:agent` | 34,548 | 20.0 |
| *-kai* | (回) | events | 703 | 25.9 | `1699:visit` | 35,050 | 20.3 |
| *-tsu* | (つ) | abstract/general objects | 565 | 20.9 | `2:concrete` | 52,921 | 30.1 |
| *-nin* | (人) | person | 298 | 11.0 | `5:person` | 8,545 | 4.9 |
| *-ko* | (個) | concrete objects | 124 | 4.6 | `854:edible fruit` | 14,380 | 8.3 |
| *-hon* | (本) | long thin objects | 52 | 1.9 | `673:tree` | 3,775 | 2.1 |
| *-mai* | (枚) | flat objects | 32 | 1.2 | `770:paper` | 2,807 | 1.6 |
| *-teki* | (滴) | liquid | 21 | 0.8 | `652:tear` | 1,219 | 0.7 |
| *-dai* | (台) | mechanic items/ furniture | 18 | 0.7 | `962:machinery` | 5,087 | 2.9 |
| *-hiki* | (匹) | animals | 12 | 0.6 | `537:beast` | 1,361 | 0.8 |
| Other | | 38 classifiers | 91 | 3.4 | | 12,813 | 7.4 |

Table 1: Japanese Numeral Classifiers and associated Semantic Classes

| Result | % | No. |
|---|---|---|
| Correctly generated | 81% | 73 |
| Incorrectly generated | 19% | 17 |
| Total | 100% | 90 |
| Breakdown of Errors | | |
| Noun needs default classifier | — | 6 |
| Target not in lexicon, bad entry | — | 4 |
| Other errors | — | 7 |

Table 2: Results of applying the algorithm

Looking at it from another point of view, the Goi-Taikei ontology, although initially designed for Japanese analysis, was also useful for generating Japanese numeral classifiers. We consider that it would be equally useful for the same task with Korean, or even the unrelated language Malay.

We generated the residual classifier *-tsu* for nouns not in the lexicon, this proved to be a bad choice for three unknown words. If we had a method of deducing semantic classes for unknown words we could have used it to predict the classifier more successfully. For example, *kikan-tōshika* "institutional investor"[5] was not in the dictionary, and so we used the semantic class for *tōshika* "investor", which was `175:investor`, a sub-type of `5:person`. Had *kikan-tōshika* "institutional investor" been marked as a subtype of company, or if we had deduced the semantic class from the modifier, then we would have been able to generate the correct classifier *-sha*. In one case, we felt the default ordering of the semantic classes should have been reversed: `673:tree` was listed before `854:edible fruit` for *ringo* "apple".

The remaining errors were more problematic. There was one example, *80,000-nin-amari-no shōmei* "about 80,000 signatures", which could be treated as referent transfer: *shomei* "signature" was being counted with the classifier for people. Another possible analysis is that the classifier is the head of a referential noun phrase with deictic/anaphoric reference, equivalent to *the signatures of about 80,000 people*. A couple were quite literary in style: for example *10nen-no toshi* "10 years (Lit: 10 years of years)", where the *toshi* "year" part is redundant, and would not normally be used. In two of the errors the residual classifier was used instead of the more specific default. Shimojo (1997) predicts that this will happen in expressions where the amount is being emphasized more than what is being counted. Intuitively, this applied in both cases, but we were unable to identify any features we could exploit to make this judgment automatically.

A more advanced semantic analysis may be able to dynamically determine the appropriate semantic class for cases of referent transfer, unknown words, or words whose semantic class can be restricted by context. Our algorithm, which ideally generates the classifier from this dynamically determined semantic class allows us to generate the correct classifier **in context**, whereas using a default listed for a noun does not. This was our original motivation for generating classifiers from semantic classes, rather than using a classifier listed with each noun as Sornlert-

[5]Institutional investors are financial institutions that invest savings of individuals and non-financial companies in the financial markets.

lamvanich et al. (1994) do.

In this paper we have concentrated on solving the problem of generating appropriate Japanese numeral classifiers using an ontology. In future work, we would like to investigate in more detail the conditions under which a classifier needs to be generated.

## 6 Conclusion

In this paper, we presented an algorithm to generate Japanese numeral classifiers. It was shown to select the correct sortal classifier 81% of the time. The algorithm uses the ontology provided by Goi-Taikei, a Japanese lexicon, and shows how accurately semantic classes can predict numeral classifiers for the nouns they subsume. We also show how we can improve the accuracy and efficiency further through solving other natural language processing problems, in particular, referent transfer, anaphor resolution and word sense disambiguation.

**Acknowledgments**

The authors thank Kentaro Ogura, Timothy Baldwin, Virach Sornlertlamvanich and the anonymous reviewers for their helpful comments.

## References

Yoshimi Asahioka, Hideki Hirakawa, and Shin-ya Amano. 1990. Semantic classification and an analyzing system of Japanese numerical expressions. *IPSJ SIG Notes 90-NL-78*, 90(64):129–136, July. (in Japanese).

A. E. Backhouse. 1993. *The Japanese Language: An Introduction*. Oxford University Press.

Francis Bond and Kyonghee Paik. 1997. Classifying correspondence in Japanese and Korean. In *3rd Pacific Association for Computational Linguistics Conference: PACLING-97*, pages 58–67. Meisei University, Tokyo, Japan.

Francis Bond, Kentaro Ogura, and Satoru Ikehara. 1996. Classifiers in Japanese-to-English machine translation. In *16th International Conference on Computational Linguistics: COLING-96*, pages 125–130, Copenhagen, August. (http://xxx.lanl.gov/abs/cmp-lg/9608014).

Francis Bond, Daniela Kurz, and Satoshi Shirai. 1998. Anchoring floating quantifiers in Japanese-to-English machine translation. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics: COLING/ACL-98*, pages 152–159, Montreal, Canada.

Pamela Downing. 1996. *Numeral Classifier Systems, the case of Japanese*. John Benjamins, Amsterdam.

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

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Goi-Taikei — A Japanese Lexicon*. Iwanami Shoten, Tokyo. 5 volumes/CDROM.

Shin-ichiro Kamei and Kazunori Muraki. 1995. An analysis of NP-like quantifiers in Japanese. In *First Natural Language Processing Pacific Rim Symposium: NLPRS-95*, volume 1, pages 163–167.

Sergei Nirenburg. 1989. KBMT-89 — a knowledge-based MT project at Carnegie Mellon University. pages 141–147, Aug. 16–18.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

Mitsuaki Shimojo. 1997. The role of the general category in the maintenance of numeral classifier systems: The case of *tsu* and *ko* in Japanese. *Linguistics*, 35(4). (http://ifrm.glocom.ac.jp/doc/s01.001.html).

Virach Sornlertlamvanich, Wantanee Pantachat, and Surapant Meknavin. 1994. Classifier assignment by corpus-based approach. In *15th International Conference on Computational Linguistics: COLING-94*, pages 556–561, August. (http://xxx.lanl.gov/abs/cmp-lg/9411027).

Shoichi Yokoyama and Takeru Ochiai. 1999. Aimai-na sūryōshi-o fukumu meishiku-no kaisekihō [a method for analysing noun phrases with ambiguous quantifiers.]. In *5th Annual Meeting of the Association for Natural Language Processing*, pages 550–553. The Association for Natural Language Processing. (in Japanese).