Classifiers in Japanese-to-English Machine Translation

Francis Bond and **Kentaro Ogura** and **Satoru Ikehara**
NTT Communication Science Laboratories
1-2356 Take, Yokosuka-shi, Kanagawa-ken, JAPAN 238-03
bond@nttkb.ntt.jp

Abstract

This paper proposes an analysis of classifiers into four major types: UNIT, METRIC, GROUP and SPECIES, based on properties of both Japanese and English. The analysis makes possible a uniform and straightforward treatment of noun phrases headed by classifiers in Japanese-to-English machine translation, and has been implemented in the MT system **ALT-J/E**. Although the analysis is based on the characteristics of, and differences between, Japanese and English, it is shown to be also applicable to the unrelated language Thai.

1 Introduction

Noun phrases in Japanese differ from those in English in two important ways. First, Japanese has no equivalent syntactic category to English determiners. Second, there is no grammatical marking of number.[1] Because of these differences, numerical expressions are realized very differently in Japanese and English. In English, countable nouns can be directly modified by a numeral: *2 dogs*. In Japanese, however, numerals cannot directly modify common nouns, instead a classifier is used, in the same way that a partitive noun is used with an uncountable noun in English: *2 pieces of furniture*. In addition, when Japanese is translated into English, the selection of appropriate determiners, such as articles and possessive pronouns, and the determination of countability and number is problematic.

Various solutions to the problems of generating articles and possessive pronouns and determining countability and number have been proposed (Murata and Nagao, 1993; Cornish, Fujita, and Sugimura, 1994; Bond, Ogura, and Kawaoka, 1995). The differences between the way numerical expressions are realized in Japanese and English has been less studied (Asahioka, Hirakawa, and Amano, 1990). In this paper we propose an analysis of classifiers based on properties of both Japanese and English. Our category of classifier includes both Japanese *josūshi* 'numeral classifiers' and English partitive nouns. We divide classifiers into four major types: UNIT, METRIC, GROUP and SPECIES. UNIT classifiers are further divided into GENERAL, TYPICAL and SPECIAL, while METRIC classifiers are divided into MEASURE and CONTAINER classifiers. Although our analysis was based on the characteristics of, and differences between, Japanese and English, we found it to be strikingly similar to the analysis for Thai proposed by Sornlertlamvanich et al. (1994), which suggests that the results may be useful for examining other languages.

The analysis introduced in this paper has been implemented in NTT Communication Science Laboratories' Japanese-to-English machine translation system **ALT-J/E** (Ikehara et al., 1991; Ogura et al., 1993) since 1994. Examples of how it has been implemented in **ALT-J/E** are woven throughout the text, although the analysis itself is not tied to any formalism or particular representation, so is adaptable to any system.

We start off by examining monolingual analyses of Japanese classifiers and English partitive expressions (Section 2). Then we introduce our bilingual analysis of classifiers and show how this analysis can be used in a Japanese-to-English machine translation system (Section 3). We also examine more complex cases where classifiers are used like normal nouns (Section 4). Finally we compare our analysis to other people's (Section 5).

Throughout the paper we use the following abbreviations: A, B or N: noun or noun phrase; C: classifier, X: Numeral, with Japanese in italics.

2 Monolingual Analyses of Classifiers

2.1 Japanese 'Classifiers'

Japanese is a numeral classifier language (Allan, 1977), in which classifiers are obligatory in many expressions of quantity. We will refer to prototypical Japanese classifiers as *josūshi* 'numerical classifiers'.

Syntactically, *josūshi* are a subclass of nouns (Miyazaki, Shirai, and Ikehara, 1995). The main property distinguishing them from normal nouns is that they can postfix to numerals, the quantifier *su* 'some' or the interrogative *nani* 'what', to form a noun phrase. Unlike normal nouns in Japanese,

[1] Japanese does not have contrasting singular and plural forms of nouns.

josūshi can not form grammatical noun phrases on their own.[2]

(1) *2-hiki* '2 animals' (Numeral)
(2) *sū-hiki* 'some animals' (Quant.)
(3) *nan-hiki* 'how many animals' (Int.)

The resulting numeral-classifier noun phrase can modify another noun phrase, either linked by *no* 'of' *'XC-no-N'*, or 'floating' elsewhere in the sentence, typically directly after the noun phrase it modifies *'NXC'*. It can also occur on its own, with anaphoric or deictic reference. Asahioka, Hirakawa, and Amano (1990) identify seven different patterns of use. In order to concentrate on the translation of classifiers and number, we will restrict our discussion to noun phrases of the type *'XC-no-N'* and not discuss the problems of resolving anaphoric reference and floating quantifiers.

Semantically, each classifier relates to a class of nouns (Kuno, 1973, 25), often fairly arbitrarily. For example *-hiki* '(small) animal' is used to count small animals excluding rabbits, which are counted with *-wa* 'bird'. There is a default classifier *-tsu* 'piece' which can be used to count almost anything.

2.2 English 'Classifiers'

In English, numerals can directly modify countable nouns 'X N'. In order to enumerate uncountable nouns, either the uncountable nouns have to be reclassified as countable nouns, or embedded in a partitive construction: *two beers* or *two cans of beer* 'X N' or 'X C *of* N' (Quirk et al., 1985, 249). This partitive construction is similar to the Japanese quantifying construction *'XC-no-N'*.

Quirk et al. (1985, 249–51) divide partitive nouns into three main categories QUALITY PARTITIVES, QUANTITY PARTITIVES, and MEASURE PARTITIVES. QUANTITY PARTITIVES are further divided into three cases, the first where the embedded noun phrase is uncountable, the second where it is plural, and the third where it is singular and countable. All the partitive nouns themselves are fully countable.

QUANTITY PARTITIVES where the embedded noun phrase is headed by an uncountable noun, the first case, are then divided into GENERAL PARTITIVES such as *piece* which serve only to quantify and TYPICAL PARTITIVES such as *grain* which are more descriptive.

[2]There are some examples of words that can be either a common noun or *josūshi*: for example *gyō* 'line' or *hako* 'box', which can follow a numeral or stand alone. These nouns can be handled in two ways: (a) as a lexical class that combines the properties of common nouns and *josūshi*, or (b) as two separate lexical entities. **ALT-J/E** follows option (b), such nouns are entered into the lexicon twice, once as a common noun and once as a *josūshi*.

3 A Bilingual Analysis of classifiers

As there is no direct fit between English and Japanese, it is necessary to categorize the Japanese and English classifiers and to define rules which will enable effective machine translation. We divide classifiers into four major types: UNIT (Section 3.1), METRIC (Section 3.2), GROUP (Section 3.3) and SPECIES (Section 3.4). The main criteria for the analysis are the restrictions placed, in English, on the countability and number of the embedded noun phrase in a partitive construction. Whether a noun is a classifier, and if so which type, is marked in the lexicon for each Japanese/English noun pair.

We distinguish between five major different noun countability preferences, based on the analysis of Allan (1980), adapted for use in machine translation by Bond, Ogura, and Ikehara (1994). 'Fully countable' nouns, such as *knife*, have both singular and plural forms, and cannot be used with determiners such as *much*. 'Uncountable' nouns, such as *furniture*, have no plural form, and can be used with *much*. Between these two extremes are nouns such as *cake*, which can be used in both countable and uncountable noun phrases. They have both singular and plural forms, and can also be used with *much*. We divide such nouns into two groups: 'strongly countable', those that are more often used to refer to discrete entities, such as *cake*, and 'weakly countable', those that are more often used to refer to unbounded referents, such as *beer*. The fifth major type of countability preference is 'pluralia tanta': nouns that only have a plural form, such as *scissors*.

3.1 Unit classifiers

UNIT classifiers are the prototypical classifiers. A UNIT classifier will be realized in Japanese as a *josūshi*. However, there are three possible translations of a Japanese noun phrase of the form *'XC-no-N'*, where *C* is a unit classifier:

Individuate: Translate as 'X N', where the classifier *C* is not translated and the numeral directly modifies the countable English noun phrase:
1-hiki-no-inu '1-piece of dog' → *1 dog.*

Part: Translate as 'X C *of* N', where the classifier is translated by its translation equivalent (from the transfer dictionary) and N is uncountable (headed by a bare singular noun):
1-tsubu-no-kome '1-grain of rice'
→ *1 grain of rice.*

Default: Translate as 'X C *of* N' where the classifier is replaced by a default that depends on the embedded noun and N is uncountable. The default is normally *piece*, but this can be over-ridden by an explicit entry for N's default classifier in the lexicon:

Table 1: Unit Classifiers

Noun Type	General	Typical	Special
Fully Countable	1 dog	1 dog	1 slice of dog
Strongly Countable	1 cake	1 crumb of cake	1 slice of cake
Weakly Countable	1 hair	1 strand of hair	1 slice of hair
Uncountable	1 piece of information	1 grain of information	1 slice of information
Pluralia Tanta (pair)	1 pair of scissors	1 pair of scissors	—

1-tsu-no-kagu '1-piece of furniture'
→*1 piece of furniture.*

The three types of UNIT classifier are summarized in Table 1.[3]

Having established three possible translations of the *'XC-no-N'* construction, we can proceed to divide UNIT classifiers into three types, depending on which of the above alternatives is most suitable. The first, GENERAL classifiers, are those that have no special meaning of their own, but are used only to quantify the denotation of a noun. Typical examples are - *tsu* 'piece' and *-ko* 'piece'. If N is fully, strongly or weakly countable, then the classifier is not translated (individuate). If N is uncountable, then the classifier is translated as the default (default). The second type of classifier, TYPICAL, consists of those classifiers which are descriptive in their own right, such as *-teki* 'drop'. If N is fully countable, then the classifier will not be translated (individuate), otherwise the classifier is translated (part). The final type of classifier, SPECIAL, is rare: classifiers which force an uncountable interpretation of even countable nouns, for example *-kire* 'slice'. N is always **part**ed: *1-kire-no-inu* '1-slice of dog' →*1 slice of dog.*

The translation of classifiers is complicated by the fact that classifiers and their relationships to nouns are both arbitrary and language dependent. Consider the Japanese classifier *-mai* 'sheet', which is used for counting flat objects. This has no direct English equivalent. As a default, it is entered in the dictionary as a GENERAL classifier with the translation *piece*. There are however several flat objects for which *piece* is inappropriate in English: food-stuffs (*slice*); paper, glass, cloth and leather (*sheet*); bacon (*rasher*); and financial contracts (*contract*). The selection of an appropriate translation is not dependent on this analysis and can be left to the normal machine translation process. In **ALT-J/E** it is done by examining the semantic category of the embedded noun. Once an appropriate translation of the classifier has been found, knowledge of its type allows the system to decide the appropriate form of the final translation.

[3]If N's countability preference is pluralia tanta then N will never be **individuat**ed. If N is **part**ed or **default**ed there are two possibilities: either, if the dictionary entry for N has the default classifier *pair* then it will be used as the classifier or, if N has no default classifier, then a different translation is searched for in the dictionary and used instead. If there is no non-pluralia tanta translation equivalent, then the translation will default to 'X C *of* N' as above, but with N headed by a bare plural noun.

3.2 Metric classifiers

The next overall category is METRIC classifiers. A noun phrase of the form *'XC-no-N'*, where C is a METRIC classifier will be translated as 'X C *of* N', where N will be plural if it is headed by a fully countable or pluralia tanta noun. We further subdivide METRIC classifiers depending on whether the resulting English noun phrase will have singular verb agreement (MEASURE classifiers), or plural verb agreement (CONTAINER classifiers) as its default.

(4) *2-kg-no-kami-ha jūbun da* '2 kg of paper-TOP enough is' → *2 kg of paper is enough*

(5) *2-hako-no-kami-ha jūbun da* '2 box of paper-TOP enough is' → *2 boxes of paper are enough*

In fact both (4) and (5) could be translated with singular or plural verb agreement. The differentiation into MEASURE and CONTAINER provides a graceful default. Examples are given in Table 2.

3.3 Group classifiers

GROUP classifiers combine with plural or uncountable noun phrases to make a countable noun phrase representing a group or set. A noun phrase of the form *'XC-no-N'*, where C is a GROUP classifier will be translated as 'X C *of* N', where N will be plural if it is headed by a fully or strongly countable noun or a pluralia tanta. Noun phrases of the form *'N-no-C'*, where C is a GROUP classifier (but not a *josūshi*) will also be translated as 'C of N' where N will be plural if it is headed by a fully or strongly countable noun or a pluralia tanta. This allows us to give a uniform treatment of noun phrases such as (6) and (7) during English generation, even though their Japanese structure is very different.

(6) *2-hako-no-pen* '2 box of pen'
→ *2 boxes of pens* *'XC-no-N'*

(7) *pen-no-hako* 'box of pen'
→ *a box of pens* *'N-no-C'*

Table 2: Container and Measure Classifiers

Noun Type	Container	Measure
Fully Countable	1 box of dogs	1 kg of ants
Strongly Countable	1 box of cake	1 kg of cake
Weakly Countable	1 box of beer	1 kg of beer
Uncountable	1 box of furniture	1 kg of furniture
Pluralia Tanta	1 box of scissors	1 kg of scissors

Table 3: Group and Species Classifiers

Noun Type	Group	Species (Si)	Species (Pl)
Fully Countable	1 set of dogs	1 kind of dog	2 kinds of dogs
Strongly Countable	1 set of cakes	1 kind of cake	2 kinds of cakes
Weakly Countable	1 set of beer	1 kind of beer	2 kinds of beer
Uncountable	1 set of information	1 kind of information	2 kinds of information
Pluralia Tanta	1 set of scissors	1 kind of scissors	2 kinds of scissors

Whether a noun is a GROUP classifier or not can also be used to help determine the number of ascriptive and appositive noun phrases. For example, in **ALT-J/E** the countability and number of two appositive noun phrases are made to match each other, unless one element is plural and the other is a GROUP classifier. For example, *many insects, a whole swarm, ...* as opposed to *many insects, bees I think, ...* (Bond, Ogura, and Kawaoka, 1995). Examples of GROUP classifiers are given in Table 3.

3.4 Species classifiers

The last type of classifier is SPECIES classifiers. SPECIES classifiers are partitives of quality and can occur with countable or uncountable noun phrases. The embedded noun phrase will agree in number with the head noun phrase if fully or strongly countable: *a kind of car, 2 kinds of cars; a kind of equipment, 2 kinds of equipment.* Examples of SPECIES classifiers are given in Table 3.

4 When is a Classifier a Classifier?

In the analysis given above for Japanese noun phrases of the form *'XC-no-N'*, we have given no consideration to the denotation of N, except for when choosing the appropriate translation for C. Thus we assume that *'XC-no-N'* will be translated as 'X C *of* N' or just 'X N' if N is countable, as in (8) or (9).

(8) *1-pai-no mizu* '1-cup of water'
→ *1 cup of water* (CONTAINER)

(9) *1-tsu-no koppu* '1-piece of cup'
→ *1 cup* (GENERAL)

However if N is a noun that denotes an attribute, such as PRICE or WEIGHT, then the translation process becomes more complicated. In the simplest case the noun phrase *'XC-no-N'* should be translated as though the classifier were a normal noun, giving '*the* N *of* X C', for example (10), (11).

(10) *1-pai-no nedan* '1-cup of price'
→ *the price of 1 cup*

(11) *1-tsu-no nedan [-ha 10en da]* '1-piece of price [-TOP 10 yen is]'
→ *the price of 1 (thing) [is 10 yen]*

In other words, if N has the attribute AMOUNT then the noun phrase should normally be translated as though C were not a classifier. The interpretation of C is, however, ambiguous. C could be used as a classifier with the amount N in its scope (12), or C could have anaphoric reference (13). **ALT-J/E** chooses the interpretation shown in example (13) as its default.

(12) *1-shū-no nedan* '1 kind of price'
→ *1 kind of price*

(13) *1-shū-no nedan* '1 kind of price'
→ *the price of 1 kind [of something]*

Further, when N is an attribute and C measures the same attribute, the interpretation is again different. For example, if C measures N's attribute then the resulting noun phrase will be indefinite by default: *a height of 10m* or *a price of 10 yen.* However if the noun phrase is used ascriptively then it should be converted either to an adjective *it is 10m high* or a prepositional phrase *it is 10 yen in price.* Finally, if a noun phrase of this type is used to modify another noun then it needs to be converted to an adjective *a 10m high building* or a post modifying prepositional phrase *a chocolate 10 yen in price.*

The combinations of nouns and classifiers mentioned above can all be translated by the machine translation system **ALT-J/E** using the analysis of classifiers presented in this paper in combination with a semantic hierarchy of 2,800 categories common to all nouns, as described in Ikehara et al. (1991). The particle *no* 'of', has many possible interpretations, Shimazu, Naito, and Nomura (1987) identify five main types of *A-no-B* expressions, and some 80

Table 4: Proposed Analysis of Classifiers

Classifier type		Example		Japanese POS	English Restriction on embedded NP
Unit	General	*-tsu*	'piece'	*josūshi*	Default classifier if uncountable head, no classifier if countable
	Typical	*-tsubu*	'grain'	*josūshi*	Translate classifier if uncountable, no classifier if countable
	Special	*-kire*	'slice'	*josūshi*	Translate classifier, force head to be uncountable
Metric	Measure	*-inchi*	'inch'	*josūshi*	Plural if possible, singular agreement
	Container	*hako*	'box'	noun/*josūshi*	Plural if possible, normal agreement
Group		*mure*	'group'	noun/*josūshi*	Plural if possible
Species		*shurui*	'kind'	noun/*josūshi*	Number agrees if possible

Table 5: A comparison of different analyses

Proposed Analysis		Quirk et al.	Kamei et al.	Sornlertlamvanich et al.
Unit	General	Quantity-General	Piece	Unit
	Typical	Quantity-Typical		
	Special			
Metric	Measure	Measure	Unit	Metric
	Container		Container	
Group		Quantity-Plural	Set	Collective
Species		Quality	Kind	
(Unit)			Times	Frequency
(Unit)				Verbal

sub types. Our analysis cuts across Shimazu et al.'s types, including at least three of the subtypes, and also makes clear some relations that are not explicitly named.

5 Comparisons with other Analyses

We summarize our analysis of classifiers in Table 4. Our analysis was based mainly on the properties of the generated English, so is naturally quite close to the division of partitive nouns proposed by Quirk et al. (1985). The analysis is also quite close to those proposed by Kamei and Muraki (1995) for Japanese and Sornlertlamvanich et al. (1994) for Thai. This supports Allan's (1977) assertion that "diverse language communities categorize perceived phenomena in similar ways". The different analyses are compared in Table 5.

We make the distinction between classifiers of frequency and other UNIT classifiers by using our general semantic hierarchy. Sornlertlamvanich et al.'s VERBAL classifiers "any classifier which is derived from a verb [...] /kraad haa muan/ 'five rolls of paper'." can be included in the METRIC category, although it may be the case that they have a different part of speech in Thai. Kamei and Muraki (1995) put UNIT classifiers into two classes: 'Counting Total Amount': *3kg of sugar* and 'Counting an Attribute Value': *a speed of 60mph*. This distinction belongs to the interpretation of the classifier in context, rather than its inherent properties, so we feel the distinction should be made during processing, as described in Section 4, rather than as part of the analysis of the classifiers themselves.

6 Conclusion

In this paper we present an analysis of classifiers, suitable for use in a Japanese-to-English machine translation system. We divide classifiers into four major types: UNIT, METRIC, GROUP and SPECIES. UNIT classifiers are further divided into GENERAL, TYPICAL and SPECIAL, while METRIC classifiers are divided into MEASURE and CONTAINER classifiers. The analysis is based on characteristics peculiar to Japanese and English, as well as the differences between them. The resulting analysis is shown to be similar to one proposed for Thai, an unrelated language, suggesting that it may be more widely applicable.

The analysis has been implemented in NTT's Japanese-to-English machine translation system **ALT-J/E** since 1994. It makes possible a uniform and straightforward treatment of noun phrases headed by classifiers.

Further work remains to be done in examining the distribution of classifiers in different domains, and possibly identifying classifiers automatically.

References

Allan, Keith. 1977. Classifiers. *Language*, 53:285–311.

Allan, Keith. 1980. Nouns and countability. *Language*, 56(3):541–67.

Asahioka, Yoshimi, 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).

Bond, Francis, Kentaro Ogura, and Satoru Ikehara. 1994. Countability and number in Japanese-to-English machine translation. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING '94)*, pages 32–38, August. (cmp-lg/9511001).

Bond, Francis, Kentaro Ogura, and Tsukasa Kawaoka. 1995. Noun phrase reference in Japanese-to-English machine translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI '95)*, pages 1–14, July. (cmp-lg/9601008).

Cornish, Tim, Kimikazu Fujita, and Ryochi Sugimura. 1994. Towards machine translation using contextual information. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING '94)*, pages 51–56, August.

Ikehara, Satoru, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. 1991. Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–. In *Proceedings of MT Summit III*, pages 101–106. (cmp-lg/9510008).

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

Kuno, Susumu. 1973. *The Structure of the Japanese Language*. MIT Press, Cambridge, Massachusetts, and London, England.

Miyazaki, Masahiro, Satoshi Shirai, and Satoru Ikehara. 1995. A Japanese syntactic category system based on the constructive process theory and its use. *Journal of Natural Language Processing*, 2(3):3–25, July. (in Japanese).

Murata, Masaki and Makoto Nagao. 1993. Determination of referential property and number of nouns in Japanese sentences for machine translation into English. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI '93)*, pages 218–25, July.

Ogura, Kentaro, Akio Yokoo, Satoshi Shirai, and Satoru Ikehara. 1993. Japanese to English machine translation and dictionaries. In *Proceedings of the 44th Congress of the International Astronautical Federation*, Graz, Austria.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, Essex.

Shimazu, Akira, Shozo Naito, and Hirosato Nomura. 1987. Semantic structure analysis of Japanese noun phrases with adnominal particles. In *Proceedings of the 25th Annual Meeting of the ACL*, pages 123–130. Association for Computational Linguistics.

Sornlertlamvanich, Virach, Wantanee Pantachat, and Surapant Meknavin. 1994. Classifier assignment by corpus-based approach. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING '94)*, pages 556–561, August.