# When and How to Disambiguate?
# — Countability in Machine Translation —

Francis BOND
**NTT Communication Science Laboratories**
1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa-ken, JAPAN 239
bond@cslab.kecl.ntt.co.jp

Satoru IKEHARA
**Tottori University**
4-101 Minami, Koyama-chou, Tottori-shi, JAPAN 680
ikehara@ike.tottori-u.ac.jp

*MIDDIM-96 Seminar, Le Col de Porte, August 1996*

## Abstract

Potentially all Japanese nouns are under-specified with respect to number and countability. To ask a user to interactively disambiguate each one is not feasible, and it can be impossible for a native speaker to do. Therefore it is important to disambiguate automatically. In addition, even if the system makes a mistake about number, it is rarely fatal. There are some ambiguities involved with the use of classifiers that are more likely to be serious and require world knowledge such as *500-gm-no ringo* which could mean either *a 500gm apple* or *500gm of apple(s)*. It may be appropriate to ask for user input to disambiguate such examples, although it may be preferable to query a knowledge base first.

## Keywords

Automatic Disambiguation, Interactive Disambiguation, Classifiers, Countability, Japanese-to-English Machine Translation

## 1 Introduction

We start off by describing the process of determining number and countability for noun phrases in our Japanese-to-English machine translation system **ALT-J/E** (Ikehara *et al.* 1991; Ikehara *et al.* 1996). This is an updated summary of (Bond *et al.* 1994; Bond *et al.* 1996). We then discuss the implications that this has for disambiguation, both for dialogue-based Machine Translation (Boitet 1996a: this volume) and the AD/ID/AD sandwich (Boitet 1996b: this volume).

There are several problems that need to be considered when we decide whether to use automatic or interactive disambiguation. The first is how critical the ambiguity is, does it

need to be resolved with 100% accuracy or can it be left vague. The second is how much of a load will this place upon the users.

## 2 Determining Number and Countability in Japanese-to-English Machine Translation.

Noun phrases in Japanese differ from those in English in two important ways. First, Japanese has no equivalent to English articles. Second, there is no grammatical marking of number.[1] This means that when Japanese is translated into English the choice of articles and number falls upon the translator. Making appropriate choices is a problem for both Japanese-to-English machine translation systems and native speakers of Japanese using English.

### 2.1 Countability

Countability is generally presented in books for second language learners as a binary property of nouns: a noun is either COUNT or MASS. Unfortunately, this generalization is close to useless in very many cases. Most nouns can appear in either a countable or uncountable noun phrase, but they can be underspecified e.g. *cake* in *I ate the cake*. Nouns exhibit a wide range of variation as to their possible use. Prototypical count nouns, such as *dog* have contrasting singular and plural forms, can be directly modified by numerals *two dogs* and other denumerators (*many, few* ...) and are used to refer to discrete individuals. Prototypical mass nouns have no plural form, can not be directly modified by numerals * *two equipment* (instead using a classifier *two pieces of equipment*), use 'mass' quantifiers (*much, little* ...) and typically refer to substances. However, potentially all 'count' nouns (except some classifiers) can be used in the same way as a mass noun, with reference to substance (*there is dog all over the road*) and most 'mass' nouns can be used in the same way as count nouns, with references to portions (*I'll have two beers*) or kinds (*There are only two beers that I like*). In addition there are other nouns with different restrictions, such as dual pluralia tantum (*scissors*) which only have plural forms (unless being used as modifiers), cannot be directly modified by numerals or denumerators, and typically refer to substances made up of two equal parts.

We distinguish between five major different noun countability preferences (Bond *et al.* 1994), based on the analysis of Allan (1980). '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 these 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 tantum': nouns that have only have a plural form, such as *scissors*.

When a noun appears in a sentence, the referentiality, countability and definiteness of the noun phrase it heads can be affected by a number of factors. We treat them as a defeasible hierarchy of rules, starting at the sentence level and then moving down to the noun phrase level and finally contextual information and world knowledge. Constraints can be placed on the referential use (**`Referentiality`**), countability/number (**`Interpretation`**) and

[1] Japanese does not have obligatory plural morphemes. Plurality can be marked but only rarely is, for example by adding a suffix such as *tachi* "and others".

| | Noun | | Countability | Default | Default |
|---|---|---|---|---|---|
| | English | Japanese | Preference | Number | Classifier |
| CO | knife | *houchou* | FULLY COUNTABLE | Si | |
| | noodles | *men* | FULLY COUNTABLE | Pl | |
| | group | *mure* | (COLLECTIVE) | Si | |
| BC | cake | *ke–ki* | STRONGLY COUNTABLE | Si | |
| BU | beer | *bi–ru* | WEAKLY COUNTABLE | Si | |
| UC | furniture | *kagu* | UNCOUNTABLE | Si | |
| | knowledge | *chishiki* | (SEMI-COUNTABLE) | Si | |
| PT | scissors | *hasami* | PLURALIA TANTUM | Pl | pair |
| | draughts | *chekka-zu* | PLURALIA TANTUM | Pl | set |
| | clothes | *ifuku* | PLURALIA TANTUM | Pl | — |

Table 1: Noun Countability Preferences

definiteness/possession. Depending on the rule's position within the program, it can be overridden by later rules, particularly those triggered by noun modifiers within the noun phrase.

### 2.2 Noun Phrase Countability

Each noun phrase has a feature **`Interpretation`** that interacts with the head noun's Noun Countability Preference.

**`Interpretation`** takes the following values:[2]

| **`Interpretation`** | Explanation | Features |
|---|---|---|
| **`I-Si`** | Individuated — Si | +b, ?i, −Pl |
| **`I-Pl`** | Individuated — Pl | +b, ?i, +Pl |
| **`Mass`** | 'Mass' interpretation | −b, ?i |
| **`Comp`** | 'Composition' interpretation | −b, ?i |
| **`Grnd`** | 'Ground' interpretation | −b, −i |
| **`Dep`** | modifier depends on NCP | see below |

Sometimes the choice of the English translation of a modifier will depend on the countability of the noun phrase. For example, *kazukazu-no* and *takusan-no* can both be translated as *many*. However, *kazukazu-no* implies that its modificant is made up of discrete entities, so the noun phrase it modifies should be translated as individuated — plural. *takusan-no* does not carry this nuance so if the noun phrase it modifies is headed by a weakly countable or uncountable noun, **ALT-J/E** will translate *takusan-no* as *much*.

**`dep`** is currently used for two kinds of modifiers: those whose translation can vary due to number: *this/these, that/those* and those whose translation can vary due to countability: *much/many, little/few, a little/a few, less/fewer, overmuch/too many*.

We show how the translation of a noun phrase depends on both the interpretation of the noun phrases and the noun phrase countability in the following 3 diagrams.

#### 2.2.1 NP Countability and Number (Bounded)

There are two bounded interpretations, which differ only in the number of the noun phrase.

[2]This table also shows the features **b** for boundedness, and **i** for salient internal structure, after Jackendoff (1991).

| Noun Countability Preference | NP **Interpretation** | |
|---|---|---|
| | **I-Si** (+b, −Pl) Individuated — Singular | **I-Pl** (+b, +Pl) Individuated — Plural |
| Fully Countable | 1 dog | 2 dogs |
| | 1 noodle | 2 noodles |
| Strongly Countable | 1 cake | 2 cakes |
| Weakly Countable | 1 beer | 2 beers |
| Uncountable | 1 piece of information | 2 pieces of information |
| Pluralia Tantum | 1 pair of scissors | 2 pairs of scissors |
| | 1 set of scales | 2 sets of scales |
| | ?* 1 piece of clothes | ?* 2 pieces of clothes |

Table 2: NP **Interpretation** (+b)

### 2.2.2 NP Countability and Number (Unbounded)

There are three unbounded intepretations. The most common (Mass), is translated as plural only for fully, strongly countable and plurilia tantum, a typical example of a noun phrase with a **mass** interpretation is the object of *collect*. Another possible translation is to only translate the noun phrase as plural if the head is fully countable or pluralia tantum. A typical example of this is the object of *made of*. In this case (**comp**), the noun phrase will normally be conceived as a substance, and the internal structure will only be expressed for nouns which, in English, are conceived of as having inherent boundaries. The last unbounded interpretation (**grnd**), is used in cases were there is no internal structure, for example the complement of *slice of*. Nouns which have internal strucure as part of their expression in English, such as *scissors* cannot appear in this environment. If you grind a pair of scissors to powder, then it is no longer a pair of scissors!

| Noun Countability Preference | NP **Interpretation** | | |
|---|---|---|---|
| | Mass (−b) Unbounded | Comp (−b) Composition | Grnd (−b, −i) Substance |
| Fully Countable | dogs | dogs | dog |
| | noodles | noodles | noodle |
| Strongly Countable | cakes | cake | cake |
| Weakly Countable | beer | beer | beer |
| Uncountable | information | information | information |
| Pluralia Tantum | scissors | scissors | ? scissors |
| | scales | scales | ? scales |
| | clothes | clothes | ? clothes |

Table 3: NP **Interpretation** (−b)

### 2.2.3 NP Countability and Number (Unknown)

Finally we come to noun phrases for which there is no information about their interpretation. We translate these based on the inherent countability of their heads.

| Noun Countability Preference | NP **Interpretation** | |
|---|---|---|
| | **Unknown** (default) | **Dep** (*much/many*) |
| Fully Countable | a dog | many dogs |
| | noodles | many noodles |
| Strongly Countable | a cake | many cakes |
| Weakly Countable | beer | much beer |
| Uncountable | information | much information |
| Pluralia Tantum | scissors | many (pairs of) scissors |
| | scales | many (sets of) scales |
| | clothes | many clothes |

Table 4: NP **Interpretation** (?b)

### 2.3 Noun Phrase Referentiality

We distinguish between three possible referential uses for noun phrases in context, generic, referential and ascriptive (Bond *et al.* 1995).

**Generic** Used to make general statements about a class.
*Mammoths are extinct.*

**Referential** Used to refer to some referent.
*Two dogs chase a cat.*

**Ascriptive** Used to ascribe a property to something.
*Hathi is an elephant.*

Noun phrases with generic reference are translated the same as noun phrases with the **comp** interpretation, that is only fully countable and pluralia tantum are translated as plural.

### 2.4 Determining Noun Phrase Countability

The determination of noun phrase countability (**interpretation**) for referential and ascriptive noun phrases is shown in figure 1. In the implementation in **ALT-J/E**, the interpretation is determined during the transfer phase, along with the noun phrase referentiality and definiteness.

The default interpretation is **unknown**, i.e. that there is no information in the Japanese text that gives any usable clues about number and countability, except for the translation of the head noun itself.

There are several top level rules, which use the meanings of verbs combined with the referential properties of noun phrases. These can all be overridden by subsequent rules. In the lexical entry for verbs, it is possible to write information about the interpretation of noun phrase countability of number for any sub-categorised noun phrase (Bond *et al.* 1993). For example, the object of the transitive verb *collect* will have a **mass** interpretation.

The next level of rules applies to noun phrases modified by adjuncts. Some adjuncts, such as those that specifiy location, give clues as to the interpretation of the noun phrase they modify.

The next level of rules is based on direct modification by Japanese plural markers such as *tachi*, *domo* or *ra*. They imply that the noun phrase so marked is bounded and plural.

1. The default is UNKNOWN
2. Top level rules
   (a) Determine according to verb:
   *wadai-ga tsukita* "I ran out of topics"
   (b) Ascriptive NPs match their subjects:
   *A computer is a piece of equipment*
3. Determine according to adjunct
   (a) Number implied by adverbial phrases
   *zenkoku-no gakkou* "schools all over the country"
4. Specific Japanese marking
   (a) If the Japanese is marked as plural:
   *tachi* ⇒plural and individuated
5. Modification within the noun phrase
   (a) Determine according to classifier:
   *hito-kire-no ke-ki* "a slice of cake"
   *hito-yama-no ke-ki* "a pile of cakes"
   (b) Determine according to quantifier:
   *ono'ono-no ke-ki* "each cake"
6. Context & world knowledge
   (a) A noun phrase judged to be co-referential has the same number and countability as its antecedent:
   *Two men, strangers, came in.*

Figure 1: Determining Noun Phrase Countability

The penultimate level of rules cover modification within the noun phrase itself, in particular quantifiers such as *onoono-no* "each" and classifiers. The use of classifiers to interpret the noun phrase is discussed in more detail in the next section.

Finally, contextual information and world knowledge can override any of the above rules. At the moment, this can only be done in one way. A noun phrase judged to be co-referential has the number and countability of its antecedent.

### 2.4.1 Classifiers

In this subsection, we will describe in some detail the use of classifiers to determine the number (and form) of the noun phrase. We found that noun phrases headed by classifiers appeared in 242 out of 981 sentences (25%) in a random selection of Japanese wire reports (Nikkei News Bulletin), and they made up 5% of the total number of noun phrases (488 out of 9454). The majority of noun phrases headed by classifiers were numerical expressions (85%), for which an accurate translation is particularly important. We use the following abbreviations:

A, B or N: noun phrase; C: classifier, X: Numeral (or equivalent).

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. The main property distinguishing them from normal nouns is that they can postfix to numerals, the quantifier *sū* "some" or the interrogative *nani* "what", to form a noun phrase. Unlike normal nouns in Japanese, *josūshi* cannot form grammatical noun phrases on their own.[3]

(1) *2-hiki* "2 animals"[2 animals] (Numeral)

(2) *sū-hiki* "some animals" (Quantifier)

(3) *nan-biki* "how many animals" (Interrogative)

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'. In order to concentrate on translating 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), 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.

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

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 2.4.1), METRIC (Section 2.4.1), GROUP (Section 2.4.1) and SPECIES (Section 2.4.1). 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.

**Unit classifiers:** UNIT classifiers are the prototypical classifiers. A UNIT classifier will be realized in Japanese as a *josūshi*. It has the effect of forcing a bounded interpretation for the noun phrase as a whole.

We further divide UNIT classifiers into two subtypes, depending on how the classifier itself is translated. 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* "pice" 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 (**`unknown`**). This translation is exactly the same as **`I-Si/Pl`**, with the number depending on the cardinality of the numeral. The second type of classifier, TYPICAL, consists of those

[3]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 seperate 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*.

that 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 two types of UNIT classifier are summarized in Table 5.

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

Table 5: Unit Classifiers

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.

Note that, as the conceptual classes are different in the two languages, it is impossible for a native speaker of one to predict which is correct in the other. For example, a Japanese native speaker with no knowledge of English will be unable to choose *rasher* as a translation of *mai* "slice" for bacon.

**Metric classifiers:** The next overall category is METRIC classifiers. We divide metric classifiers into three sub types. The first two, MEASURE and CONTAINER classifiers are both translated as 'X C *of* N', where N will have a **`comp`** interpretation. In fact METRIC classifiers do not specify whether the measured noun is countable or not. You could equally well have *a box of cake* or *a box of cakes*. As a default, we give the former translation, as it is more general, *a box of cakes* is always a *a box of cake*, but not vice versa.

The third type of classifier, PORTION, forces an interpretation with no internal structure, for example *-kire* "slice". The classifier is always translated, and the embedded noun phrase has the **`grnd`** interpretation: *1-kire-no-inu* "1-slice of dog" $\Rightarrow$*1 slice of dog*. In Bond *et al.* (1996) PORTION classifiers were treated as a special kind of UNIT classifier, however the fact that they are always translated, and force an unbounded interpretation makes it more natural to treat them as METRIC classifiers.

The difference between MEASURE and CONTAINER classifiers is that, in English, noun phrases headed by container classifiers can have two interpretations, denoting either the substance bounded by the container (4) or the container itself (6), although Japanese does not have this ambiguity.

(4) 2-hai-no biiru-wo kobosita
*2-C*-GEN *beer*-ACC *spilt*
[I] spilt 2 cups of beer (CONTAINER)

(5) * 2-hai-no biiru-wo kowasita
*2-C*-GEN *beer*-ACC *broke*
[I] broke 2 cups of beer (CONTAINER)

(6) 2-tsu-no biiru-no koppu-wo kowasita
*2-C*-GEN *beer*-GEN *cup*-ACC *broke*
[I] broke 2 cups of beer (UNIT)

Japanese does have an ambiguity with metric classifiers as shown in (7), which can have three possible interpretations, depending on whether the classifier phrase is interpreted as quantifying (like a classifier) or attributive (like an adjective). There are no syntactic clues to distinguish between the two interpretations. At present **ALT-J/E** defaults to the quantifying interpretation.

(7) 1-kg-no ringo
*1-C*-GEN *apple*
1kg of apples (default)
1kg of apple
a 1kg apple (attributive)

**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 'A-*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 (8) and (9) during English generation, even though their Japanese structure is very different.

(8) *2-hako-no-pen* "2 box of pen" ⇒*2 boxes of pens* 'XC-*no*-N'

(9) *pen-no-hako* "box of pen" ⇒*a box of pens* 'N-*no*-C'

Examples of GROUP classifiers are given in Table 6.

| 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 |

Table 6: Group and Species Classifiers

**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 6.

## 3 Interactive Disambiguation

All the processing described above has been implimentd in **ALT-J/E** as part of its automatic disambiguation. A success rate of 80% (number of noun phrases correctly translated with respect to number and determiners/total number translated) has been obtained. Analysing the errors, about 10% are due to errors elsewhere in the system, and thus outside the scope of this article. Of the remaining 10%, half of them can improved by further tuning, in particular the use of knowledge from other sentences. The remaining 5% require world knowledge that is not currently available.

Could we improve our results by using interactive disambiguation? The answer is a qualified yes. It would be possible to reduce the erros, but probably not completely, and at a cost that appears to prohibitive.

Boitet (1996a) introduced two new paradigms for interactive disambiguation: 'Dialogue-Based MT for Monolinguals' and the 'AD/ID/AD sandwich'. In the first, a monolingual user interactively disambiguates their own utterances fully, and it is then translated into one or more languages. In the second, the system automatically disambiguates as far as it is confident (possibly to a threshold which the user can set ) then asks a user structured questions (preferably in order of importance) and disambiguates automatically again after each one. Finally, when the user is tired of answering questions, the system does what it can to solve any remaining ambiguity and then translates.

Let us first consider the 'Dialogue-Based MT for Monolinguals'. There are two problems. The first, and most critical, is the time involved. From our experience, it is not possible to automatically identify which noun phrases are problematic, therefore all noun phrases would have to be fully disambiguated. Translating from Japanese to English would require the user to specify the following, for each noun phrase, **after** the translation has been chosen (which proabably also requires some questions): Referential Property, Interpretation, Definiteness, Noun Countability Preference. This is clearly unrealistic. The second problem is that non-native speakers are not able to distinguish between countability preferences as they are culture dependent. Consider *scales*: in British English they are pluralia tantum, with the default classifier *pair*; in Australian English they also pluralia tantum, but with the default classifier *set*, while in American English they are fully countable (Wierzbicka 1996:387–390). The countability depends on the shared culture and conventions of the speakers, and cannot be reliably predicted by a non-native.

The 'AD/ID/AD sandwich' approach is more attractive. A good implementation could reduce the number of questions needed considerably. A quick look at the results of our system, however, suggests that over half of the noun phrases have the default value for their Referential Property, which is still one of the major sources of errors. Therefore the only way to disambiguate this property correctly is to ask the user for every noun phrase. Aassuming that an appropriate translation for the head noun has been made, and its referential property is known, the situation looks more promising. Considering only the **`Interpretation`**, the number of questions can be greatly reduced, as it is generally only necessary to ask about noun phrases headed by strongly and weakly countable nouns, which appear less than 25% of

the time. This still leaves a lot of questions. A final approach, and the only one that I think is feasible, would be to limit the questions to problems of structural disambiguation, such as the difference between *a 500gm apple* or *500gm of apple(s)*. These show critical differences in meaning and are easy to identify. After they have been resolved, it would seem to be most efficient to go on to the final level of automatic disambiguation, where the noun phrase interpretation is combined with the head noun's countability preference, and leave any other disambiguation to the post editor.

## 4 Conclusion

My [Francis Bond] personal opinion is that there is still a lot that can be done to improve the automatic disambiguation of text,[4] and that this is the most important area of research for machine translation. Automatic disambiguation requires an enormous amount of knowledge, as outlined by Ikehara *et al.* (1996), but the knowledge is cumulative and residual. Once you've got it, it's there and can be used forever.

## References


ALLAN, KEITH. 1977. Classifiers. *Language* 53.285–311.

——. 1980. Nouns and countability. *Language* 56.541–67.

BOITET, CHRISTIAN. 1996a. Dialogue-based machine translation for monolinguals and future self explaining documents. In *Proceedings of the International Seminar on Multimodal Interactive Disambiguation (MIDDIM '96)*.

——. 1996b. Dialogue-based machine translation for monolinguals and future self explaining documents. In *Proceedings of the International Seminar on Multimodal Interactive Disambiguation (MIDDIM '96)*.

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)*, 32–38. (cmp-lg/9511001).

——, ——, and ——. 1996. Classifiers in Japanese-to-English machine translation. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING '96)*, 125–130. (cmp-lg/9608014).

——, ——, ——, and SATOSHI SHIRAI. 1993. Using the meanings of verbs to select the countability of English noun phrases. In *Proceedings of the 1993 IEICE Fall Conference*, 6:61–62. IEICE.

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)*, 1–14. (cmp-lg/9601008).


[4] Even problems like resolving the difference between *a 500gm apple* or *500gm of apple(s)* can potentially be done by a large scale knowledge base.

IKEHARA, SATORU, SATOSHI SHIRAI, and FRANCIS BOND. 1996. Approaches to disambiguation in **ALT-J/E**. In *Proceedings of the International Seminar on Multimodal Interactive Disambiguation (MIDDIM '96)*.

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*, 101–106. (cmp-lg/9510008).

JACKENDOFF, RAY. 1991. Parts and boundaries. In *Lexical and Conceptual Semantics*, ed. by Levin and Pinker. Elsevier.

KUNO, SUSUMU. 1973. *The Structure of the Japanese Language*. Cambridge, Massachusetts, and London, England: MIT Press.

QUIRK, RANDOLPH, SIDNEY GREENBAUM, GEOFFREY LEECH, and JAN SVARTVIK. 1985. *A Comprehensive Grammar of the English Language*. Essex: Longman.

WIERZBICKA, ANNA. 1996. *Semantics: Primes and Universals*. Oxford University Press.