# Determining Countability Classes *

Scott Grimm
*University of Rochester*

Aeshaan Wahlang
*University of Rochester*

**Abstract** This paper provides a corpus-driven investigation into establishing classes of nouns based on grammatical environments relevant to countability, such as combination with cardinal modifiers or appearing as a bare singular. We investigate the countability environments of Allan (1980) and assess their predictive power across a large corpus (350 million words). We show, by applying machine learning methods, that while the environments Allan (1980) distinguishes are predictive, the occurrence of nouns as bare singular and/or bare plural is substantially more powerful as a diagnostic. Using the most important environments, we induce, through automatic clustering, a set of countability classses, which distinguish between varieties of countable, non-countable and pluralia tantum nouns.

## 1 Introduction

Most works on countability quite sensibly begin with a set of grammatical and/or semantic diagnostics which isolate different classes of nouns, named mass, count or other labels. Despite the uniformity among works in the literature in possessing initial discussions on how to determine countability classes, there is a large variability in what results from this discussion. A range of classification schemes have been proposed for countability, based on many different criteria, and, as such, many questions immediately arise: How comparable are the different criteria? How much do they overlap? Are the different countability schemes picking up on distinct aspects of the problem?

To give a sense of the variation, we single out three approaches. The first, the paradigm example of which is Allan (1980), is strictly based on grammatical and/or syntactic contrasts, such as that numeral modifiers are often incompatible with substance nouns like *gold*. A set of these syntactic criteria then determine a number of countability classes. A second approach recognizes syntactic diagnostics, but adds semantic diagnostics, such as cumulative reference (Quine 1960) or quantificational behavior with comparatives (Barner & Snedeker 2005), which in turn may

* The authors would like to thank the current and former members of the Quantitative Semantics Lab at the University of Rochester, especially Yufei Du, Rebecca Friedman, Elizabeth Lee, Caleb New, Isabelle Schmit and Xuan Tang. We also would like to thank the organizers and audience of "The Count-Mass Distinction: A Linguistic Misunderstanding?" for their stimulating feedback.

partially correlate with the nouns' syntactic behavior or differentiate into different semantically-motivated classes nouns that appear, from their syntactic behavior, to be similar. A third approach, exemplified in Wierzbicka (1988), is based in groupings of intuitively similar lexical items, which manifest a cluster of syntactic and semantic correlates.

The varying applications of different diagnostics lead to very different answers to how many distinct classes of nouns are recognizable in their countability behavior. In fact, the researchers just cited come to quite different conclusions as to the number of countability classes: Quine (1960) establishing two classes, Barner & Snedeker (2005) three, Allan (1980) eight, and Wierzbicka (1988) fourteen.

In addition to each study using its own set of diagnostics, the set of nouns used differs across studies and in all cases is very limited—normally to several dozen in the more data-intensive studies—in comparison to the many thousands of common nouns in the English vocabulary. This limitation to a small set of nouns is an unfortunate necessity of traditional methodologies, and in each study it is a reasonable limitation to make. Collectively, however, it is challenging to compare in detail different studies' conclusions in the face of varying diagnostics and data sets.

This paper argues that common techniques from the fields of data science and machine learning can assist to increase the scale of countability studies and provide techniques to compare different diagnostics and classifications. In this, it joins other large-scale studies, such as Kulkarni et al. (2013), Kiss et al. (2014) and Kiss et al. (2016), which contribute studies of semantic and syntactic diagnositics across a large portion of the vocabulary. The focus in this paper, however, is on the common grammatical diagnostics for countability and does not invoke semantic diagnostics at all.

We take as our starting point the study of Allan (1980), who through using a battery of diagnostics argues that nouns do not divide cleanly into countable and non-countable, instead, many sub-classes arise, which can in fact be ordered in terms of their "degree of countability" with respect to the grammatical diagnostics employed. We discuss Allan's (1980) approach in section 2 and critically examine the environments for diagnosing countability across a large corpus in section 3. We then assess the importance of different countability environments in predicting whether a noun is (non-)countable using machine learning methods in section 4, observing, among other things, that bare occurrences of nouns are the strongest predictor of countability status. Section 5 performs clustering on the data to automatically induce a countability classification based on the countability environments. Section 6 concludes, highlighting the implications for the semantic contrasts of countability.

## 2 Background: Countability Preferences (Allan 1980)

The wide-ranging and pioneering study in Allan (1980) argued that a binary, or even ternary, countability classification understates the variation of the nominal domain. Using a set of syntactic tests, he argues for (at least) 8 degrees of countability "preferences" nouns may have. We first provide a discussion of the diagnostics Allan (1980) argues for before discussing his eventual classification. Like most diagnostics, a fair amount of care needs to be taken when applying them, and the reader is directed to Allan (1980) for discussion of a variety of nuances that arise in applying these diagnostics.

Allan first distinguishes a class of elements in the nominal phrase that he terms "denumerators". Denumerators include cardinal numbers, but also quantifiers such as *every* or *both*, and their defining characteristic is that they presuppose that the noun refers to a number of discrete entities. Thus, *each* is a denumerator, since it presupposes discrete entities which in turn can be counted (*each boy/*each sand*), while *some* is not a denumerator, since it is does not presuppose discrete individuals (*some boy/some sand*). Allan argues for the following generalization: If the head constituent of an NP falls within the scope of a denumerator, it is countable. Thus, a noun's countability status follows from its co-occurrence with denumerators.

Allan distinguishes three subtypes of denumerators, UNIT (A+N), FUZZY (F+N), and OTHER (O-DEN). The unit denumerators consist of only the indefinite determiner *a(n)* and *one*. Allan argues that while some nouns, such as *admiration*, reject combining with most denumerators, they have licit uses with unit denumerators, as in (1).

(1) Penelope's is an admiration that I treasure. (Allan 1980: ex. 27)

Fuzzy denumerators include quantifiers and other terms which specify an imprecise number of entities. Allan exemplifies this group of denumerators with *(a) few*, *several*, *many*, *a dozen or so*, *about fifty*, and high round numbers as *five hundred cattle*, *70,000 cattle*.

Allan (1980) argues that this class of denumerators distinguishes pluralia tantum nouns such as *cattle*, which reject precise cardinals, such as *two* or *four*, yet accept fuzzy denumerators, such as *many* or *about fifty*, as shown in (2-a) and (2-b).[1]

(2) a. *Two cattle were severely injured by the falling wall. (Allan 1980: ex. 28)
b. Many cattle died in the cyclone. (Allan 1980: ex. 28)

1 Although fuzzy denumerators and approximative numbers have been less at issue in the literature on countability in general, certain members of this class, such as *hundreds of*, have turned out to be very important testing grounds for classifier languages such as Japanese. See Sudo (2016) for discussion.

The class of "other denumerators", abbreviated as O-DEN, is defined as all the denumerators which are neither unit nor fuzzy denumerators. This, as Allan notes, is a heterogeneous collection, including cardinal numerals and quantifiers such as *each* or *both*.

Allan (1980) isolates two other environments which determine a noun's countability preference. First, certain nouns (and noun phrases) are morphologically undifferentiated from a singular form, but have plural reference, as with *sheep* or *the poor*, which is detectable as they license plural agreement elsewhere in the clause. Allan (1980: p. 551) states that this diagnostic identifies "an NP as countable if it governs plural external number registration", and abbreviates it as EX-PL. This is shown in (3).

(3) Three sheep *were* nibbling the carrot tops when farmer Giles noticed *them*. (Allan 1980: ex. 38)

Second, Allan observes that a distinguishing environment for uncountable nouns is the co-occurence in the singular with the universal quantifier *all*, as shown in (4-a), while countable nouns do not permit this, as shown in (4-b).

(4) a. All lightning is caused by the discharge of electricity from the clouds. (Allan 1980: ex. 52)
b. *All car is 20th century man's horse. (Allan 1980: ex. 53)

Together, these five environments provide a classification over nouns which is potentially very fine-grained. Although the number of potential combinations is rather large ($2^5 = 32$), not all are completely independent from one another, nor are they all of the same discriminatory power. Allan argues that there is an ordering among three of the diagnostics, namely nouns which may occur in O-DEN environments may also occur in F+N environments, and those that occur in F+N environments may also occur in EX-PL environments. The converse relation among these environments does not hold. Through analyzing the behavior of several dozen nouns in the various environments, Allan (1980) adduces eight sets of nouns which represent countability classes, or classes of countability preferences, shown in Table 1. The nouns are cross-classified by each environment in which they are able to occur, here marked by a $+$, or in the case of All+N, fail to occur (since *not* occurring in the All+N environment is diagnostic of countable nouns).

From the table, one can distinguish highly countable nouns, such as *car*, and highly uncountable nouns, such as *equipment*, and between those two poles spans a range of nouns with mixed countability properties. This result is notable for several reasons. First, this indicates that there may be a range of distinctions in play for countability, not all of which lead to a simple bifurcation of nominal meaning into

| Noun Environment | *car* | *oak* | *cattle* | *Himalayas* | *scissors* | *mankind* | *admiration* | *equipment* |
|---|---|---|---|---|---|---|---|---|
| EX-PL | + | + | + | + | + | + | | |
| A+N | + | + | | + | | + | + | |
| *All*+N | + | | + | + | + | | | |
| F+Ns | + | + | + | | ? | | | |
| O-DEN | + | + | | | | | | |

Table 1: Countability preferences of select nouns across the 5 environments (from Allan 1980: p. 562)

countable and non-countable. Second, and not unrelated, the range of nominal data included is wider than most studies, as it includes pluralia tantum, proper nouns, and abstract nouns, all of which are rarely seen in the countability literature. Thus, Allan (1980) argues that the challenge of determining countability contrasts is not simply limited to understanding the contrast between, e.g., objects (*dogs*) and substances (*water*) but is far broader and more nuanced. Finally, this result, as Allan (1980) remarks, is purely syntactic, and except for the initial definition of denumerators, hardly any relation to nominal meaning is asserted.

While Allan (1980) is clearly a pioneering effort, there are several avenues left open for investigation. The study reports on a larger set of data than typically used, yet the amount of data is still quite restricted. Similarly for the different countability environments isolated, these are exemplified by a handful of lexical elements rather than exhaustively tested, nor is much claimed about elements that are not denumerators, namely if they are able to aid in discerning countability preferences. We take these issues up in the next section.

## 3 Quantifying Countability Environments

In this section, we provide a corpus-driven investigation of the differing preferences among nouns for different grammatical environments which bear on countability. We first discuss the details of the corpus, its processing, and the automatic coding of the various countability environments, then analyze the findings and in turn examine a more fine-grained classification of countability environments.

### 3.1 Methodology: Data processing and annotation

We constructed a database of grammatical behavior of nouns to assess the varying countability behaviors of nouns. All data comes from the Corpus of Contemporary American English (COCA) corpus (Davies 2009). COCA is a useful resource since it

presents a collection of well-balanced texts which are controlled for quality, and does not inject the sort of uncertainty into studies that, say, raw internet data or Twitter data might. This study focuses on using 4 of the 5 genre types in the corpus: *Fiction*, *Popular Magazines*, *Newspaper*, and *Academic*. (We set aside the *Spoken* genre as it results in too many parsing errors.) In total, the study spans over a roughly 350 million word portion of the 450 million word corpus. This size of a corpus evades many issues related to data sparsity. While some extremely rare words do not occur in the corpus, in practice, it is uncommon not to find a noun of interest.

We developed an NLP pipeline to process the data and populate a database containing all relevant information. First, it is parsed with the CoreNLP suite (Manning et al. 2014), which includes dependency parsing (De Marneffe et al. 2006) that proves critical for efficiently identifying grammatical patterns. Subsequent processing with a Python script extracts from the parsed output all relevant grammatical relations and represents them as features in the database. More concretely, if the output from the dependency parser contains the dependency DET(DOG, THE), then the script will extract the determiner *the* and, then, in the relevant row of the database representing this occurrence of *dog* in the corpus, mark that the determiner was *the*.[2]

Various post-processing steps were taken to insure the quality of data. For instance, an enormous number of words get tagged as a "noun" by the part-of-speech tagger which may have been abbreviations, brand names, or even unusual punctuation. We filtered the nouns that populated the database so as to consist of only the nouns which occur in the CELEX database (Baayen et al. 1996), which is a large and representative sample of standard English vocabulary. (One drawback of this technique is it will exclude more recent innovations like *bling*.) Of the sentences which contained a noun recognized by this criteria, further exclusion criteria were applied, the most important being the exclusion of instances of the noun where it serves as a modifier in a compound, e.g. compounds such as *school bus* were excluded from the analysis of *school*.[3]

It is worth noting that such a method, while applied to nouns and to the COCA corpus, is very general and could be applied to investigate any part of speech on any corpus. Further, since we employ "Universal Dependencies" (De Marneffe & Manning 2014), that is, dependency annotations that are designed to be cross-linguistically comparable, this general strategy can be applied to a large number of languages in a comparable way.

2 More information was extracted than is at issue in this paper, such as position in the clause, modifiers and all other aspects of the grammatical distribution detectable through the dependencies. This information is not used in this study, however.

3 Further sentences in the corpus were not included in the final database due to limitations of the NLP tools, such as sentences which were too long for the parser or contain html code which make the parser fail.

| Noun Environment | *car* | *oak* | *cattle* | *Himalayas* | *scissors* | *mankind* | *admiration* | *equipment* |
|---|---|---|---|---|---|---|---|---|
| EX-PL | + | + | + | + | + | + | | |
| A+N | + | + | | + | ✓ | + | + | |
| *All*+N | + | | + | + | + | | | |
| F+Ns | + | + | + | | | | | |
| O-DEN | + | + | ✓ | | ✓ | | | |

Table 2: Countability preferences of select nouns across the 5 environments as recognized in the database

In order to quantitatively assess Allan's (1980) countability classification, we transposed his countability environments to a set of search patterns over the corpus. There were several challenges in carrying this out. First, only a handful of denumerators are discussed in the text of Allan (1980), thus a major task was simply extending the classification to all naturally-occurring elements of the nominal phrase. While the A+N and *All*+N environments were straightforward to detect, the *Fuzzy* and *Other* denumerators are essentially unlimited in number. Additionally, it was necessary to delimit the class of non-denumerators.[4] Finally, while almost all of the countability environments could be detected in the corpus, recognizing the EX-PL environment was not possible to do in a quantitatively reliable way, due to the fact a very large number of occurrences simply show ambiguous agreement, thus it is not discernible if it is external singular or plural agreement.

### 3.2 Assessing Allan (1980)

We are now in a position to assess if Allan's (1980) claims hold up across a much larger set of data. First, we examine the distribution of just the 8 nouns from Table 1. A straightforward comparison with Table 1 is given in Table 2, which shows the divergence between Allan's claims and what was found in the corpus. Occurrences of nouns in environments other than those claimed by Allan (1980) are marked with a ✓.

The nouns were observed to occur in nearly all the environments just as discussed in Allan (1980), the only exception being that no instances of *scissors* with fuzzy denumerators were found. At the same time, several nouns passed diagnostics in the corpora that they were asserted not to pass in Allan (1980). In particular, *cattle* and *scissors* were able to be used in more countable environments than expected.

4 A full listing of the mapping from elements of the noun phrase to the Allan categories, as well as all code, models and dataframes, is available at https://quantitativesemanticslab.github.io/.

*cattle* appeared (not infrequently) with numerals, and as such is licit in O-DEN environments, as shown in (5).[5]

(5) "My father had **27 cattle**, which I looked after."

*scissors* appeared with both numerals and with indefinite articles, thus were observed in both O-DEN and A+N environments, as shown in (6) and (7), respectively.

(6) "He belted on the leather shoulder-holster he had custom-made for **his three silver-plated styling scissors**."

(7) "First, careful cutting of young leaves, with **a scissors**, will encourage the plant to continue producing more leaves well into the summer."

Although the occurrence of these nouns in these environments was unexpected given what was reported in Allan (1980), the general classificatory result remains: As Table 2 shows, eight distinct noun classes remain of differing compatibility with the countability environments.

A different perspective on the data can be gained by examining the frequency with which these nouns occur in the different environments. Figure 1 shows the quantitative distribution of the different denumerator environments (A+N, F+N, O-DEN) as well as occurrences in "non-denumerator" environments over a larger set of nouns. This set of nouns contains most of the nouns discussed in Allan (1980), although excludes proper names. Looking at this distribution permits examining if the categories asserted in Tables 1 and 2 are representative across larger groups of nouns. Non-denumerator environments include determiners, quantifiers, and all other material that may occur in a noun phrase, e.g., comparatives, but that do not qualify as denumerators (since they do not presupose discrete individuals). (The class of non-denumerators does not include bare plural or singular occurrences, to which we will turn shortly.)

Several trends are visible in Figure 1. First, there is a split between nouns which permit denumerators and those which do not, which corresponds to the traditional intuition of the count/non-count distinction. Thus, *equipment*, *measles*, *furniture*, and *evidence* do not show the presence of any denumerators, and for *water* and *lightning* their presence is exceedingly rare. To this latter group should be added *mankind*, for which there were only 4 occurrences of an indefinite determiner and all other occurrences were in the bare singular. (Since no other (non-)denumerators co-occur with *mankind*, it has a peculiar position in Figure 1.)

Second, of the different types of denumerators, the A+N is the most frequently found, clearly due to its role as an indefinite determiner and not just as a signal of

5 All examples are from the COCA corpus.

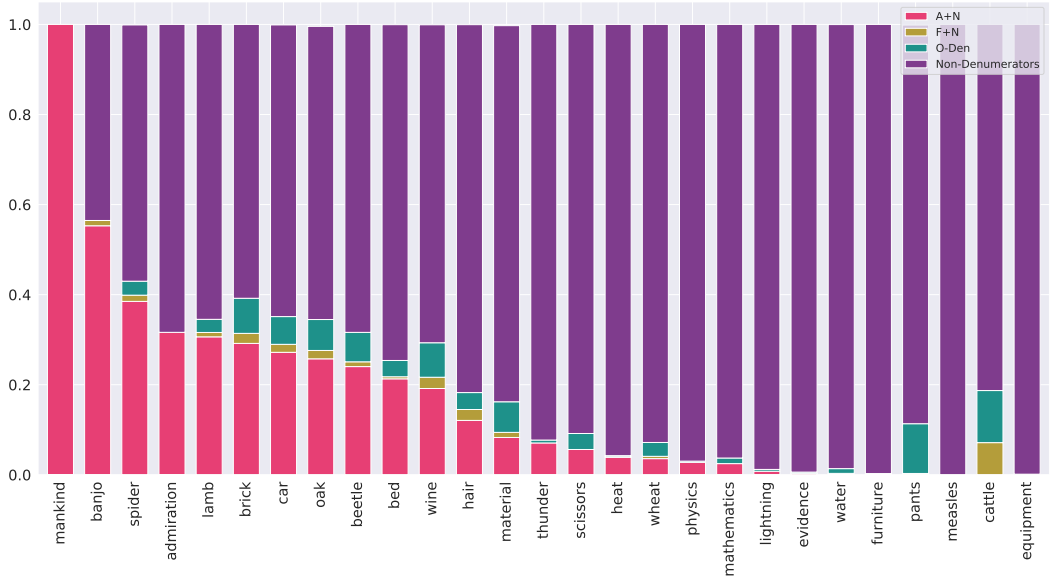


Figure 1: Distribution of nouns across Allan environments

quantity. Allan (1980) argues for the presence of an asymmetry amongst the possible combinations between denumerators and nouns. One of these asymmetries clearly holds, nouns that permit A+N do not necessarily permit the other denumerator types to occur in its environment. On the other hand, the proposed generalization that if a noun occurs in an O-DEN environment then it will also occur in an F+N environment appears less clear. Nouns may occur in the O-DEN environment, but show no evidence of occurring in the F+N environment, as is the case for *physics* and *thunder*, exemplified in (8-a) and (8-b), respectively. Both examples show non-typical uses of the noun at issue. In (8-a), the interpretation is a type-level interpretation, as it is discussing types of physics, while in (8-b), it is a metaphorical extension, that is, *thunder* does not refer to the natural event but a sound event that can be described as "thunder". It seems prudent not to assume that fuzzy denumerator uses come for free simply because a nouns has O-DEN uses.

(8) a. With the Multi-field solver, **each physics** can have totally independent meshes and solution settings.
b. As he cleared the last embattled pair of behemoths he heard **another thunder** of flesh headed into the battle.

We now turn to decomposing the Allan environments to detect more fine-grained generalizations.

| Allan Environment | Subtype |
|---|---|
| Unit | *a/an*<br>*one* |
| Fuzzy | imprecise quantifier (*few, many*)<br>plural numeral (*hundreds*)<br>approximative (*about 50*)<br>round numbers (*100*, *1000*)<br>comparative values (*more than 10*) |
| O-DEN | numbers (*seven,* . . . )<br>digits (*27,* . . . )<br>precise quantifiers (*both, every,* . . . ) |
| Non-Denumerators | *the*<br>measures *half of, quarts of*<br>non-denumerating quantifiers *most, all,* . . .<br>non-quantificational *enough, more than just,* . . . |

Table 3: Correspondence between Allan environments and subtypes thereof

### 3.3 Decomposing the Allan (1980) Environments

While examining nominal countability through the environments argued for by Allan (1980) provides a more nuanced view on the distribution of nouns with respect to syntactic properties of countability, there is much internal variation in each of the environments. For instance, O-DEN includes both numerals and quantifiers while F+N includes even more types of elements, such as modified numerals (*about 50*). As such, the original classification scheme in Allan (1980) might be obscuring further patterns in the data. Yet anther type of information that we aim to keep track of are other elements of the nominal phrase that do not strictly qualify as denumerators, including the definite determiner, quantifiers which do not presuppose individuals (*some*, *any*), measure terms (*kilo of*, *half of*), or special terms like *pair of*. Table 3 shows the correspondence between the Allan environments and the subtypes of those environments.

Figure 2 shows the distribution of the different subtypes of the same nouns from Figure 1. As expected, the distribution of the different subtypes yields a yet more nuanced view than visible with the more coarse-grained Allan environments. First, there are some additional outliers that are now visible. *Measles* is similar to other non-countable nouns by rejecting all denumerators, but is dissimilar to, say, *equipment* or *furniture*, since it also rejects all non-denumerators save the definite determiner *the*. *Cattle*, still an outlier, is an outlier in a different way: it manifests

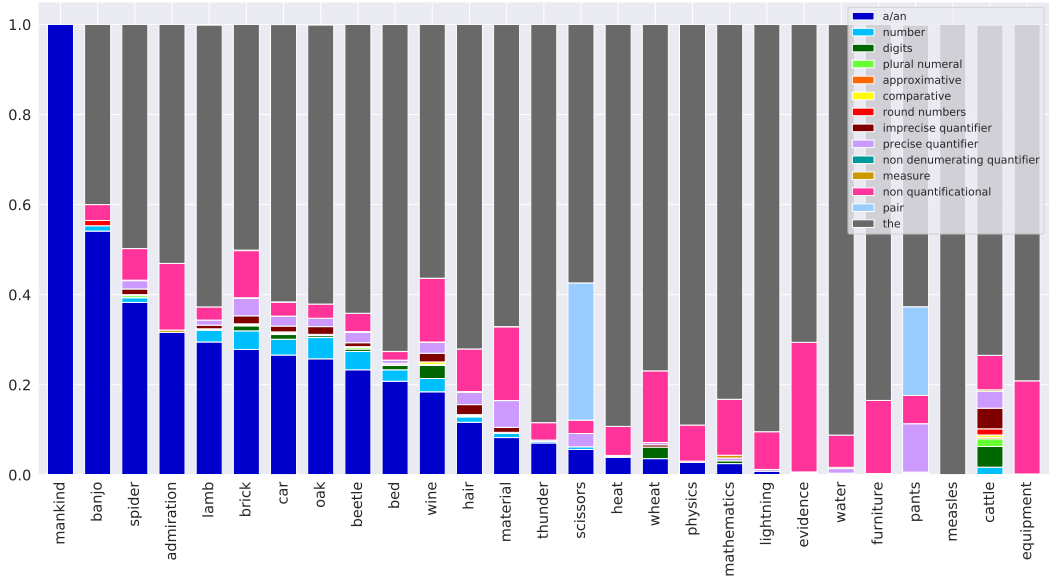


Figure 2: Distribution of nouns across subtypes of Allan environments

the most diversity of different types of quantificational and non-quantificational elements, excepting, of course, the indefinite determiner.

Further contrasts can be seen that are not present in the Allan classification. For instance, tracking not only whether a noun accepts cardinal modification, but also whether it is represented textually as a word or as a digit shows a further distinction between nouns that are tracked in large quantities, such as *cattle* and *wine*, which have a substantial number of uses with digits. Additionally, as would be expected, tracking nouns' co-occurrence with *pair of* isolates certain pluralia tantum nouns (*pants, scissors*) far more clearly than the other environments.

From one perspective, there is not a radical contrast between the distribution in Figure 1 and Figure 2: there are highly countable nouns (*banjo, car*) and highly non-countable nouns (*equipment, furniture*) and a range of behaviors between, some of which are distinctive, in the case of pluralia tantum nouns, and some of which are simply quantitatively different, in the case of *beetle* and *spider*, which do not differ in any categorical way although do so quantitatively. On the other hand, a clearer contrast emerges between nouns which occur in a high diversity of environments, most obviously *cattle* and *wine*, and those which occur in a much more restricted set of environments, such as *admiration* or *lightning*. This is revealing in that while, e.g., *furniture* and *cattle* may have been thought of as semantically similar in that they are both non-countable nouns which have individuals in their denotations, in fact their grammatical behavior is quite divergent, with *cattle* hosting a range of quantifiers

and other elements while *furniture* is limited to non-quantificational elements and the definite determiner.

Having performed a rather detailed assessment of these environments, both the original countability environments from Allan (1980) and more fine-grained subtypes, we now turn to a broad-scale assessment of these environments. Section 4 assesses how predictive the different environments are of countability status using a supervised machine learning method, namely using a form of random forest classification. Section 5 examines if these environments can be used to automatically induce countability classes through unsupervised clustering.

## 4 Assessing the predictive strength of countability environments

This section examines the influence of the individual environments on determining whether a noun is countable or non-countable. Using the environments described in the last section, both the original environments from Allan (1980) and the subtypes of those environments elaborated in section 3.3, we assess through machine learning methods which environments are most predictive.

We use a gradient boosted ensemble learning algorithm similar to random forest classification, XGBoost or "extreme gradient boosting" (Chen & Guestrin 2016). The core method is random forest classification, which yields a classification by means of constructing a multitude of decision trees (see Hastie et al. 2009 for discussion and references). An advantage of using random forest classification is that it reduces the effect of overfitting on training data which is common for decision tree algorithms. Gradient boosting is technique to build a strong predictor model from an ensemble of weak predictors, in our case, an ensemble of decision trees, wherein it attempts to minimize a loss function as it adds each tree to the ensemble. The XGBoost model trains a random forest with gradient boosting. For our purposes, this technique allows us to robustly measure the importance of each environment in a classification task.

We use this method in a supervised fashion, that is, we assume to know whether a noun is countable or not, and then assess what features influence its countability status. We make use of the countability classification performed in the CELEX database (Baayen et al. 1996), which labels each noun as countable, uncountable or both countable and uncountable. Thus, we analyze two cases: (i) what is predictive of nouns that are labeled as countable and (ii) what is predictive of nouns that are labeled as uncountable. In addition, we construct one set of models using the Allan environments and one set using the subtypes of those environments. We also assess how these environments compare with information that can be gleaned from two other syntactic environments which are diagnostic of countable or non-countable status, namely occurrence as a bare plural or bare singular, respectively. The bare plural and bare singular occurrences have a much higher rate of occurrence than

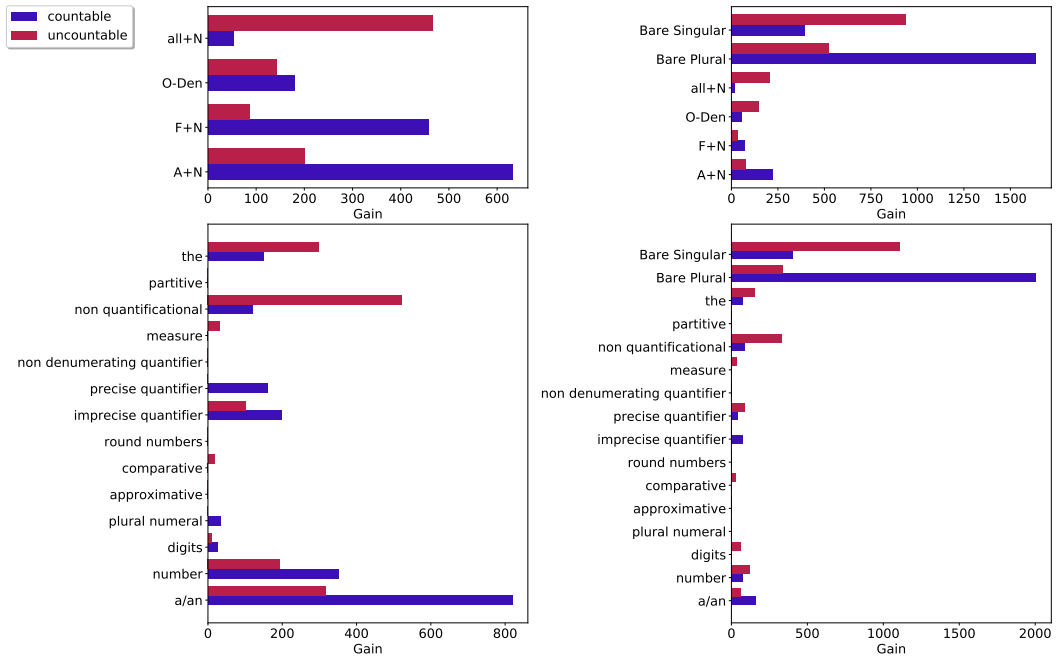


Figure 3: Variable importance in classifying nouns as countable or uncountable across Allan (top) and subtype (bottom) environments with (right) and without (left) bare plural and bare singular included

| Environment | Class | Accuracy |
|---|---|---|
| Allan | Countable | 73.26% |
| Allan | Uncountable | 66.11% |
| Allan + Bare | Countable | 81.17% |
| Allan + Bare | Uncountable | 73.45% |
| Subtypes | Countable | 73.24% |
| Subtypes | Uncountable | 68.72% |
| Subtypes + Bare | Countable | 81.24% |
| Subtype + Bare | Uncountable | 73.91% |

Table 4: Model Accuracy Results Predicting Countable and Uncountable Nouns

any given Allan or Subtype environment, risking obscuring any effect of the latter. To adjust for this, we weight the Allan and Subtype environments by using the calculated proportion of occurrence with respect to all (non-)denumerators, whereas we use the proportion of bare singular and bare plural uses of the noun with respect to all occurrences of the noun.

Table 4 shows the models' results classifing whether a noun was labeled countable or uncountable in CELEX from the different environments and their combinations. Regardless of which set of environments were used as features in the classifier, it is much more difficult to predict whether a noun is uncountable than it is to predict if it is countable, by a 5%-8% difference. Table 4 also shows that there is little difference in the classification accuracy whether the Allan environments are used or the Subtype environments are used; however, there is a significant increase in accuracy in both cases if information from bare plurals and singulars is added.

Figure 3 shows the overall information gained relative to each feature (the gain) for each of the eight models. Examining the Allan environments (upper left panel of Figure 3), it is notable that they perform reasonably close to what could be expected based on the discussion of Allan (1980). The A+N, F+N and O+DEN environments all contribute to classifying a noun as countable, and in an ordering of importance reminiscent of Allan's (1980) claims. Additionally, the high performance of the *All*+N environment indicates that it is indeed a robust diagnostic for uncountability.

Comparing the models without and with the bare plural and bare singular included, the left and right panels of the figure, respectively, shows the overwhelming importance of bare plural and bare singular for predicting countability. While it is not surprising that the bare singular is highly significant for predicting uncountable nouns, it is more surprising that the bare plural outperforms, and many times over, all other environments that have been observed to signal countable nouns, a point to which we will return in section 6.

## 5 Clusters of Countability

We now turn to assessing the relevance of these environments to developing countability classes. We use unsupervised clustering to examine which nouns cluster together in terms of the relevant countability properties. Unlike the last experiment where we assumed a gold standard annotation for countability provided by Baayen et al. (1996), in this experiment we provide no information about countability external to the occurrence of the nouns in the different environments and attempt to induce countability classes directly from that information. Essentially, this is an update on Allan's (1980) original approach using machine learning techniques.

We applied the *Density-based spatial clustering of applications with noise (DBSCAN)* algorithm (Ester et al. 1996) to explore latent countability classes in our data. This algorithm is particularly suited to exploratory work with this sort of data. Operating over a given a set of points in some space—here each noun is a point in the space determined by the values of occurrence in each environment—the algorithm groups together points that are close together.[6] Intuitively, if two nouns behave similarly in terms of the different environments, they will fall under the same cluster.

We explored multiple ways of clustering the data (and also multiple algorithms) but discuss here two of those clusterings that are most directly related to the claims being tested in this paper. For these clustering models, we clustered a total of 6872 nouns. On one experiment, we clustered solely on the environments isolated in Allan (1980): *A+N*, *F+N*, *O-Den*, and *all+N*. This returned a model with a relatively small number (12) of clusters, each of relatively large size (average 485 words per cluster).[7] Only a small percentage were classified as noise (561 nouns, 8.2%), that is, were not identified with a particular cluster. This clustering was able to identify some of the countability contrasts one would expect to see. For instance, the resulting clustering identified *cattle*, *oak* and *wine* as belonging to one cluster, intuitively representing nouns with primarily a non-countable use yet which also have substantial occurrence with indefinites and numerals, in contrast to a separate cluster for more uniformly non-countable nouns such as *equipment*, *thunder*, *evidence*, and *furniture*. We performed a qualitative assessment of this clustering result, however, which indicates that it is both too coarse-grained and consistently conflates nouns which would seemingly be distinct in terms of countability. For instance, a third

6 This method is non-parameteric, so there are no assumptions of a particular distribution, e.g. normal, underlying the data.

7 The parameters for this model were set at EPS=0.9 (a parameter for the maximum distance between two samples in the clustering), minimum samples=8 (that is, clusters must contain at least 8 members), where the Canberra distance metric was used. The parameters for the subsequent model are identical save for a lower EPS value (.8) to promote conservative clustering in the face of a higher number of features.

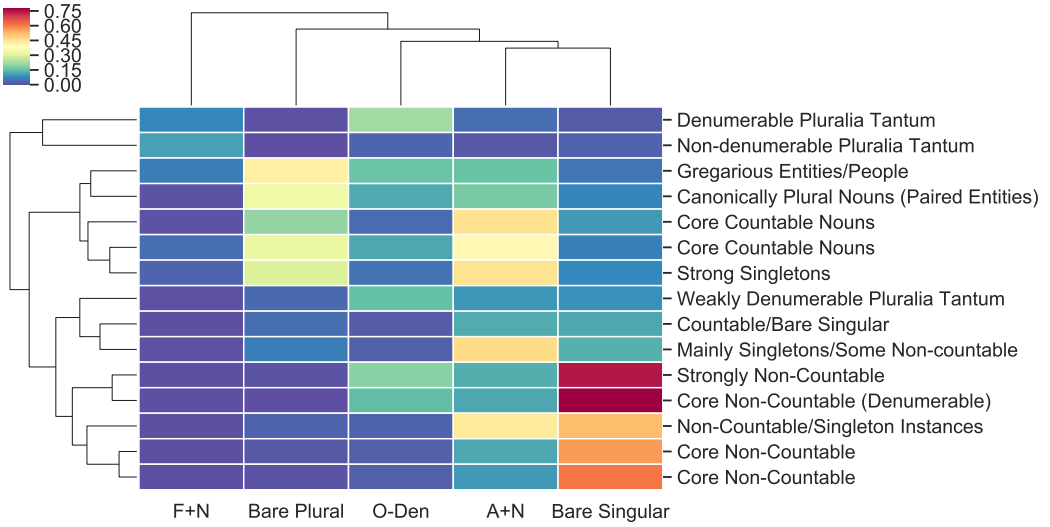


Figure 4: Heatmap Representing the Clusters' Distributional Tendencies Across Countability Environments

cluster contains *beetle*, *lightning* and *scissors*, which intuitively, and according to Allan (1980), should be separated into different countability classes. Given the findings of section 4, this is not altogether surprising as two important environments are not taken into account, the bare singular and the bare plural.

A second clustering was performed with all the Allan environments previously used in conjunction with the bare singular and bare plural environments. This returned a larger number (23) of clusters of smaller size than the preceding model (average 180 words per cluster). This model classified a higher number of nouns as noise (2545 nouns, 37.0%).

While the coverage of this model is not as high as the previous model, a qualitative assessment of the clustering indicated that it was identifying a large number of the countability contrasts, and also classes, that appear in the literature. This clustering again groups together *equipment*, *thunder*, *evidence*, and *furniture*, but correctly separates *lightning* (which occurs in a cluster together with *heat*, *physics* and *water*) from *scissors* (which occurs in a cluster together with certain other pluralia tantum terms which are potentially denumerable, such as *binoculars* and *handcuffs*).

We list below the major clusters identified and provide labels for them. While 23 clusters were identified, we list 15, having excluded clusters that were uninformative being either very small (6 clusters) or not coherent (2 clusters). The list below also presents for each cluster a small number of nouns for to indicate the trends in each cluster. (The full results can be viewed at https://quantitativesemanticslab.github.io/).

Figure 4 displays a heatmap which represents the distribution of the clusters across each environment. The greater the percentage of occurrences in a given environment, the higher the hue, as shown in the legend. In the following, both in the text and the labels of the clusters, we use the terms SINGLETON and GREGARIOUS to indicate the tendency of a noun to refer to single or multiple entities, respectively.

**Denumerable Pluralia Tantum**: Nearly all pluralia tantum nouns, which occur frequently with numerals and other O-DEN denumerator (*briefs, cattle, clothes, fries, singles, species, spectacles, supplies, troops*)

**Non-Denumerable Pluralia Tantum**: Nearly all pluralia tantum nouns or similar, which do not generally occur with denumerators (*belongings, brethren, clergy, dealings, furnishings* )

**Gregarious Entities/People**: Entities or person types which frequently occur in large numbers or groups (*freckle, noodle, petal, photon, tentacle; customer, delegate, fundamentalist, recruit, refugee, spectator*)

**Canonically Plural Nouns (Paired Entities)**: Nouns that have very high rates of plurals, including paired entities and some vegetables (*artichoke, beet, boot, cheekbone, datum, goal, leek, pea, shoe, slipper, standard, yolk*)

**Core Countable Nouns**: A wide range of fully countable nouns which have differing preferences in frequency of occurrence in denumerator environments (Trending towards singleton: *basket, bouquet, contest, ditch, dream, kite, paradigm, return, splash, sum, tornado*; Trending towards gregariousness: *bead, bullet, cookie, follower, hue, impediment, parasite, skyscraper, undergraduate,weapon*)

**Strong Singletons**: Countable nouns which preferentially occur in A+N environments, primarily professional titles (*anthropologist, banker*) and clearly delimited physical objects or events (*asteroid, handbag, mistake, puddle, reward, tattoo*)

**Weakly Denumerable Pluralia Tantum**: Nearly all pluralia tantum nouns, which may occur with numerals and other O-DEN denumerators (*cheek, goggles, gymnastics, jeans, pants, proceedings, savings*)

**Countable/Bare Singular**: Nouns which have a significant use both in bare singular and countable environments, significant subgroups include locations, which often have a bare singular use with prepositions (*bed, deck*), group nouns (*committee, commission, crew* ) and some pluralia tantum nouns (*handcuffs, proceeds, scissors*)

**Mainly Singletons/Some Non-countable**: Nouns which occur primarily with *A+N* and some bare singular uses, highly varied semantic domains (*axe, belly, briefcase, cello, convent, necklace, reputation, rope, rhythm, skirt, spatula*), but body parts and artifacts are frequent

**Strongly Non-Countable**: Nouns which nearly only have a bare singular use (*aviation, fennel, ginger, homelessness, modernism, parenthood, profanity, urbanization*) along with various proper names (*Gregory, Havana*)

**Core Non-Countable (Denumerable)**: Comprised primarily of substances, which have dominant use in the bare singular, yet some may be counted or quantified in certain contexts, e.g., financial contexts (*gold*, *oil*, *barley*), along with various proper names (*Frank, Holland*)

**Non-Countable/Singleton Instances**: Nouns with a primary bare singular use but also a significant use of singulars, often designating an instance of a quality or material (*addiction, awareness, breakfast, calm, guitar, ham, shame, straw*)

**Core Non-Countable**: Primarily standard non-countable nouns, often substances and abstract entities (*awe*, *bacon*, *candlelight*, *colonization*, *despair*, *foliage*, *freedom*, *nutmeg*)

The organization of the clusters, as displayed in Figure 4, serve as a validation of one of Allan's (1980) insights, namely that there are degrees of countability that can be detected through different grammatical environments. Further, as indicated by the dendograms along the x- and y-axes of Figure 4, there is a structure to the different clusters and countability environments. The dendogram on the y-axis of Figure 4 shows three coarse-grained groups among the clusters: (i) pluralia tantum (denumerable and non-denumerable), (ii) countable (plural, core countable and strong singleton nouns), and (iii) non-countable (weakly denumerable pluralia tantum, bare singular and non countable nouns). On the other hand, the dendogram on the x-axis of Figure 4 is based on the similarity of the various environment distributions across the clusters. We see no distinct groups formed in the hierarchy, which implies that these environments are distinct from one another and make independent contributions to the classifications, as can be verified by looking at the distributional patterns in the heatmap. On examining the heatmap we can see that F+N is useful for determining pluralia tantum nouns, as argued by Allan (1980), although not in all cases, as there are three classes of pluralia tantum nouns identified: Denumerable and Weakly Denumerable which have a relatively high proportion of F+N occurrences, but also Non-Denumerable Pluralia Tantum nouns, which do not appear proportionately more in F+N environments.

## 6 Outlook: Implications for the Semantics of Countability

In this section, we redirect our focus from the syntactic distribution of nouns to the semantic implications of this study's results, and the implications for the countability literature more broadly. In particular, we consider three issues: (i) current semantic

models' underfitting the space of variation, (ii) varieties of non-countable nouns and (iii) the importance of bare plurals as a diagnostic.

The most general point arising from this study is that there is much greater variation in nominal behavior than generally acknowledged in theoretical models. Many popular approaches, e.g., Bale & Barner (2009) or Deal (2017), advocate a primarily three-way division between nouns (countable, substance, and *furniture*-type nouns), while others note other semantic classes, such as countable nouns like *fence* (Rothstein 2010). Even more flexible approaches, such as Grimm (2018), do not provide specific analyses of substantially more classes. Yet, reviewing the data and the classes induced in section 5 indicates a much higher degree of variability in nominal behavior and many unexplored semantic contrasts. As such, our semantic models are likely severely underfitting the true variation in nominal meaning as regards countability. One possible response is that this variation is innocent and merely fluctuates in ways that are uninteresting for theoretical linguistics; however, usage frequencies and variation in one language, in this study English, have long been know to correspond to grammatical distinctions in other languages (Bresnan et al. 2001). Thus, these quantitative distinctions observed in English may be crucial to unraveling cross-linguistic variation in countability classification, both in related languages, such as French or German, but also in languages which have a much more elaborate and overt nominal classification system, such as Niger-Congo languages.

More pointedly, a contrast that came clearly to light in section 5 is the different types of non-countable nouns. The contrast between substances (*water*) and non-countable nouns with individuals (*furniture*) is well-known. Thus far not discussed in the literature however are nouns such as *parenthood* or *urbanization*, those which are strongly non-countable and whose grounds for being non-countable elude the current state of the art of nominal semantics. Even the contrast between *cattle* and *furniture*, two nouns which have garnered a fair amount of discussion, remains obscure, as noted in section 3.3. In sum, the semantic work needed to understand the contrasts present in a realistically large swath of English vocabulary is substantial. Importantly, both lexical and formal semantic work are critical for this enterprise.

Finally, one of the more surprising results from the study in section 4 was that the bare plural served as a more informative environment to determine if a noun was countable or not than any other diagnostic. Although to our knowledge, this has not been discussed, there is a clear intuitive basis for this, as bare plural usage would appear to be even more discriminative than combinations with the indefinite article or numerals. In particular, some weakly countable nouns, such as *admiration* or *thunder*, may take an indefinite article or occasionally numbers, in contexts, such as packaging or sorting, where individual entities or instances must be identified; however, the bare plural form, whose most well-known use is for generic reference, is less likely to be licensed due to particular contextual needs. In addition, there

is likely competition from the use of the bare singular for any required generic uses.[8] Thus, another clear avenue for future research is to examine more closely countability phenomena in generic contexts.

## References

Allan, Keith. 1980. Nouns and countability. *Language* 56(3). 41–67.

Baayen, R.H., R. Piepenbrock & L. Gulikers. 1996. *Celex2*. Philadelphia, PA: Linguistic Data Consortium.

Bale, Alan & David Barner. 2009. The interpretation of functional heads: Using comparatives to explore the mass/count distinction. *Journal of Semantics* 26. 217–252.

Barner, David & Jesse Snedeker. 2005. Quantity judgments and individuation: Evidence that mass nouns count. *Cognition* 97. 41–66.

Bresnan, Joan, Shipra Dingare & Christopher D Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG01 conference*, 13–32. CSLI Publications Stanford, CA.

Chen, Tianqi & Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794. ACM.

Davies, Mark. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14(2). 159–190.

De Marneffe, Marie-Catherine, Bill MacCartney & Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *LREC*, vol. 6, 449–454.

De Marneffe, Timothy Dozat Natalia Silveira Katri Haverinen Filip Ginter Joakim Nivre, Marie-Catherine & Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *LREC*, vol. 14, 4585–92.

Deal, Amy Rose. 2017. Countability distinctions and semantic variation. *Natural Language Semantics* 25. 125–171.

Ester, M, HP Kriegel, J Sander & Xu Xiaowei. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second international conference on knowledge discovery and data mining (kdd-96)*, Menlo Park, CA, USA: AAAI Press.

8 See Grimm (2016) for discussion of how this competition plays out for the noun *crime*, which possesses both a non-countable and countable usage, and interestingly appears to only permit kind-level reference with the bare singular, while the bare plural preferably refers to specific crime events or subkinds of crime.

Grimm, Scott. 2016. Crime investigations: The countability profile of a delinquent noun. *Baltic International Yearbook of Cognition, Logic and Communication* 11(1). 4.

Grimm, Scott. 2018. Grammatical number and the scale of individuation. *Language* 94(3). 527–574.

Hastie, Trevor, Robert Tibshirani & Jerome Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction (2nd edition)*. Springer-Verlag.

Kiss, Tibor, Francis Jeffry Pelletier, Halima Husic, Roman Nino Simunic & Johanna Marie Poppek. 2016. A sense-based lexicon of count and mass expressions: The bochum english countability lexicon. In *LREC 2016*, vol. 10, 2810–2814.

Kiss, Tibor, Francis Jeffry Pelletier & Tobias Stadtfeld. 2014. Building a reference lexicon for countability in english. In *LREC*, vol. 9, 995–1000.

Kulkarni, Ritwik, Susan Rothstein & Alessandro Treves. 2013. A statistical investigation into the cross-linguistic distribution of mass and count nouns: Morphosyntactic and semantic perspectives. *Biolinguistics* 7. 132–168.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard & David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for computational linguistics (ACL) system demonstrations*, 55–60.

Quine, W. V. O. 1960. *Word and object*. Cambridge, MA: MIT Press.

Rothstein, Susan. 2010. Counting and the mass/count distinction. *Journal of Semantics* 27. 343–397.

Sudo, Yasutada. 2016. The semantic role of classifiers in Japanese. *Baltic International Yearbook of Cognition, Logic and Communication* 11(1). 10.

Wierzbicka, Anna. 1988. *The semantics of grammar*. Amsterdam/Philadelphia: John Benjamins.