Using an Ontology to Determine English Countability

Francis Bond[*] and **Caitlin Vatikiotis-Bateson**[**]
[*] `bond@cslab.kecl.ntt.co.jp` [**] `caitlinvb@yahoo.com`
2-4 Hikari-dai Seika-cho, Kyoto, Japan 619-0237
NTT Communication Science Laboratories,
Nippon Telegraph and Telephone Corporation

Abstract

In this paper we show to what degree the countability of English nouns is predictable from their semantics. We found that at 78% of nouns' countability could be predicted using an ontology of 2,710 nodes. We also show how this predictability can be used to aid non-native speakers to determine the countability of English nouns when building a bilingual machine translation lexicon.

1 Introduction

In English, nouns heading noun phrases are typically either **countable** or **uncountable** (also called **count** and **mass**). Countable nouns can be modified by denumerators, prototypically numbers, and have a morphologically marked plural form: *one dog*, *two dogs*. Uncountable nouns cannot be modified by denumerators, but can be modified by unspecific quantifiers such as *much*, and do not show any number distinction (prototypically being singular): ** one equipment*, *some equipment*, **two equipments*.

Knowledge of countability is important when translating from a source language without obligatory number and countability distinctions to a target language that does make number distinctions. Some examples are Japanese-to-English (Ehara and Tanaka, 1993; Bond, 2001), Japanese-to-German (Siegel, 1996), and Chinese-to-English.

For a system generating English, it is important to know the countability of the head noun, as this determines whether it can become plural, and the range of possible determiners. Knowledge of countability is particularly important in machine translation, because the closest translation equivalent may have different countability from the source noun. Many languages, such as Chinese and Japanese, do not mark countability, which means that the choice of countability will be largely the responsibility of the generation component.

In this paper, we measure how well semantic classes predict countability. Obviously, the answer depends both on how many countability distinctions are made, and how many semantic classes are used. If every sense of every word belongs to its own semantic class, then semantic classes will uniquely, although not usefully, predict countability. This is effectively the position taken by Wierzbicka (1988), where the semantics of a noun, given in the Natural Semantic Metalanguage, always provides enough information to predict the countability. On the other hand, if there are only a handful of semantic classes, then they will have little predictive power. We first define countability, and discuss its semantic motivation (§ 2). Then we present the lexical resources used in our experiment (§ 3), including the ontology of 2,710 semantic classes. Next, we describe the experiment, which uses the semantic classes of words in a Japanese-to-English transfer dictionary to predict their countability (§ 4). Finally, we present the results and discuss the theoretical and practical implications in (§ 5).

2 Linguistic Background

Grammatical countability is motivated by the semantic distinction between **object** and **substance** reference (also known as **bounded/non-bounded** or **individuated/ non-individuated**). Imai and Gentner (1997) show that the presence of countability in English and its absence in Japanese influences how native speakers conceptualize unknown nouns

[**] This research was done while the second author was visiting the NTT Communication Science Laboratories

as objects or substances. There is definitely some link between countability and conceptualization, but it is a subject of contention among linguists as to how far grammatical countability is motivated and how much it is arbitrary. Jackendoff (1991) assumes countability and number to be fully motivated, and shows various rules for conversion between countable and uncountable meanings, but does not discuss any of the problematic exceptions.

The prevailing position in the natural language processing community is to effectively treat countability as though it were arbitrary and encode it as a lexical property of nouns. Copestake (1992) has gone some way toward representing countability at the semantic level using a type **form** with subtypes **countable** and **uncountable** with further subtypes below these. Words that undergo conversion between different values of **form** can be linked with lexical rules, such as the **grinding** rule that links a **countable** animal with its **uncountable** interpretation as meat. These are not, however directly linked to a full ontology. Therefore there is no direct connection between being an animal and being countable.

Bond et al. (1994) suggested a division of countability into five major types, based on Allan (1980)'s noun countability preferences (NCPs). Nouns which rarely undergo conversion are marked as either `fully countable`, `uncountable` or `plural only`. Nouns that are non-specified are marked as either `strongly countable` (for count nouns that can be converted to mass, such as *cake*) or `weakly countable` (for mass nouns that are readily convertible to count, such as *beer*). Conversion is triggered by surrounding context. Noun phrases headed by uncountable nouns can be converted to countable noun phrases by generating classifiers: *one piece of equipment*, as described in Bond and Ikehara (1996).

Full knowledge of the referent of a noun phrase is not enough to predict countability. There is also language-specific knowledge required. There are at least three sources of evidence for this: the first is that different languages encode the countability of the same referent in different ways. To use Allan (1980)'s example, there is nothing about the concept denoted by *lightning* that rules out **a lightning* being interpreted as *a flash of lightning*. In both German and French (which distinguish between countable and uncountable uses of words) the translation equivalents of *lightning* are fully countable (*ein Blitz* and *un éclair* respectively). Even within the same language, the same referent can be encoded countably or uncountably: *clothes/clothing*, *things/stuff*, *jobs/work*. The second evidence comes from the psycholinguistic studies of Imai and Gentner (1997) who show that speakers of Japanese and English characterize the same referent in different ways depending on whether they consider it to be countable (more common for English speakers) or uncountable (more common for Japanese speakers). Further evidence comes from the English of non-native speakers, particularly those whose native grammar does not mark countability. Presumably, their knowledge of the world is just as complete as English native speakers, but they tend to have difficulty with the English specific conceptual encoding of countability.

In the next section (§ 3) we describe the resources we use to measure the predictability of countability by meaning, and then describe our experiment (§ 4).

3 Resources

We use the five noun countability classes of Bond et al. (1994), and the 2,710 semantic classes used in the Japanese-to-English machine translation system **ALT-J/E** (Ikehara et al., 1991). These are combined in the machine translation lexicons, allowing us to quantify how well semantic classes predict countability.

3.1 Semantic Transfer Dictionary

We use the common noun part of **ALT-J/E**'s Japanese-to-English semantic transfer dictionary. It contains 71,833 linked Japanese-English pairs. A simplified example of the entry for *usagi* "rabbit" is given in Figure 1. Each record of the dictionary has a Japanese index form, a sense number, an English index form, English syntactic information, English semantic information, domain information and so on. English syntactic information includes the part of speech, noun countability preference, default number, default article and whether the noun is inherently possessed. The semantic information includes common and proper noun semantic classes. In this example, there are two semantic classes: `animal` subsumed by `living thing`, and `meat` subsumed by *foodstuff*.

Because the dictionary was developed for a Japanese-to-English machine translation system, many of the English translations are longer than the Japanese source terms: many concepts encoded in a single lexical item in Japanese may need multiple words in English. Of the 71,833 entries, 41,285 are multi-word expressions in English (57.4%).

3.2 Semantic Ontology

ALT-J/E's ontology classifies concepts to use in expressing relationships between words. The meanings of common nouns are given in terms of a semantic hierarchy of 2,710 nodes. Each node in the hierarchy represents a semantic class. Edges in the hierarchy represent **is-a** or **has-a** relationships, so that the child of a semantic class related by an **is-a** relation is subsumed by it. For example, `organ` **is-a** `body-part`. The semantic hierarchy and the Japanese dictionary marked with it have been published as Goi-Taikei: A Japanese Lexicon (Ikehara et al., 1997).

The semantic classes are primarily used to distinguish between word-senses using the selectional restrictions which predicates place on their arguments. Countability has not been used as a criterion in deciding which word should go into which class. In fact, because the dictionary has been built mainly by native Japanese speakers, who do not have reliable intuitions on countability, it was not possible to use countability to help decide into which class to put a given word.

Although the dictionary has been extensively used in a machine translation system, errors still exist. A detailed examination of user dictionaries with the same information content, made by the same lexicographers who built the lexicon, found errors in 11–21% of the entries (Ikehara et al., 1995). A particularly common source of errors was words being placed one level too high or low in the hierarchy. The same study found that 90% of words entered into a user dictionary could be automatically assigned to lexical classes with 13–25% errors, although words were assigned to too many semantic classes 32–56% of the time (the range in errors is due to different results from different domains: newspapers and software manuals).

3.3 Noun Countability Preferences

Nouns in the dictionary are marked with one of five major countability preference classes: `fully countable, strongly countable, weakly countable, uncountable` and `plural only`, described at length in Bond (2001). In addition to countability, default values for `number` and classifier (`cl`) are also part of the lexicon. The classes and additional features are summarized in Table 1, along with their distribution in **ALT-J/E**'s common noun dictionary.[1] The most common NCP is `fully countable`, followed by `uncountable`.

The two most basic types are `fully countable` and `uncountable`. Fully countable nouns such as *knife* have both singular and plural forms, and cannot be used with determiners such as *much, little, a little, less* and *overmuch*. Uncountable nouns, such as *furniture*, have no plural form, and can be used with *much*.

Between these two extremes there are a vast number of nouns, such as *cake*, that can be used in both countable and uncountable noun phrases. They have both singular and plural forms, and can also be used with *much*. Whether such nouns will be used countably or uncountably depends on whether their referent is being thought of as made up of discrete units or not. As it is not always possible to determine this explicitly when translating from Japanese to English, we divide these nouns into two groups: `strongly countable`, those that refer to discrete entities by default, such as *cake*, and `weakly countable`, those that refer to nonbounded referents by default, such as *beer*. At present, these distinctions were made by the lexicographers' intuition, as there are no large sense-tagged corpora to train from.

In fact, almost all English nouns can be used in uncountable environments, for example, if they are given the ground interpretation. The only exception is classifiers such as *piece* or *bit*, which refer to quanta, and thus have no uncountable interpretation.

Language users are sensitive to relative frequencies of variant forms and senses of lexical items (Briscoe and Copestake, 1999, p511). The division into `fully, strongly, weakly`

[1] We ignore the two subclasses in this paper: `collective` nouns are treated as `fully countable` and `semi-countable` as `uncountable`.

$$\begin{bmatrix} \textsc{Index} & \textit{usagi} \\ \textsc{sense 1} & \begin{bmatrix} \textsc{English Translation} & \textit{rabbit} \\ \textsc{Part of Speech} & \texttt{noun} \\ \textsc{Noun Countability Pref.} & \texttt{strongly countable} \\ \textsc{Default Number} & \texttt{singular} \\ \textsc{Semantic Classes} & \begin{bmatrix} \textsc{common noun} & \texttt{animal, meat} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 1: Japanese-English Noun Lexical Entry (*usagi* ⇔ *rabbit*)

Table 1: Noun Countability Preferences

Noun Countability Preference	Code	Example	Default Number	Default Classifier	#	%
`fully countable`	CO	knife	`sg`	—	47,255	65.8
`strongly countable`	BC	cake	`sg`	—	3,110	4.3
`weakly countable`	BU	beer	`sg`	—	3,377	4.7
`uncountable`	UC	furniture	`sg`	*piece*	15,435	21.5
`plural only`	PT	scissors	`pl`	*pair*	2,107	2.9

and `uncountable` is, in effect, as a coarse way of reflecting this variation for noun countability.

The last major type of countability preference is `plural only`: nouns that only have a plural form, such as *scissors*. They can neither be denumerated nor modified by *much*. `plural only` are further divided depending on what classifier they take. For example, *pair* `plural only` nouns use *pair* as a classifier when they are denumerated: *a pair of scissors*. This is motivated by the shape of the referent: *pair* `plural only` nouns are things that have a bipartite structure. Such words only use a singular form when used as modifiers (*a scissor movement*). Other `plural only` such as *clothes* use the plural form even as modifiers (*a clothes horse*). In this case, the base (uninflected) form is *clothes*, and the plural form is zero-derived from it. The word *clothes* cannot be denumerated at all. If clothes must be counted, then a countable word of similar meaning is substituted, or *clothing* is used with a classifier: *a garment, a suit, a piece of clothing*.

Information this detailed about noun countability preferences is not found in standard dictionaries. To enter this information into the transfer lexicon, a single (Australian) English native speaker with some knowledge of Japanese examined all of the entries in Goi-Taikei's common-noun dictionary and determined appropriate values for their countability preferences.

4 Experiment and Results

To test how well the semantic classes predict the countability preferences, we carried out a series of experiments.

We ran the experiments under several conditions, to test the effect of combinations of semantic classes and single-word or multi-word entries. In all cases the baseline was to give the most frequently occurring noun countability preference (which was always `fully countable`).

In the experiments, we use five NCPs (`fully, strongly, weakly countable, uncountable` and `plural only`), we do not consider default number in any of the experiments.

For each combination of semantic classes in the lexicon, we calculated the most common NCP. Ties are resolved as follows: `fully countable` beats `strongly countable` beats `weakly countable` beats `uncountable` beats `plural only`. For example, consider the semantic class `910:tableware` with four members: *shokki* ⇔ *tableware* (`UC`), *youshokki* ⇔ *dinner set* (`CO`), *youshokki* ⇔ *Western-style tableware* (`UC`) and *toukirui* ⇔ *crockery* (`UC`).

Conditions	Entries	%	Range	Baseline
Training=Test	all	77.9	76.8–78.6	65.8
Tenfold Cross Validation	all	71.2	69.8–72.1	65.8
Tenfold Cross Validation	single-word	66.6	65.6–67.7	58.6
Tenfold Cross Validation	multi-word	74.8	73.9–75.8	71.1

Table 2: Results

The most common NCP is `UC`, so the NCP associated with this class is `uncountable`.

In our first experiment, we calculated the percentage of entries whose NCP was the same as the most common one. For example, the NCP associated with the semantic class `910:tableware` is `uncountable`. This is correct for three out of the four words in this semantic class. This is equivalent to testing on the training data, and gives a measure of how well semantic classes actually predict noun countability in **ALT-J/E**'s lexicon: 77.9% of the time. This is better than the base-line of all `fully countable` which would give 65.8%. All the results are presented in Table 2.

In order to test how useful countability would be in predicting the countability of unknown words, we tested the system using stratified ten-fold cross validation. That is, we divided the common noun dictionary into ten sets, then tested on each set in turn, with the other nine-tenths of the data used as the training set. In order to ensure an even distribution, the data was stratified by sorting according to semantic class with every 10th item included in the same set. If the combination of semantic classes was not found in the test set, we took the countability to be the overall most common NCP: `fully countable`. This occurred 11.6% of the time. Using only nine tenths of the data, the accuracy went down to 71.2%, 5.4% above the baseline. In this case the training set for `910:tableware` will still always contain a majority of `uncountable` nouns, so it will be associated with `UC`. This will be correct for all the words in the class except *youshokki* ⇔ *dinner set* (`CO`).

Finally, we divided the dictionary into single and multiple word entries (looked at from the English side) and re-tested. It was much harder to predict countability for single words (66.6%) than it was for multi-word expressions (74.8%). We will discuss the reason for this in the next section.

5 Discussion

The upper bound of 78% was lower than we expected. There were some problems with the granularity of the hierarchy. In English, the class names of heterogeneous collections of objects tend to be uncountable, while the names of the actual objects are countable. For example, the following terms are all hyponyms of *tableware* in Wordnet (Fellbaum, 1998): *cutlery, chopsticks, crockery, dishware, dinnerware, glassware, glasswork, gold plate, service, tea set, ...* Most of the entries are either uncountable, or multi-word expressions headed by group classifiers, such as *service* and *set*. The words below these classes are almost all countable, with a sprinkling of plural only (like *tongs*). Thus in the three levels of the hierarchy, two are mainly uncountable, and below that mainly countable. However, **ALT-J/E**'s ontology only has two levels here: `910:tableware` has four daughters, all leaf nodes in the semantic hierarchy: `911:crockery`, `912:cookware`, `913:cutlery` and `914:tableware (other)`. The majority NCPs for all four of these classes are `fully countable`. The question arises as to whether words such as *cutlery* should be in the upper or lower level. Using countability as an additional criterion for deciding which class to add a word to makes the task more constrained, and therefore more consistent. In this case, we would add *cutlery* to the parent node `910:tableware`, on the basis of its countability (or add a new layer to the ontology).

Adding countability as a criterion would also help to solve the problem of words being entered in a class one level too high or too low, as noted in Section 3.2.

We were resigned to getting almost all of the *pair* `plural only` wrong, and we did, but they amount to less than 3% of the total. Although there are some functional similarities, such as a large percentage of `820:clothes for the lower body`, it was more common to get one or two in an otherwise large group, such as *tongs* in the `913:cutlery` class, which is overwhelmingly `fully countable`. Because the major differentiator is physical shape, which is not included in our semantic hierarchy, these words cannot be learned by our method. This is another argument for the importance of representing physical shape so that it is accessible for linguistic processing.

We had expected single word entries to be easier to predict than multiple word entries, because of the lack of influence of modifiers. However, the experiment showed the opposite. Investigating the reason found that single word entries tended to have more semantic classes per word (1.38 vs 1.34) and more varied combinations of semantic classes. This meant that there were 5.1 entries per combination to train on for the multi-word entries, but only 3.7 for the single word entries. Therefore, it was harder to train for the single word entries.

As can be seen in the case of *tableware* given above, there were classes where the single-word and multi-word expressions in the same semantic class had different countabilities. Therefore, even though there were fewer training examples, learning the NCPs differently for single and multi-word expressions and then combing the results gave an improved score: 72.0%.

Finally, there were also substantial numbers of genuine errors, such as ソフトカラー *sofuto karā* which has two translations *soft colour* and *soft collar*. Their semantic classes should have been `hue` and `clothing` respectively, but the semantic labels were reversed. In this case the countability preferences were correct, but the semantic classes incorrect.

An initial analysis of the erroneous predictions suggested that the upper bound with all genuine errors in the lexicon removed would be closer to 85% than 78%. We speculate that this would be true for languages other than English because is not specifically tuned to English, it was developed for Japanese analysis. Unfortunately we do not have a large lexicon of French, German or some other countable language marked with the same ontology to test on.

5.1 Further Work

First, we would like to look more at multiword expressions. There is a general trend for the head of a multiword expression to determine the overall countability, which we did not exploit. Modifiers can also be informative, particularly for quantified expressions such as *zasshoku* ⇔ *various colors* whose English part must be countable as it is explicitly denumerated.

Second, we would like to investigate further the relation between under-specified semantics and countability. Words such as *usagi* ⇔ *rabbit* are marked with the semantic classes for `animal` and `meat`, and the single NCP `strongly countable`. It may be better to explicitly identify countability with the animal sense, and uncountability with the meat sense. In this way, we could learn NCPs for each semantic class individually (ignoring `plural only`) and look at ways of combining them, or of dynamically assigning countability during sense disambiguation. Learning NCPs for each class individually could also help to predict NCPs for entries with idiosyncratic combinations, for which training data may not be found.

Finally, from a psycho-linguistic point of view, it would be interesting to test whether unpredictable countabilities (that is those words whose countability is not motivated by their semantic class) are in fact harder for non-native speakers to use, and more likely to be translated incorrectly by humans.

5.2 Applications

In general, many errors in countability that had been overlooked by the lexicographers in the original compilation of the lexicon and its subsequent revisions became obvious when looking at the words grouped by semantic class and noun countability preference. Most entries were made by Japanese native speakers, who do not make countability distinctions. They were checked by a native speaker of English, who in turn did not always understand the Japanese source word, and thus was unable to identify the correct sense.

Adding a checker to the dictionary tools, which warns if the semantic class does not predict the assigned countability, would help to avoid such errors. Such a tool could also be used for fine tuning the position of words in the hierarchy, and spotting flat-out errors.

Another application of these results is in automatically predicting the countability of unknown words. It is possible to automatically predict semantic classes up to 80% of the time (Ikehara et al., 1995). These semantic classes could then be used to predict the countability at a level substantially above the baseline.

6 Conclusions

Even with a limited ontology and noisy lexicon, semantics does predict countability around 78% of the time. Therefore countability is shown to correlate with semantics. This semantic motivation can be used to build tools to (a) automatically predict countability for unknown words, and (b) serve as a check on consistency when building a dictionary.

Acknowledgments

The authors would like to thank the other members of the NTT Machine Translation Research Group, as well as the NTT Linguistic Media Group, Timothy Baldwin and Ann Copestake. This research was supported by the research collaboration between the NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University.

References

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

Francis Bond and Satoru Ikehara. 1996. When and how to disambiguate? — countability in machine translation —. In *International Seminar on Multimodal Interactive Disambiguation: MIDDIM-96*, pages 29–40, Grenoble. (Reprint of MIDDIM-1996).

Francis Bond and Kentaro Ogura. 1998. Reference in Japanese-to-English machine translation. *Machine Translation*, 13(2–3):107–134.

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. 1994. Countability and number in Japanese-to-English machine translation. In *15th International Conference on Computational Linguistics: COLING-94*, pages 32–38, Kyoto. (`http://xxx.lanl.gov/abs/cmp-lg/9511001`).

Francis Bond. 2001. *Determiners and Number in English contrasted with Japanese — as exemplified in Machine Translation*. Ph.D. thesis, University of Queensland, Brisbane, Australia.

Ted Briscoe and Ann Copestake. 1999. Lexical rules in constraint-based grammars. *Computational Linguistics*, 25(4):487–526.

Ann Copestake. 1992. *The Representation of Lexical Semantic Information*. Ph.D. thesis, University of Sussex, Brighton.

Terumasa Ehara and Hozumi Tanaka. 1993. Kikaihonyaku-ni-okeru shizengengo shori (natural language processing in machine translation). *Journal of Information Processing Society of Japan*, 34(10):1266–1273. (in Japanese).

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

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. 1991. Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–. In *Third Machine Translation Summit: MT Summit III*, pages 101–106, Washington DC. (`http://xxx.lanl.gov/abs/cmp-lg/9510008`).

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, Francis Bond, and Yoshie Omi. 1995. Automatic determination of semantic attributes for user defined words in Japanese-to-English machine translation. *Journal of Natural Language Processing*, 2(1):3–17. (in Japanese).

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.

Mutsumi Imai and Dedre Gentner. 1997. A crosslinguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition*, 62:169–200.

Ray Jackendoff. 1991. Parts and boundaries. In Beth Levin and Steven Pinker, editors, *Lexical and Conceptual Semantics*, pages 1–45. Blackwell Publishers, Cambridge, MA & Oxford, UK.

Kazumi Kawamura, Yasuhiro Katagiri, and Masahiro Miyazaki. 1995. Multi-dimensional thesaurus wth various facets,. In *IEICE Technical Report NLC94-48*, pages 33–40. (in Japanese).

Melanie Siegel. 1996. Definiteness and number in Japanese to German machine translation. In D. Gibbon, editor, *Natural Language Processing and Speech Technology*, pages 137–142. Mouton de Gruyter, Berlin.

Anna Wierzbicka. 1988. *The Semantics of Grammar*. John Benjamins, Amsterdam.