A Statistical Investigation into the Cross-Linguistic Distribution of Mass and Count Nouns: Morphosyntactic and Semantic Perspectives

Ritwik Kulkarni, Susan Rothstein & Alessandro Treves

We collected a database of how 1,434 nouns are used with respect to the mass/count distinction in six languages; additional informants characterized the semantics of the underlying concepts. Results indicate only weak correlations between semantics and syntactic usage. In five out of the six languages, roughly half the nouns in the database are used as pure count nouns in all respects; the other half differ from pure counts over distinct syntactic properties, with fewer nouns differing on more properties, and typically very few at the pure mass end of the spectrum. Such a graded distribution is similar across languages, but syntactic classes do not map onto each other, nor do they reflect, beyond weak correlations, semantic attributes of the concepts. Considerable variability is seen even among speakers of the same language. These findings are in line with the hypothesis that much of the mass/count syntax emerges from language- and even speaker-specific grammaticalization.

Keywords: cross-linguistic variability; language universals; mass/count syntax; multi-dimensional scaling; mutual information

1. Introduction

The mass/count distinction between nouns, in various languages, has been discussed in the linguistic literature since Jespersen (1924), and has received considerable attention in particular in the last 35 years (see the bibliography in Bale & Barner 2011). This distinction between mass and count nouns is a grammatical difference, which is reflected in the syntactic usage of the nouns in a natural language, if it makes the distinction at all (as has been often noted, not all

This study benefited from many contributors at various stages of its seven-year course, starting with Moshe Abeles, who encouraged us to interact. Our gratitude goes first and foremost to the informants who committed to the tedious work of compiling the database, including the ten or so (spanning six more languages) whose responses had to be discarded half-way through. Special thanks to Judit Gervain (Hungarian native speaker), who helped us select the nouns, Emilio Kropff (Argentinian Spanish), who carried out a daring and still unpublished study of mass count syntax in Mapudungu, and Mohammed Katran (Iraqi Arabic), who set up an automated testing system to facilitate data collection. Limited financial support at one stage was provided by a MIUR 'Internazionalizzazione' grant.

ISSN 1450–3417

language do; in the Chinese language family, for example, all nouns are mass). For example, in English, mass nouns are associated with quantifiers like *little* and *much* and require a measure classifier (*kilos*, *boxes*) when used with numerals; on the other hand, count nouns are associated with determiners like *a(n)*, quantifiers like *many/few* or *each*, and can be used with numerals without a measure classifier.

These syntactic properties are intuitively correlated with semantic properties. Typical count nouns denote sets of individual entities, as in *girl*, *horse*, *pen*, while typical mass nouns denote 'substances' or 'stuff', for example, *mud*, *sand*, and *water*. It has often been noted that the correlation is not absolute, and that there are mass nouns which intuitively denote sets of individuals (e.g., *furniture*, *cutlery*, *footwear*). Nonetheless, the correlation seems non-arbitrary and there has been much discussion of this correlation in the linguistics literatures as well as in the psycholinguistics literature (e.g., Soja *et al*. 1991, Prasada *et al*. 2002, Barner & Snedeker 2005, Bale & Barner 2009) and in the philosophical literature (e.g., Pelletier 2011 and references cited therein).

Within the semantics literature, a seminal attempt to ground the syntactic distinction semantically is Link (1983). Link proposed that mass nouns are associated with homogeneity and cumulativity, while count nouns are associated with atomicity. Homogeneity, cumulativity, and atomicity are properties which can be associated with matter or with predicates. An object is atomic when it has a distinguishable smallest element which cannot be further divided without compromising the very nature of the object, and an atomic predicate denotes a set of atomic elements. Thus *boy* is an atomic predicate, since we can easily identify atomic boys, parts of which do not count as boys. Homogeneity is a property by which, when parts of an object are separated, each individual part holds the entire identity of the original object, and a homogeneous predicate is one which denotes entities (or quantities of matter) of this kind. For example, any part of something which is water is water, thus *water* is a homogeneous predicate. Cumulativity is the property that a predicate has if two distinct entities in its denotation can be combined together to make a single entity in the denotation of the same predicate. For example, if A is water and B is water, then A and B together are water. Cumulativity and homogeneity can be seen as different perspectives on the same phenomenon, though linguistic research has shown that the difference between them is important in certain contexts (see e.g., Landman & Rothstein 2012). However, for our purposes, we can ignore these differences. The generalization emerging from Link (1983) is that mass nouns are non-atomic and exhibit properties of being homogeneous and cumulative, whereas count nouns are atomic.

Link's proposal has been hugely influential, giving a representation to the intuition that the syntactic expression of the mass/count distinction correlates with a real semantic or ontological contrast. Expressions of this intuition are widespread. Thus Koptjevskaya-Tamm (2004) writes about the mass/count distinction: "In semantics, the difference is between denoting (or referring to) discrete entities with a well-defined shape and precise limits vs. homogeneous undifferentiated stuff without any certain shape or precise limits" (p. 1067).

Despite this ingrained intuition, it has been generally recognized that it is not possible to postulate a simple projection of the homogeneous/atomic or undifferentiated/discrete distinction onto mass/count syntax (seem e.g., some recent references such as Gillon 1992, Chierchia 1998, 2010, Barner & Snedeker 2005, Nicolas 2010, Rothstein 2010, Landman 2010, as well as Koptjevskaya-Tamm 2004). There are various pieces of evidence which show this. In the first place, there are mass nouns which denote sets of atomic entities, such as *furniture* and *kitchenware*, and some of these have synonyms in the count domain as in the English pairs *change*/*coin*(*s*), *footwear*/*shoe*(*s*), *carpeting*/*carpet*(*s*) which denote roughly the same entities. Conversely, there are also count nouns such as *fence* and *wall* which show properties of homogeneity (Rothstein 2010). Secondly, nouns stems may have both a count and mass realization in a single language, with the choice depending on context. In some cases, both count and mass usage are equally acceptable, as with *stone* and *brick* and *hair* in English. In other cases, one of the uses is considered non-normative, for example, when a count noun like *dog* is used as a mass noun in *After the accident there was dog all over the road*. Thirdly, items which are comparable in terms of lexical content do not have stable expressions cross-linguistically as either mass or count. The much cited examples is *furniture*, which is mass in English but count in French (*meuble*/*s*), while in Dutch and Hebrew, the comparable lexical item has both a mass and a count realization (Hebrew: count *rehit*/*im* vs. mass *rihut*, Dutch: count *meuble*/*s* vs. mass *meubiliar*).

The received wisdom therefore oscillates between these two perspectives, with much recent research trying to mediate between them, both capturing the basic generalization, while accounting for the variations both cross-linguistically and within a single language. Chierchia (2010) suggests that the mass/count distinction is based on whether or not the noun is envisaged to have a set of stable atoms. Rothstein (2010) argues that semantic atomicity is context dependent. Pires de Oliveira & Rothstein (2011) argue that the mass/count alternation is a reflection of whether the noun relates to its denotata as a set of entities to be counted or as a set of quantities to be measured.

However, in the midst of all this discussion, certain basic facts remain unclear. In particular, how great is the cross-linguistic variation in mass/count syntax? Clear evidence that the syntactic mass/count distinction is not a projection of a semantic or ontological distinction has stayed at the level of the anecdotal, with discussion focusing on a few well-known and well-worn examples (see, e.g., Chierchia 1998 and Pelletier 2010 for reviews). As a consequence, most discussions of the basis of the mass/count distinction have been based on some explicit and some tacit assumptions, which have not been verified empirically. In particular, it is often assumed that the mass/count distinction is essentially binary, that is, that a noun is classified as mass or as count or as ambiguous. (This is explicit in accounts which assume that nouns are labeled as mass or count in the lexicon, and implicit in accounts such as Borer (2005) which assume that noun roots are not classified lexically but naturally appear in either a count or a mass syntactic context.) Another, related, common assumption is that in a language with a mass/count distinction, most nouns are either mass or count, with the syntax reflecting the homogeneous/atomic distinction, and that cross-linguistic variation occurs in a lexically defined 'grey area' in the middle, which includes nouns which are not easily classifiable. But crucially, discussion of the facts of the matter has not gone far beyond the anecdotal. The semantics literature has discussed in great depth the syntactic properties of nouns like *furniture* and comparing it syntactically and semantically with its cross-linguistic counterparts, but despite very few more in-depth, but still narrow, studies (e.g., Wierzbicka 1988), we have little sense of how representative nouns like this actually are.

An answer to the question to what degree there is cross-linguistic variation in the expression of the mass/count distinction is essential to the discussion of its cognitive and semantic basis. If there is ultimately little cross-linguistic variation, then we are entitled to hypothesize that there may be some general strong correlation between properties of the denotata (e.g., as atomicity and homogeneity) and the grammatical distinction. In this case, the grammatical mass/count distinction may have a sound cognitive/perceptual foundation, and its semantic interpretation would reflect this. The task of linguistics would then be to characterise precisely the semantic basis of the grammatical distinction, to identifying 'exceptional' areas where the correlation does not hold and/or where cross-linguistic variation naturally appears, and to try and explain why these occur. This is an approach which has been exploited especially with respect to 'furniture nouns' which has been identified as 'super-ordinates' (Markman 1985) or functional artifacts (Grimm & Levin 2011). On the other hand, if cross-linguistic variation is wide, then the basis for assuming that there is a correlation between cognitive/perceptual features and the grammatical distinction is considerably weakened. Then questions that linguistics should be asking will depend directly on the nature of the patterns, or lack of them, that an analysis of the cross-linguistic facts of the matter reveals. The lack of any quantitive data on the extent of cross-linguistic variation is thus highly problematic.

With the goal of remedying this lack of data and contributing to understanding the cognitive aspects of mass/count syntax and the relation between grammatical, semantic, and cognitive differentiations in this domain, we have conducted a statistical cross-linguistic empirical study based on a quantitative approach, and also a corpus study on the Browns section of the CHILDES database (MacWhinney 1995). We hope with this to be able to begin to answer several basic questions: To what extent is the mass/count distinction a straightforward reflection of the semantic properties of nouns? Is the variability across languages in any degree predictable, or is the grammatical division into mass and count arbitrary? Furthermore, is the division into mass and count absolute, or are some nouns 'more count' or 'more mass' than others? Do differences in the semantic explanations essentially arise due to the multi-dimensional nature of the semantic (as well as the syntactic) space? And if so, can the multi-dimensional aspect provide useful insights in the acquisition of mass/count syntax in humans?

Our study aims to go some way to providing empirically substantiated answers to these questions. We carried out a relatively large scale analysis of the mass/count classification of nouns cross linguistically. Count nouns are usually distinguished from mass nouns by a number of different syntactic properties, for example, co-occurrence with numerical expressions, co-occurrence with distributive quantifiers like *each*, and so on, but the specific tests vary from language to language. We focused on several issues:

(i) To what extent can mass/count syntax be predicted in language A on the basis of knowledge of language B?

(ii) To what extent is mass/count syntax a binary division (i.e. if a noun classifies as count on one test, what are the odds that it will classify as count on all tests)?

(iii) To what extent can mass/count syntax be predicted on the basis of real-world semantic properties?

2. Methods

2.1. *Data Collection*

2.1.1. Noun List

Binary syntactic usage tables were compiled for a list of 1,434 common nouns in English, which included 650 abstract and 784 concrete nouns. The list was derived from a longer list of 1,500 very frequent English nouns, originally extracted from the CELEX database (see http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC96L14) for a different project, integrated with about 150 additional nouns often used in linguistics to study the mass/count domain, after translating the nouns into the five other languages included in our study, and eliminating over 200 nouns for which either the identification of the common semantic concept, or the syntactic classification in at least one language, as described below, were unclear or problematic. At the translation stage, each noun/concept was provided with a sample usage sentence, to disambiguate its potentially divergent meanings; thus trying to ensure that each language had the same semantic concept translated, for the same context, into a corresponding noun.

2.1.2. Usage Tables

A set of yes/no questions was then prepared, in each language, to probe the usage of the nouns in the mass/count domain. The questions asked whether a noun from the list could be associated with a particular morphological or syntactic marker relevant in distinguishing mass/count properties. Some questions were designed to give positive properties of count nouns (e.g., can N be directly modified by a numeral?) and some to give positive properties of mass nouns (e.g., can the noun appear in the singular with measure expressions?). Since the mass/count distinction is marked by different syntactic properties cross-linguistically, the questions were dependent on the particular morpho-syntactic expressions of mass/count contrast in each language. For example, in English we asked whether a noun could appear with the indefinite determiner *a(n)* but this was obviously an inappropriate question to ask in Hebrew where there is a null indefinite determiner. The questions in English are shown in Table 1 below.

The questions were answered by native speakers of each of the languages in our study. Thus each noun was associated, for each informant, with a string of binary digits, 1 indicating yes and 0 indicating no, reporting how that particular noun is used (or predominantly used) in the mass/count domain, by that informant. Such usage tables (a tiny portion of an English usage table is shown as Table 3 below) were compiled by Armenian, English, Hebrew, Hindi, Italian, and Marathi informants (at present, we have complete data for 16 informants; Armenian: AN, AR, GR, GY, RF; Italian: LE, FR, GS, RS, BG; Marathi: SN, TJ, SK; English: PN; Hebrew: HB; Hindi: MN). Although the choice of languages was ultimately determined by the available informants, the languages studied represent a spread across language families. The five Indo-European languages come from distinct branches: Germanic (English), Romance (Italian), Northern Indo-Aryan (Hindi), southern Indo-Aryan (Marathi), and Armenian, which constitutes a branch of its own. Hebrew comes from a distinct phylum, the Semitic family.

No.	Syntactic Questions
1.	Can the noun be used in bare form?
2.	Can the noun be used with *a/an*?
3.	Can the noun be pluralized (in a morphological distinct form)?
4.	Can it be used with numerals?
5.	Can the noun be used with *every/each*?
6.	Can the noun be used with *many/few*?
7.	Can the noun be used with *much/little*?
8.	Can the noun be used with *not much*?
9.	Can the noun be used with *a lot of*?
10.	Can the noun be used with a numeral modifier + plural on kind?
11.	Does the noun appear in the singular with a classifier or measure phrase?

Table 1: List of questions used in English to compile the usage table.
The questions probe whether a particular noun is associated with certain typical syntactic markers, important in English for the mass/count distinction. Similar questions were used for other languages, formulated according to the morphosyntactic properties of the languages in question. These are listed in tables A1–A5 in the Appendix.

2.1.3. Semantic Table

A similar table was prepared by five informants (KM, RI, SL, SU, and TJ, four native Marathi and one Hindi speaker) using the English database to describe the properties of the denotations of the nouns in the list. These questions probed aspects of the denotations which were plausibly related to the more general semantic properties of atomicity, homogeneity and cumulativity discussed above. The questions asked (also supplied with an example to each, to clarify the meaning) are shown in Table 2. The questions were purposely formulated in informal terms, since we were interested in the correlation between mass/count syntax and what is often taken as the 'intuitively obvious' basis for the distinction. We will somewhat loosely refer to these as 'semantic questions'.

No.	Semantic Questions
1.	Is it Alive irrespective of context?
2.	It is an Abstract Noun?
3.	Does it have a single Unit to represent itself ?
4.	Does it have a definite Boundary, visually or temporally?
5.	Does it have a stable Stationary shape (only if concrete)?
6.	Can it Flow freely (only if concrete)?
7.	Does it take the shape of a Container (only if concrete)?
8.	Can it be Mixed together indistinguishably (only if concrete)?
9.	Is the identity Degraded when a single unit is Divided (only if concrete)?
10.	Can it have an easily defined Temporal Unit (only if abstract)?
11.	Is it an Emotion /Mental process (only if abstract)?
12.	Can it have an easily defined Conceptual Unit (only if abstract)?

Table 2: Questions used to probe the semantic properties of the nouns.

The questions are based on the properties of atomicity, homogeneity and cumulativity, if nouns are concrete. For abstract nouns, the semantics is based on how easy it is to define a unit of the concept. The questions were asked without elaboration, with only a reference example; in the case of question 8, for example, applicable to concrete nouns: Can it be mixed together indistinguishably? [e.g., butter *as opposed to* man*].*

Both syntactic and semantic tables were then processed through the analysis described below.

2.2. *Analysis*

Nouns in the syntactic usage table of a particular informant were clustered together according to the binary string associated with them. In this way, nouns which have the exact same binary string are grouped together, reflecting the fact that their mass/count syntactic behavior is (considered by that informant to be) the same. Thus each group formed in the usage table is identified with a unique binary string. Informants for each language of course group the nouns according to their own syntactic rules, hence the clusters formed in different languages inform us about mass/count phenomenology in that language. The same grouping procedure can be applied to the semantic table, generating 'semantic classes' (relative to the main features putatively underlying mass/count syntax across languages). The resulting distributions of nouns/concepts in syntactic or semantic classes were analyzed, with the measures described below, for both syntactic and semantic tables.

Noun	Context	1	2	3	4	5	6	7	8	9	10	11
ability	Ability is more desirable than wealth.	1	1	1	1	1	1	1	1	1	0	0
accident	The crash was an accident, not intentional.	0	1	1	1	1	1	0	0	0	0	0
acid	Acid stains clothes.	1	1	1	1	1	1	1	1	1	1	1
act	The flood was an act of nature.	0	1	1	1	1	1	0	0	0	0	0
act of crime	Murder is always an act of crime.	0	1	1	1	1	1	0	0	0	0	0
activity	A favorite activity was spitting cherry stones.	1	1	1	1	1	1	1	1	1	1	0
actor	Any good actor can play Tarzan.	0	1	1	1	1	1	0	0	0	0	0

Table 3: A small section of the usage table for English as filled by a native informant. Numbers in the top row refer to the syntactic questions in Table 1.

2.2.1. Hamming Distance Scale

The data in the usage tables is in principle high-dimensional, containing distinct contributions from each of several syntactic markers. It is possible, however, that much of the relevant mass/count syntax might be organized along one main dimension. We consider the hypothesis that this most important dimension may be defined as the 'distance' from a pure count string, where nouns at different distances might be associated with characteristic combinations of syntactic markers (see Fig. 1 below).

To probe this potential organizing dimension, the high dimensional data is collapsed onto a single dimension. This is obtained by calculating the Hamming distance, or fraction of discordant elements, of each noun (i.e. of each syntactic group) from a bit string representing a pure count noun. A pure count string is one which has 'yes' answers for all count questions and 'no' answers for all mass questions. Hence a noun that has distance 0 from a pure count string is a proper count noun, whereas a noun with all its bits flipped with respect to a pure count string is a mass noun, and has a normalized distance of 1 from the pure count string. Such a noun has answers 'no' to all count questions and 'yes' to the mass questions. By plotting the distribution of nouns on this dimension we expect to be able to visualize the main mass/count structure, to relate easily with a linguistic interpretation. This measure does not strictly reflect the categorical nature of groups defined by a unique syntactic string, in the sense that all nouns with a syntactic string differing at 3 bits from the pure count string are clustered together, irrespective of which are the 3 syntactic markers for each noun. This allows for a coarser but perhaps more intuitive and linguistically more transparent comparison between languages than the mutual information measure discussed below, which is a fine-grained comparison between languages, taking into account all the existing dimensions.

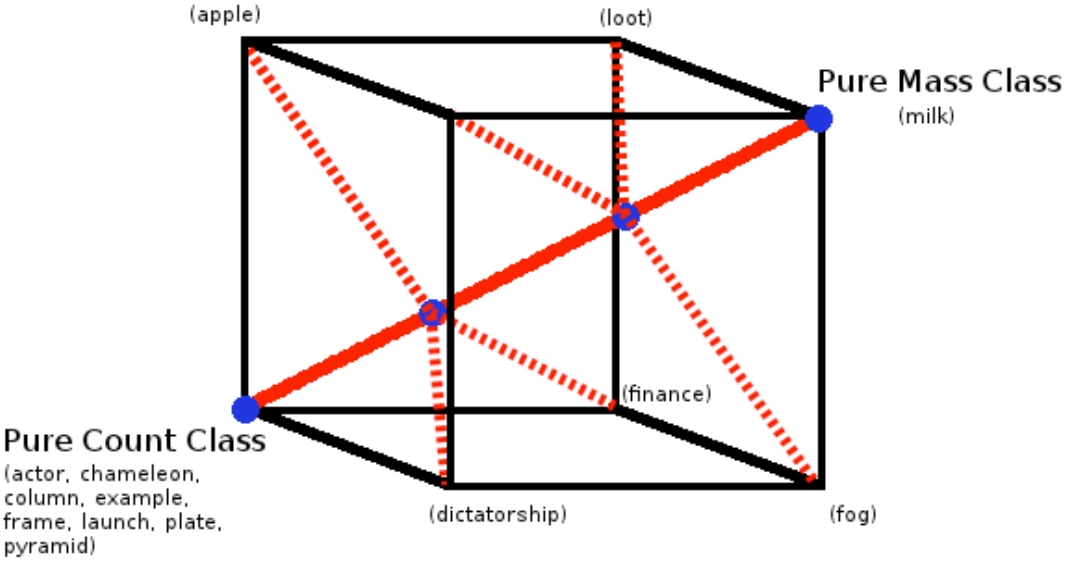

Figure 1: Schematic representation of the Hamming distance scale.

Nouns are located in an N-dimensional space (here only three dimensions are represented) and the Hamming Distance scale projects these points onto the mass/count dimension (red diagonal), going from the bit string of pure count to that of pure mass.

Agreement between two languages is estimated as a variance measure, $\langle x^2\rangle + \langle y^2\rangle - 2\langle xy\rangle$ which is simply a sum of squares of the difference between the Hamming distances *x* and *y* of a noun from the pure count class, as found in the two languages concerned. This measure has a strict upper bound of 1, if Hamming distances are expressed as fractions of discordant bits, which is attained when each noun is either pure count in one language and pure mass in the other, or vice versa; clearly a rather implausible occurrence. A more natural reference value, although not strictly speaking an upper bound, can be estimated by calculating the variance measure between the Hamming distances in a language and those of randomly shuffled nouns in another language, $\langle x^2\rangle + \langle y^2\rangle - 2\langle x\rangle\langle y\rangle$. The random shuffling simulates the case of a total absence of any relation between the position of the nouns along the main mass/count dimension in the two languages, while respecting the distribution of Hamming distances in each. Thus by comparing the actual value with the reference value, we can get an understanding of how the languages match each other in broadly classifying nouns on the main mass/count dimension. Each language however has different number of questions analyzing its mass/count structure and hence the Hamming distance space for a language is populated only at intervals of 1/Nth of a bit, where N is the number of questions in a language. To minimize the effect of different intervals we estimate a true minimum of variance between languages (which in an ideal case is 0) by calculating the variance between two languages when all the nouns are ordered in the same way in their position on the Hamming distance scale. We adjust the raw variance by simply subtracting the minimum variance for that pair, and then normalize it by dividing it by the (adjusted) effective maximum value as mentioned above.

2.2.2. Clustering and Information Measures

Information theory provides us with useful tools to quantify aspects of the clustering observed in the data. The entropy of a variable, which can take a certain set of values, quantifies the uncertainty in predicting the value it can take in terms of its possible values and their probabilities. A variable which always takes a single value is perfectly predictable and has an entropy of 0 bits. A binary variable has an entropy of 1 bit when it has 50% probability to take either value, e.g. 1 or 0. We can apply this measure to the grouping structure formed around the mass/count distinction in the languages we study. In our case, the variable *G* is which group any given noun or concept has been associated to in a particular table, taking values 1,…,*i*,…,*n*, where *n* is the total number of groups observed in that table. The probability *p*(*i*) is determined for our purposes as the relative frequency of nouns/concepts assigned to group *i*. The entropy of the table is then calculated as:

$$H(G) = -\sum_{i=1}^{n} p(i)\log_2 p(i)$$

H(G) informs us about the overall syntactic variability expressed (by an informant) in a language, and can be regarded as the logarithm of an equivalent number of significant syntactic classes.

To make cross lingual comparisons, we quantify the extent to which the groups formed by informants in one language overlap with the groups formed by those in another. This amounts to defining equivalence classes, whereby two nouns are grouped together if and only if they are members of the same syntactic usage group in the two languages. For example, if the nouns *water* and *wine* are a part of the same group in language X and also fall in one group in language Y, whatever the syntactic usage questions that define groups in the two languages, they are members of the same equivalence class. For analyzing syntactic-semantic relations, language Y is replaced by the semantic table. To give a limiting case, if two languages were to behave exactly the same in classifying nouns in the mass/count domain, the equivalence classes would coincide with the groups formed in the individual languages, reflecting the exact match between groups produced by language X and Y. At the other extreme, if two languages were to share no commonality, there would be no relation whatsoever between the groups in the two languages, and membership in a group in one language would not be informative about membership in the other language.

The mass/count similarity between X and Y can be quantified by the mutual information *I(X;Y)*, a measure that quantifies the mutual dependence of two variables. If two variables share no common information then the mutual information between them is 0, which is the lower bound for *I*, whereas the upper bound on mutual information is the lower between the entropies of the two variables (the shared information between two variables cannot be more than the total information content in one variable, i.e. its entropy). Mutual information is calculated using the joint entropy of the two variables in question, which in our case is the entropy of the groups, by the relation

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

which can be written also

$$I(X;Y) = \sum p(i,j)\log_2\left[p(i,j)/\,p(i)p(j)\right]$$

and where *H(X,Y)* is the joint entropy of the two variables, at least equal to the higher of the two individual entropies. In the limit case in which the syntactic groups are identical, *H(X)=H(Y)=H(X,Y)=I(X;Y)*, whereas in the opposite limit case, in which there is no relation whatsoever between the groups each table, *p(i,j)=p(i)p(j)*, expressing independent assignments, and then *H(X,Y)=H(X)+H(Y)*, so that *I(X;Y)*=0.

Mutual information measures suffer from a bias due to limited sampling (Panzeri & Treves 1996) related to the number of equivalence classes actually occupied compared to the total possible ($2^{\mathrm{Nq1}} \times 2^{\mathrm{Nq2}}$) classes, where Nq1 and Nq2 are the number of questions for the two languages in the pair. The correction to mutual information is estimated by calculating the mutual information between the pairs of languages when the nouns for one pair are randomly shuffled, thus simulating the lack of correlation between the two languages, and then averaging the value over 50 such shuffles. The correction is then subtracted from the raw value calculated for a pair.

2.2.3. Artificial Syntactic String Generation

To test the importance of the mass/count dimension and its link with semantics, an artificial syntactic usage table was also generated, wherein the 'yes/no' decision to a syntactic question was decided by a stochastic algorithm based on the position of the noun on the main semantic mass/count dimension. This algorithm generates a 0 or 1 for each of a string of N_l 'pseudo-syntactic' questions, one string per language, where N_l is the number of syntactic questions in that language. To do so, it uses two reference points, namely the syntactic pure count string for that language and the position of the noun concept along the semantic mass/count dimension, which is taken to be a language universal. The latter is quantified by the Hamming distance from the pure count semantic string, i.e. by the fraction d=D/N of semantic features that differ, for that concept, from those of the pure count. Each bit of the artificial string is then assigned, one by one, for a given noun, the value the bit has in the pure count string with probability (1–d), and the other value with probability d.

Syntactic questions, for this purpose, are empty of content, and simply refer to distinct bits of a pseudo-syntactic usage string. Such bits are determined, for a particular language, by the specific configuration of the pure count string for that language. If the noun is semantically close to a pure count then the probability to generate a syntactic pure count, or something close to it, is higher. The Hamming distances of the artificial strings from the pure count string have a certain distribution (a convolution with exponentials of the semantic Hamming distance distribution) which resembles that of the real syntactic strings, in most cases (except for Marathi, see below); while the position of each noun along the artificial syntactic mass/count dimension is strongly correlated with the position of the noun along the semantic mass/count dimension. The variance measure between the pseudo-usage table of any language and the semantics table provides us with a lower reference value for the variance itself, in contrast to the upper reference value obtained by random shuffling of the nouns. We are then able to better gauge the significance of the mass/count dimension and the importance of semantics with respect to the mass/count syntax. Also, the mutual information between natural usage tables and semantics can be compared to the mutual information between the pseudo-usage table and semantics, to allow a better estimate of what is the contribution of sheer semantics to the mass count syntax (by providing what for the mutual information scale is a more realistic upper value, see Fig. 12 below). The entropy for a particular language depends also on the number of questions used to investigate the mass/count syntax. By looking at the entropies of the artificial syntax we can see how the entropy measure scales with the number of questions.

2.3. Corpus Study of the Mass/Count Distinction in English

Brown's section of the CHILDES corpus was also used, in an additional component of the study, to obtain mass/count information about nouns occurring in a natural English language corpus. For this purpose all nouns were collected, in the adult-produced sentences of the corpus, which co-occurred with a set of predefined mass/count markers. The co-occurrence frequency of a noun and the set of mass/count markers was recorded and normalized to the total occurrence frequency of the noun. Thus, for each noun, there was a set of numbers which indicated the statistical distribution of syntactic markers for that noun. The markers that were used to measure co-occurrence frequency were *a(n)*, *every/each*, pluralization, *many, much, some* + sing. N, and *a lot of* + sing. N. This study contains a total of 1,506,629 word tokens and 27,304 word types.

The usage table obtained from the CHILDES corpus was analyzed with multi-dimensional scaling, and the distribution of the nouns on the mass count dimension. Multi-dimensional scaling projects high dimensional data on a lower dimensional space while preserving the inter-data-point distance, allowing to visually identify structural information in the data. By analyzing the distribution and clusters in the projected space one can gain information about statistically important dimensions and markers. Moreover, the data from the CHILDES corpus was analyzed in terms of distribution of distances from the pure count class and of entropy measures, after binarizing the table indicating the frequency of each marker. Thus, for example, if a noun was found at least once in plural form, this was taken as evidence that it could be pluralized; if found at least once with *a* or *an*, that it could take the indefinite article, and so on. In this way, the same analyses could be applied as for our database.

3. Results

3.1. *Individual Syntactic Rules and Semantic Attributes Do Not Match*

The starting point of our analysis is the observation that, at least in five out of the six languages we considered, roughly half the nouns in the sample can be easily classified as pure count nouns. The exact numbers in each language are Armenian: 1058, English: 693, Hebrew: 757, Hindi: 994, Italian: 863, and Marathi: 255. For example, in both Italian and English nouns like *act, animal, box, country* (as the territory of a nation), *house, meeting, person, shop, tribe, wave*; and *accident, cell* (as in biology), *loan, option, pile, question, rug, saint, survey, zoo* were classified as count in all respects by our informants. In Marathi, while the first 10 examples were also classified as count, the second 10 tested positive on all count properties except one, usually the property of having a morphologically distinct plural form. Marathi appears to stand out from the group in other ways, as reported below. For all other languages, clearly the focus has to be on the remaining proportion of non-pure-count nouns.

Among the informant responses, we observed cases of nouns that were regarded as pure count in English but cannot be normally used with numerals in Italian (*back, forum, grin*), or vice versa that test as pure counts in Italian but cannot be normally used with numerals in English (such as *behavior* or *disgrace*). Interestingly, when considering only usage with numerals and with distributive *each/every* (*ogni* in Italian), our informants classified as 'count' in English nouns the translation of which failed both tests in Italian: *love, noon, youth* have count usages in English, but not in Italian. The converse is also found: there are count nouns in Italian that, translated into English, failed both the numerals test and the test "can be used with *each/every*": *advice, blame, literature, trust, wood*. There were cases where the impression of one of the authors was that his or her judgment might differ from the informants, or the informants disagreed among themselves. Since we are interested in this study in an overall quantitative analysis of cross-linguistic usage judgments, we did not subject these differences of judgments to in-depth linguistic analysis, but entered the judgments of the majority. We note that that there were a significant number of such cases. Overall, there were only 116 nouns that were classified as pure count in all six languages, and still only 392 when excluding Marathi. We thus proceeded to a quantitative analysis, without further questioning the responses by the informants on a noun-by-noun basis.

For a quantitative analysis, we first assessed whether, in any of the languages in the database, a particular syntactic usage rule can be taken to reflect in a straightforward manner a particular semantic attribute of the noun. While in many cases the yes/no answer to a syntactic question turns out to be significantly or highly significantly correlated with a specific semantic attribute, we found no cases where the correspondence could be described as expressing a 'rule', even a rule with a few exceptions. To present quantitative results, we focused on cases where the semantic-syntactic correspondence was higher. The notion of high correspondence is somewhat arbitrary, because for example, one may contrast a case where among 10% of nouns with a particular semantic attribute, 90% admit a certain syntactic construct, with another case where those proportions are 30% and 70%. In our sample, the first 'quasi-rule' appears stricter, but it applies to only 129 nouns in the sample, whereas the second one, while laxer, applies to 301 nouns. For consistency with later analyses, we focus on relative (normalized) mutual information as a measure of correspondence, while reporting also the number of nouns for which syntax matches semantics. The relative mutual information measure ranges from 0 to 1 and it quantifies the degree to which the variability in the syntax, across nouns, reproduces that in the semantic attributes, both of which are quantified by entropy measures.

Language	++	+–	–+	—	H(Lang)	H(Sem)	MI(S,L)	Norm MI
Armenian	24	31	686	43	0.451	0.366	0.080	0.218
Italian	26	29	662	67	0.536	0.366	0.053	0.145
Marathi	25	30	559	170	0.819	0.366	0.020	0.054
English	29	26	668	61	0.503	0.366	0.046	0.126
Hebrew	29	26	682	47	0.447	0.366	0.055	0.150
Hindi	28	27	686	43	0.434	0.366	0.062	0.170

Table 4: A case of relatively high correspondence between a semantic attribute and a syntactic rule. Semantic question 8, applied only to 784 concrete nouns, asked whether the noun denotes an entity (or individual quantity) that can be mixed with itself without changing properties. (This somewhat loosely phrased question makes reference to the homogeneity and cumulativity properties discussed in section 1, since it can be interpreted either as asking whether proper parts can be permuted without changing the nature of the object, or whether instantiations can be collected under the same description.) The syntactic question considered was whether the noun can be used with numerals, and it was present in all languages. The largest group of concrete nouns, in the –+ class, denote objects that are not homogeneous, and the nouns can be used with numerals. The relative proportion of nouns in each of the four classes, however, yield meager normalized information values, indicating that individual attributes are insufficient to inform correct usage of specific rules, even in this 'best case' example.

Table 4 shows that most concrete nouns in our database (729/784) denote entities that, according to our informants, cannot be 'mixed' while retaining their properties as instantiations of the noun. Most of these nouns can be counted in the sense that they can be preceded with numerals, across languages (with a somewhat less disproportionate bias in Marathi). Nevertheless, among the nouns for which the answer to question 8 was positive, i.e. that displayed properties of either cumulativity or homogeneity, roughly half can be used with numerals, again across languages, yielding rather low values of mutual information between semantics and syntax, as quantified in the last column of the table. Normalized MI values are much closer to zero than to one.

Even though the correspondence with the particular semantic attribute of cumulativity is low, the results above suggest that there might be a high degree of correspondence among the syntactic usage with numerals across languages, at least when excluding Marathi. After all, across languages it is roughly half the nouns denoting entities which intuitively are cumulative, which can be used with numerals, and half which cannot. Is it roughly the same half?

Language pair	++	+–	–+	—	H1	H2	I(1:2)	Norm. MI
Arm–Ita	662	48	26	48	0.451	0.536	0.124	0.275
Arm–Mar	560	150	24	50	0.451	0.819	0.059	0.131
Arm–Eng	666	44	31	43	0.451	0.503	0.106	0.235
Arm–Heb	675	35	36	38	0.451	0.447	0.095	0.212
Arm–Hin	683	27	31	43	0.451	0.434	0.129	0.297
Ita–Mar	548	140	36	60	0.536	0.819	0.062	0.115
Ita–Eng	654	34	43	53	0.536	0.503	0.131	0.261
Ita–Heb	652	36	59	37	0.536	0.447	0.068	0.152
Ita–Hin	661	27	53	43	0.536	0.434	0.102	0.235
Mar–Eng	547	37	150	50	0.819	0.503	0.041	0.082
Mar–Heb	553	31	158	42	0.819	0.447	0.034	0.076
Mar–Hin	556	28	158	42	0.819	0.434	0.037	0.086
Eng–Heb	669	28	42	45	0.503	0.447	0.119	0.266
Eng–Hin	675	22	39	48	0.503	0.434	0.144	0.331
Heb–Hin	675	36	39	34	0.447	0.434	0.078	0.181

Table 5: The correspondence between languages is not higher.

In the same case of relatively high correspondence between a semantic attribute and a syntactic rule, entropy and mutual information between languages yield the relatively low normalized MI values listed in the fifth column, which indicate that a broadly applicable syntactic question ("Can the noun be used preceded by a numeral?") selects different subsets of nouns across different languages.

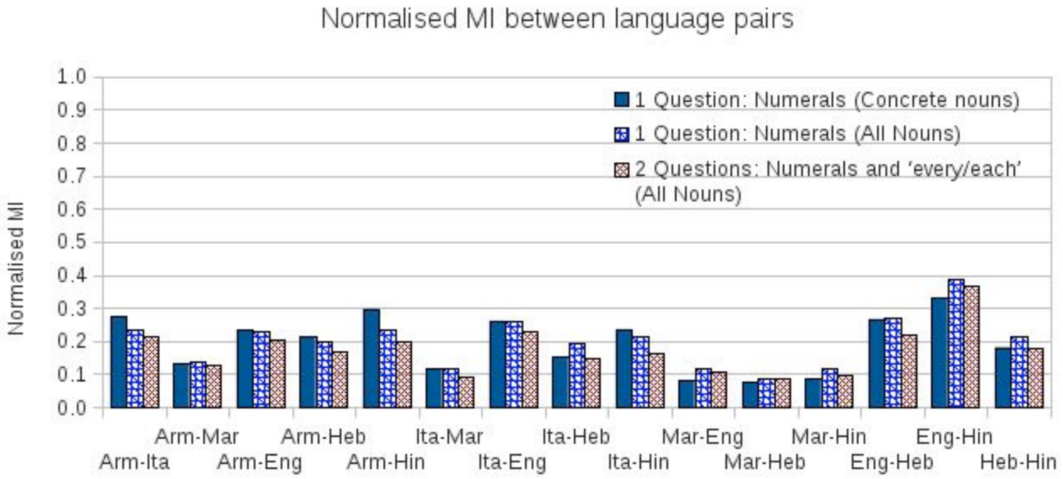

Figure 2: Agreement across languages remains low, however it is measured.

The solid bars show the normalized mutual information between pairs of languages for a single question, on usage with numerals, for concrete nouns only. The stippled bars are for the same measure over all nouns in the database, both concrete and abstract. The patterned bars are for pairs of questions, on both the use of numerals and that of distributive quantifiers such as each/every in English (see text).

Table 5 and Figure 2 show that the naïve expectation is not met by the data. The syntactic correspondence in the usability with numerals is weak across languages, even irrespective of any semantic attribute it may originate from. The congruence (number of concrete nouns in the same syntactic class when translated across languages) appears relatively high, because most nouns can be used with numerals anyway, but, properly quantified in terms of normalized mutual information, the degree of correspondence even excluding the special case of Marathi is roughly in the 15–30% range, with English and Hindi reaching a peak value of 33%. When considering all nouns in the database, including abstract nouns, the degree of correspondence does not change much (stippled bars in Fig. 2). Again excluding the special case of Marathi, it falls roughly in the 20–27% range, with English and Hindi reaching a peak value of 39%.

One may ask whether the low MI values with semantics, in Table 4 above, may be due to the lack of exact match between the semantic attribute considered and the specific syntactic rule. Similarly, one may ask whether the weak correspondence in the pattern of usage with numerals may also be due to the fact that numerals might point in different directions, so to speak, in the syntactic space of each distinct language, for example, atomicity vs. non-homogeneity. To approach these issues, we have begun by considering pairs of attributes, and pairs of syntactic rules. The degree of correspondence of each language with semantics does not change much, and in fact it tends to slightly decrease. For example, when asking whether the object denoted by the noun can flow freely, and also whether it is cumulative, and on the other hand whether the noun can be used with numerals and whether it can be used with distributive quantifiers like 'each' in English, we find that the normalized MI decreases with respect to the above analysis with one attribute and one rule, in all cases except for Marathi (data not shown). The decrease is entirely due to the increase in the entropy that appears in the denominator of the normalization (see Methods). In terms on non-normalized mutual information, instead, adding dimensions reveals perforce more variability.

Similarly, the match between languages, independently of semantic attributes, does not increase when considering two syntactic rules instead of one. Table 6 and Figure 2 report the data, this time for concrete and abstract nouns together, when considering the two syntactic rules above.

Table 6 shows that normalized mutual information values are low, all below 0.23 except for the English–Hindi match, even though 'congruence' values appear high. Congruence is the sum of the number of nouns that are used in the same way in both languages, with respect to the two syntactic constructs considered. Except for pairs including Marathi, between 70–81% of nouns are congruent across pairs. Yet mutual information is low because many of the congruent nouns are simply pure count nouns in either language, accepting both numerals and distributives, and their permanence in the largest class is not very informative about mass/count syntax in the other classes. As considering two questions rather than one does not affect results, it is interesting to ask what happens when considering all available questions together. We first focus on the main mass count dimension.

Language pair	H1	H2	I(1:2)	Norm. MI	Congruency
Arm–Ita	0.862	1.129	0.186	0.215	1119
Arm–Mar	0.862	1.427	0.109	0.127	825
Arm–Eng	0.862	0.872	0.176	0.204	1154
Arm–Heb	0.862	0.940	0.143	0.166	1152
Arm–Hin	0.862	1.242	0.172	0.200	1046
Ita–Mar	1.129	1.427	0.106	0.094	849
Ita–Eng	1.129	0.872	0.199	0.228	1132
Ita–Heb	1.129	0.940	0.141	0.150	1081
Ita–Hin	1.129	1.242	0.182	0.161	1011
Mar–Eng	1.427	0.872	0.094	0.108	882
Mar–Heb	1.427	0.940	0.082	0.087	826
Mar–Hin	1.427	1.242	0.122	0.098	812
Eng–Heb	0.872	0.940	0.191	0.219	1157
Eng–Hin	0.872	1.242	0.320	0.367	1099
Heb–Hin	0.940	1.242	0.169	0.179	1037

Table 6: Congruency and mutual information between languages.

The correspondence between languages is not higher when considering pairs of rules at a time. Here we considered whether a noun can be used with numerals, and whether it can be used with a distributive quantifier such as each/every *in English.*

3.2. *Hamming Distance*

Plotting the data on the main mass/count dimension (Fig. 3) as the distance from the pure count string shows that a very high proportion of the nouns are at a distance zero from the pure count class (groups are labeled from 1 to N+1 at an increasing Hamming distance of a single bit, where N is the number of questions and group 1 represents pure count nouns). Overall there is an exponential-like decreasing trend in the group frequencies, as we go further from the pure count, for all languages but Marathi. Since this measure does not distinguish between different classes that are at the same distance from the pure count class but vary in the questions that define them, we use different colors in the bars to show the proportions of particular classes at that specific distance from the pure count class. The number of questions, N, is 9 for Armenian, 8 for Italian, 5 for Marathi, 11 for English, 9 for Hebrew, and 5 for Hindi. Since there are N+1 possible groups, we see that for Italian the 9th group is empty, whereas for Hebrew the last two groups are empty.

Distributions in Figure 3 seem to reflect the nature of the nouns as brought out by the questions used to investigate them. In the case of Marathi, the distribution is seen to have two groups of high frequency, at the distance of 1 bit from the pure count class and mass class, respectively. In each of these high frequency groups there is one class that accounts for most of the nouns. The class making up most of 5th group differs from the pure mass class in answering 'no' to the question regarding use of measure classifiers. Upon closer inspection, we find that, out of the 411 nouns that form the largest class in the 5th group, 332 are abstract nouns, hence answering 'no' to the measure classifier question. The question that differentiates, instead, the largest class in the 2nd group from the pure count class is 'Pluralization with morphological change', to which for nouns in the largest class the answer is 'no'. Figure 4 shows the same distribution, only for Marathi, but restricted to the 650 abstract and to the 784 concrete nouns, respectively. Notice the changes in the frequencies of the 5th group for both concrete and abstract nouns, as compared to Figure 3. For other languages, the distributions restricted to concrete and to abstract nouns look similar to the overall distribution, with a quasi-exponential downward trend (not shown).

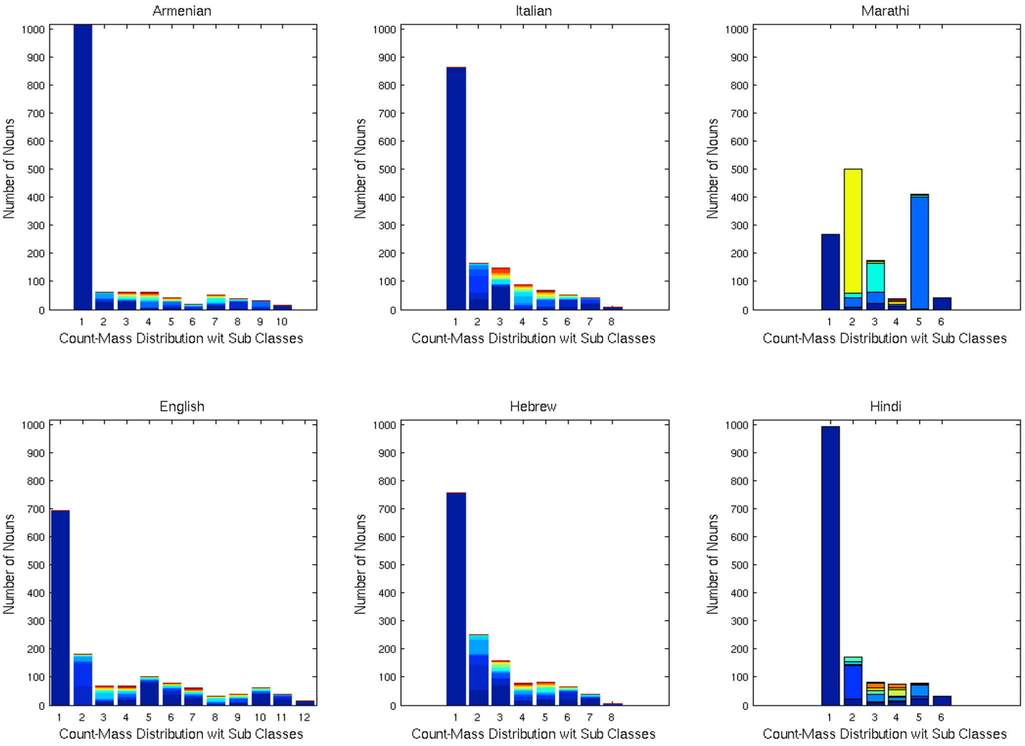

Figure 3: Distribution of nouns along the mass/count dimension.

Each histogram reports the frequency of nouns in the database, for a particular language, at increasing distances from pure count usage (1) and towards pure mass usage (N+1), where N is the number of syntactic question for the language. Colors in the bars indicate the proportion of nouns in each of the syntactic classes occurring at the same Hamming distance from the pure count.

In summary, the distribution of mass/count syntactic properties is undoubtedly graded rather than binary, as might have been intuitively expected. Most common nouns are strictly count in nature, in five of the six languages considered, with mass features increasingly rarer as they approach the pure mass ideal. Marathi differs from the other languages, and it remains of be examined whether it is representative of several other natural languages not considered in this study.

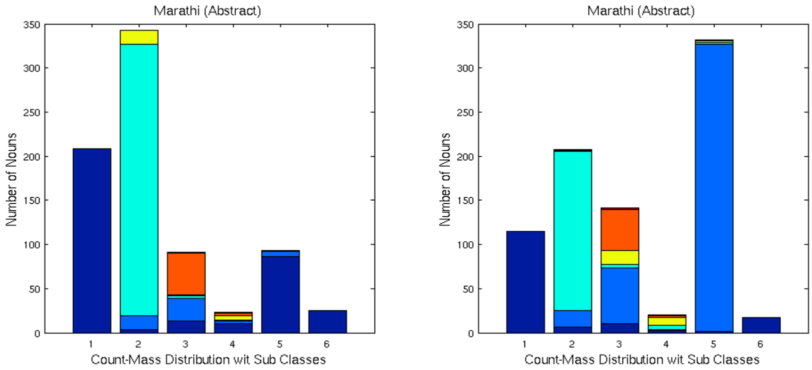

Figure 4: The distribution for Marathi, restricted to concrete and to abstract nouns.

As the 5th group in the histogram in Figure 3 (top right) includes mostly abstract nouns, it dominates the abstract noun distribution (right), while it is considerably reduced for concrete nouns (left).

How distributed are the semantic features characterizing these same nouns? Figure 5 shows that, for concrete nouns, semantic features define a monotonically decreasing distribution from the pure count class, roughly similar to that observed for syntactic usage features in five out of six languages. *Prima facie*, this might suggest that concrete semantics might be the common source that influences the global structure of the mass/count classification, at least for concrete nouns. For abstract nouns, the two semantic questions considered leave most nouns in the 'ambiguous group' — in particular, in the class which includes abstract nouns without an easily definable conceptual unit or a temporal unit. Since the semantics of abstract nouns does not have as clear a definition as concrete nouns, it may not have a strong independent influence on the mass/count syntax of abstract nouns.

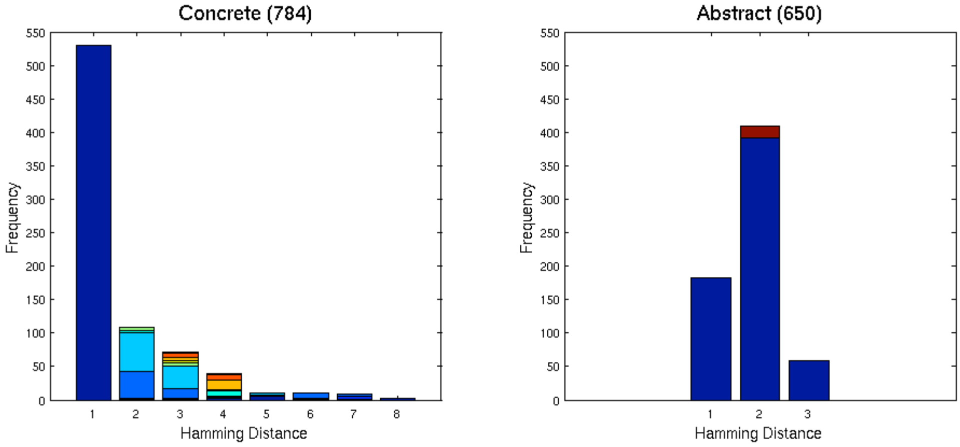

Figure 5: Distribution of nouns on the main mass/count dimension for semantics.

Given the different applicable semantic questions, it is shown separately for concrete (left) and abstract nouns (right). Concrete nouns show an exponential like shape similar to most of those in Figure 3, whereas most abstract nouns are in the ambiguous group.

If semantics serves as the common source of the mass/count syntax for concrete nouns, we expect not only the distributions to look similar overall, but also to include individual nouns at similar positions along the main mass/count dimension, both when comparing semantics with syntax for each of the 'well-behaved' languages, and when comparing the syntax of two such languages. This can be assessed through our variance measure, which quantifies the overall difference between such positions.

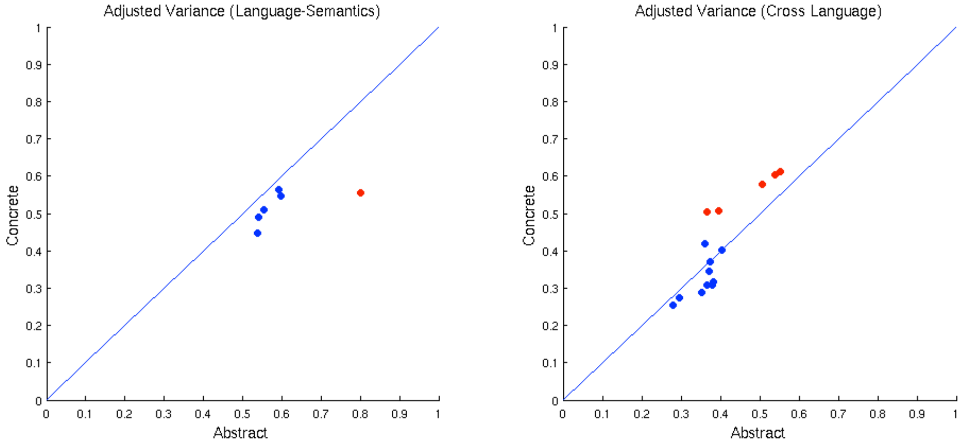

Figure 6: Scatter plots of variance values along the main mass/count dimension.

The adjusted and normalized variance (see Methods) in the position of individual nouns is shown between semantics and syntax (left) and between syntax in pairs of languages (right). In each plot, the adjusted and normalized variance for concrete nouns is on the y-axis and for abstract nouns on the x-axis. Pairs that include Marathi are indicated in red.

This expectation is only weakly borne out by the data. Figure 6 shows our relative variance measure, which is normalized to range between zero (when individual nouns are identically ordered in terms of their distance from the pure count class, either in both of two languages or between semantics and the syntax of one language) and one (when the relative orders are completely unrelated to each other). For concrete nouns, Figure 6 (left) shows that relative variance from semantics hovers around 0.5, halfway to complete lack of any relationship. For abstract nouns, variance values from semantics are somewhat higher. Marathi is an outlier, with yet higher variance from semantics, for abstract nouns. It appears therefore that the relationship to the underlying semantics is not very strong, even when considered solely along the main mass/count dimension.

Between languages, the adjusted variance is less than 40% of the adjusted maximum for all the pairs except for those including Marathi, both for concrete and abstract nouns. Marathi is an outlier, with higher variances from semantics (for concrete nouns) and from other languages. Marathi has higher variance since it does not have the exponential-like distribution as the rest of the languages. The variance values calculated over the entire database tend to be, for each pair of languages, close to the average between the values calculated over concrete and over abstract nouns, separately (not shown). The fact that between languages variance values are relatively low, relative to those from semantics, may indicate that there is an overall agreement across languages in classifying the nouns in the mass/count domain. This is shown only a gross level, however, along the main mass/count dimension, since here we do not take into account the fine grained differences between the classes at the same distance from the pure count class.

3.3. Mutual Information along the Main Mass/Count Dimension

The results from the analysis of the variance can be verified by considering an alternative measure of the correspondence in the classification, the mutual information. Along the main mass/count dimension, the mutual information can be calculated, e.g. between two languages, by grouping nouns at each Hamming distance from the pure count nouns, rather than each syntactically defined class. The mutual information (Table 7) ranges between zero (when there is no correspondence whatsoever in the groupings) and the minimum of the two entropy values, where entropy is calculated also by putting together all nouns in a group, i.e., at the same Hamming distance from pure count nouns.

Language	**Entropy**
*Armenian	1.63
*Italian	1.96
*Marathi	2.15
English	2.66
Hebrew	2.11
Hindi	1.54
*Semantics	(2.01) 1.58 (C) 1.24 (A)

Table 7: Language–entropy relations

*Entropy values along the main mass/count dimension in the six languages, and for semantics. The * sign indicates an 'average' over five informants (three for Marathi), taken by assigning to each question and each noun the yes/no answer chosen by the majority. For semantics, the overall value (in parenthesis) has little significance, because concrete nouns are assigned to eight distinct groups and abstract to only three, and combining them distributes the abstract nouns into the two extreme concrete groups and one central group.*

Figure 7 confirms, on the different quantitative scale of mutual information measures, the results obtained with the analysis of variance. The normalized mutual information with semantics is quite low, and lower for abstract than for concrete nouns, corresponding to higher variance values. It is at its lowest, 0.016, for abstract nouns in Marathi, which had the highest variance values. Between languages, mutual information is somewhat higher, and not markedly different between abstract and concrete nouns.

To better appreciate the significance of the relatively high variance values we measured, and of the relatively low MI values, we contrasted the values obtained between different languages with those obtained 'within' languages, i.e. measuring the correspondence between different informants of the same language. These data are available for five informants each for Armenian and Italian, and three for Marathi, and also five for the semantics classification. They give thus rise to 10 informant pairs in three cases and three pairs for Marathi.

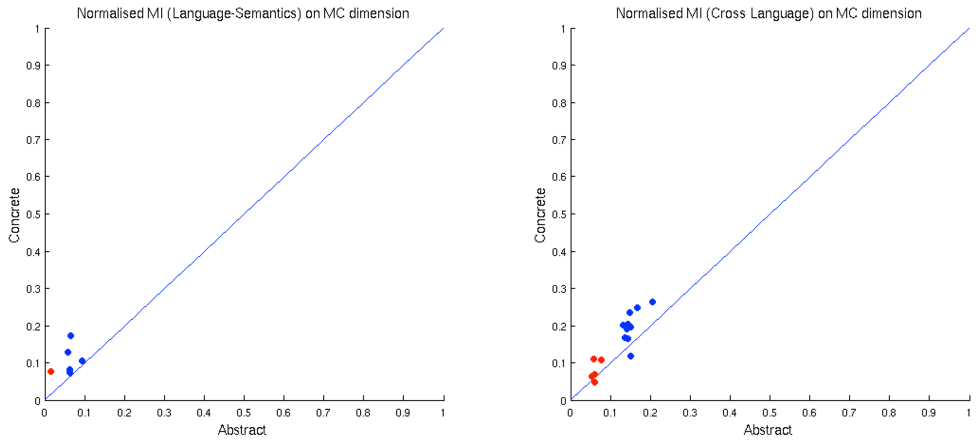

Figure 7: Scatter plots of mutual information values along the main mass/count dimension.

The normalized mutual information (see Methods) between the groups of individual nouns is shown between semantics and syntax (left) and between the syntax of pairs of languages (right). In each plot, the normalized mutual information for concrete nouns is on the y-axis and for abstract nouns on the x-axis. Pairs that include Marathi are indicated in red.

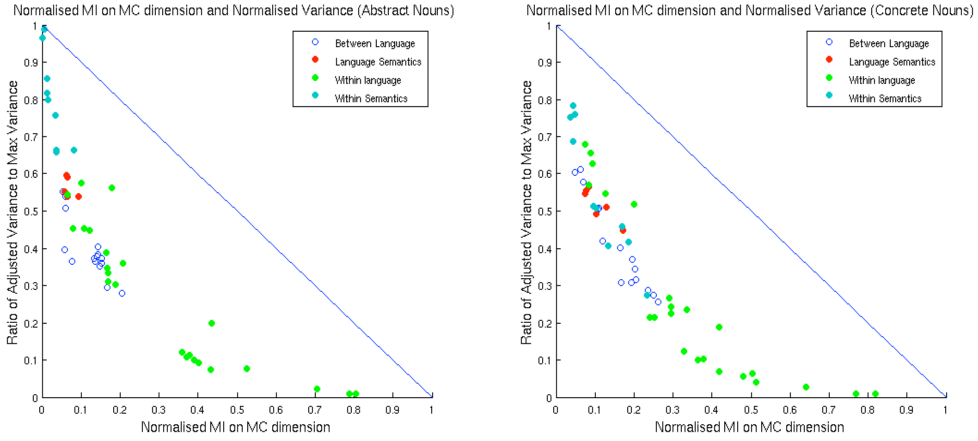

Figure 8: Scatter plots comparing variance with mutual information values.

The normalized mutual information (along the main mass/count dimension) is shown on the x-axis with the corresponding normalized variance value on the y-axis, for abstract nouns (left) on and for concrete nouns (right). Different colors denote data points between the syntax of pairs of languages (empty circles), between the semantics and syntax (red), within language (green) for 10 Armenian, 10 Italian, and three Marathi data points, and within different semantics informants (10 light blue data points).

Figure 8 shows, first of all, that the MI measure and the Variance measure are broadly equivalent. Their relation is (very roughly) Var ~ $(1\text{–MI})^4$. This occurs despite the different nature of the two measures: the mutual information is not sensitive to distance along the mass/count dimension, only to group membership, whereas variance has limited sensitivity to small differences in the exact classification of each noun, as long as its position on the mass/count dimension does not vary too much. Variance turns out to be a more informative measure with our data, which better span its 0–1 range, but mutual information can be easily generalized beyond the main mass/count dimension.

Second, the within language data show mostly more agreement (higher MI and lower Var) than the between language data. Exceptions are due to one Armenian informant (yielding four data points) and one Marathi informant (yielding two more data points) that differ sensibly in their syntactic judgment from the rest. The 'average' data for both Armenian and Marathi, however, due to the majority rule effectively disregards their peculiarities. Thus both measures overall indicate more agreement between informants of the same language than between languages, although this is very far from a clear cut all-or-none difference. Confronted with the requirement to answer yes or no to a set of binary questions, speakers of the same language vary substantially in their responses.

Third, the informants who contributed the semantic classification show the least agreement, particularly for abstract nouns. Even though there were just two questions to answer for abstract nouns, the responses to those two questions are effectively random, with the variance between informants close to its random reference value (in one case exceeding it), and the mutual information close to zero. This suggests that while the semantic properties that should inform the mass/count syntactic usage are already not that salient and self-evident for concrete nouns, they are completely irrelevant for abstract nouns.

3.4. *Mutual Information across the Complete Syntactic Classification*

Mutual information is however higher when all the dimensions are considered (Fig. 11), even in relative terms, i.e. when taking into account that the entropy values are higher for the full classification (Table 8). Entropy values, as discussed in the Methods section, inform us about the logarithm of the equivalent number of significant classes found in the data. Table 8 shows that the entropies of the languages are in the range of 2–4 bits, which indicates the presence of something equivalent to 2^2–2^4 equi-populated classes of nouns (from slightly above 4 for Hindi to just below 16 for English). In a hypothetical case where there were just two significant classes of mass and count the entropy would have been in the range of 1 bit, in fact even less if the count class were, as it turns out to be in most cases, much more populated. This provides a quantitative estimate of the variability that exists in the mass/count classification, which is much higher than may have been intuitively expected.

Language	Entropy
*Armenian	2.29
*Italian	3.02
*Marathi	2.71
English	3.92
Hebrew	3.40
Hindi	2.12
*Semantics	(3.72) 2.94 (C) 2.34 (A)

Table 8: Entropy values for the full classification in the 6 languages, and for semantics.

*The * sign indicates an 'average' over five informants (three Marathi), taken by assigning to each question and each noun the yes/no answer chosen by the majority.*

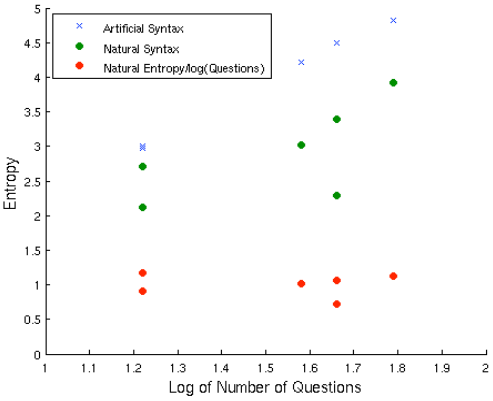

Figure 9: The entropy scales up with the number of questions.

Both when calculated for natural syntax and for the artificial syntactic strings used as controls, entropy values turn out to be roughly proportional to the logarithm of number of questions, hence to yield almost the same value, around 1, when divided by that number.

It is important to note that the entropy and mutual information values obtained with our procedure are influenced by the number of questions used for each language. The scale of the entropy of the 'artificial syntax' depends solely on the number of questions, and we can see from Figure 9 how also the entropy values for natural syntax are strongly correlated with the logarithm of number of questions. Dividing the entropy of natural syntax (Table 8) by the logarithm of the number of question all the entropy values get together at around the 1 bit mark.

The limited agreement that there is, is somewhat stronger for concrete than for abstract nouns except for the 10 within Italian pairs. Figure 10 indicates that this holds within languages, between languages, and much more so when including semantics. As noted above in the case of measures restricted to the main mass/count dimension, the semantic classification of abstract nouns is so arbitrary that agreement among the five informants that filled the questionnaire is extremely low, and the correspondence of their majority response with any natural syntax is also low.

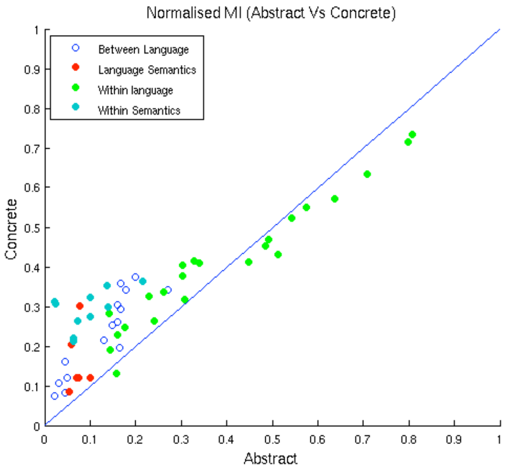

Figure 10: Scatter plots of mutual information values for abstract and concrete nouns.

The normalized mutual information is shown for abstract nouns on the x-axis with the corresponding value for concrete nouns on the y-axis. Different colors denote data points between the syntax of pairs of languages (empty circles), between the semantics and syntax (red), within language (green) for 10 Armenian, 10 Italian, and three Marathi data points, and within different semantics informants (10 light blue data points).

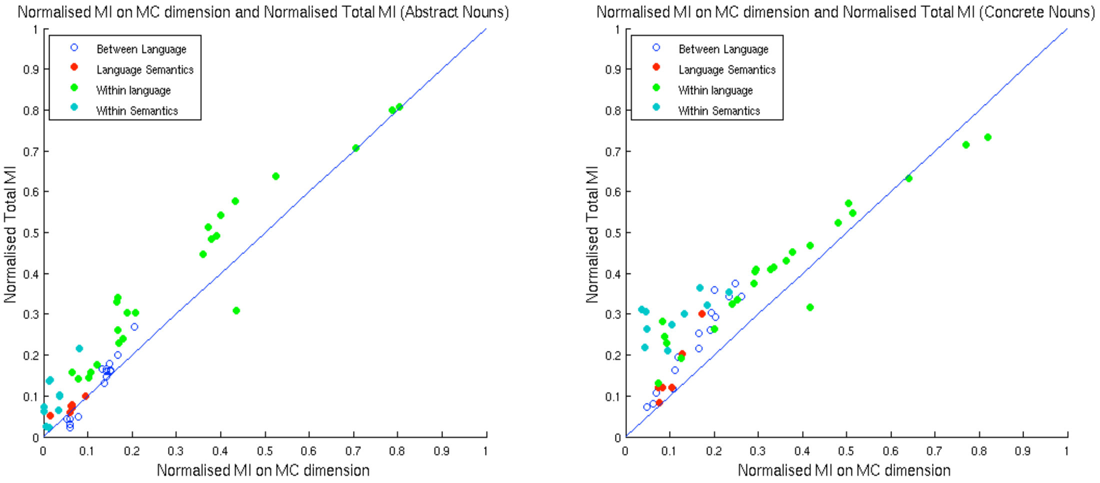

Figure 11: Scatter plots comparing mutual information on the MC dimension with total mutual information values.

The normalized mutual information along MC dimension is shown on the x-axis with the corresponding normalized mutual information including all dimensions on the y-axis, for abstract nouns (left) on and for concrete nouns (right). Different colors denote data points between the syntax of pairs of languages (empty circles), between the semantics and syntax (red), within language (green) for 10 Armenian, 10 Italian, and three Marathi data points, and within different semantics informants (10 light blue data points).

From Figure 11 we see that relative mutual information, when all the dimensions are taken together, is only slightly higher than when just the main mass/count dimension is considered, telling us that most of the variability is present along the main MC dimension. Again, abstract nouns show a larger variability between and within languages, and this difference is particularly strong within semantics. The source of the variability is most likely to be the degrees of freedom left in the syntactic or semantic classification task, applied to the abstract nouns. Even though the nouns and their meanings were disambiguated with a reference sentence, informants were still free to frame the sentences while deciding whether a particular marker can be used with a particular noun. Hence part of the variability may come as a result of the somewhat arbitrary determination of the exact meaning used by different informants when adapting their abstract cognitive categories to the classification of nouns, or of individual differences in the manipulation of context (Raymond *et al.* 2011).

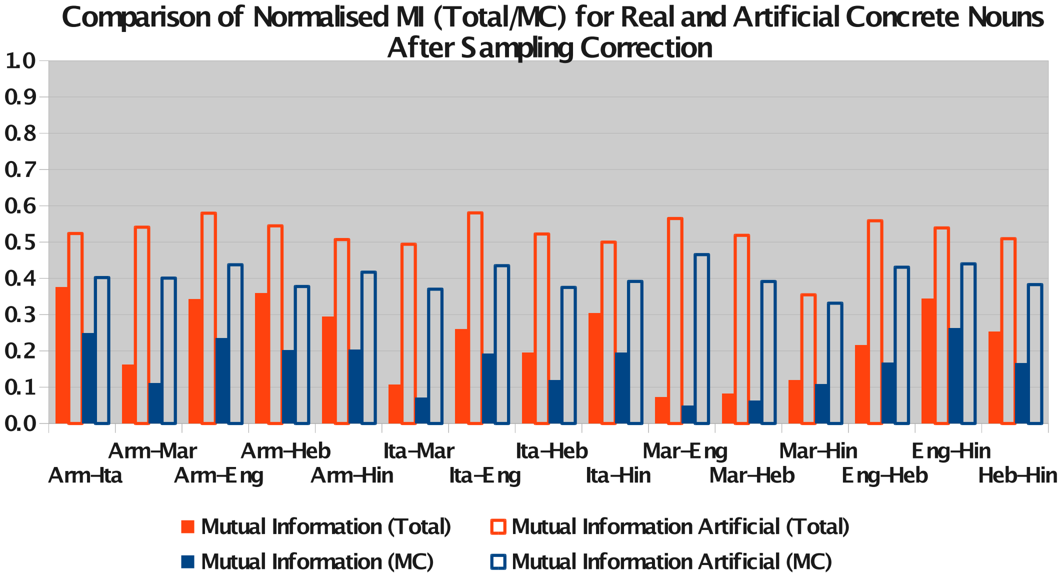

Figure 12: Mutual information between language pairs vs. artificially generated control values. Normalized mutual information between language pairs (red solid) are in the 0.33–0.52 range, except for pairs including Marathi, for which they are around 0.2. These values can be contrasted with the higher values obtained by generating a pseudo usage table, based solely on semantic properties (red empty), as explained in Methods. A similar comparison is shown for the normalized mutual information but only on the MC dimension (blue-solid for real and blue empty for artificial).

Figure 12 tells us that there while relative values are higher than when computed only along the main mass/count dimension (Fig. 7), still there is little agreement across languages even on a finer scale, as the MI values are mostly less than half of the lower of the two entropies. Mutual information is a strict measure, wherein a single bit difference will put a noun in a different equivalence class and lower the mutual information. In contrast, however, artificial syntactic strings produced from the semantic ones, with the stochastic procedure outlined in Methods, share around 50% mutual information, relative to the lower of their entropy values. Artificial syntactic strings also 'suffer' from a sensitivity to single fluctuating bits, hence the contrast between their 50% agreement and the 20–30% (roughly) agreement of the real syntax tells us that real agreement is genuinely low, and it is not all due to using a bizarre measure. The low mutual information of the natural syntax suggests that there is considerable syntactic variability along different dimensions of the syntactic domain, although most of the variability is already in the main mass/count dimension, since when restricted along that dimension agreement is even lower (Fig. 11–12).

3.5. CHILDES Corpus Study

With a method analogous to the Hamming distance measures, we analyze the Brown's section of the CHILDES corpus (only for the adult sentences) on the main mass/count dimension. We simply count the frequency of occurrence of a noun with mass markers out of the total occurrence of the noun in the corpus. There are 1551 nouns in this study out of which 522 nouns (151 are abstract and 371 concrete) are in common with the nouns used for the analyses above. Figure 13 plots the distribution of all the nouns on this main mass/count dimension. In a similar trend to Figure 3, we see the nouns to be distributed all across the spectrum from count to mass, with an overall decreasing trend in frequency going from count to mass (except for the pure mass class). Nouns with pure count usage are very many compared to the rest of the groups. We do find, however, a higher number of pure mass nouns in the corpus, as compared to the English syntactic data obtained from an informant.

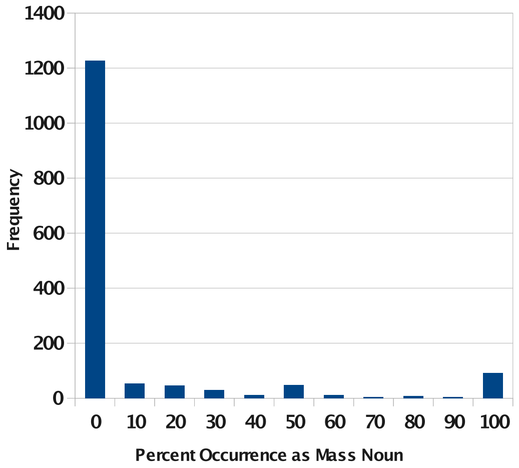

Figure 13: Distribution of nouns from the CHILDES corpus on the main mass/count dimension. Count occurrences of nouns are very frequent as compared to mass occurrences, with nouns lying along the entire spectrum.

A multi-dimensional analysis of the corpus data brings forward four markers as salient, two count ('*a(n)*' and Pluralization) and two mass markers (bareness and '*some* + singular noun'). Nouns mostly lie along the vertices connecting these four markers. Figure 14 shows the most significant dimensions in terms of the co-occurrence frequencies found in the corpus, for example, along the edge connecting the vertices '*a(n)*' and 'pluralization', close to the '*a(n)*' vertex there are nouns that occur almost always with '*a(n)*' but seldom in plural form, in the corpus, while close to the 'pluralization' vertex there are nouns with the opposite occurrence, with the rest of the nouns occurring in between these two extremes. All nouns along this edge are in any case classified as pure count nouns, in the first bin of Figure 13. The density of nouns along the count edge is much higher than along the 'mass edge' (defined by the properties of appearing in bare form, at one vertex, and appearing with 'some' + singular noun at the other vertex). These four markers have the highest variance in their frequency of occurrence across the nouns in the corpus (Table 9).

bare	***a/an***	***every/each***	***many***	**pluralization**	***much***	***some***	***a lot of***
0.0485	0.1556	0.0044	0.0015	0.1177	0.0034	0.0275	0.0010

Table 9: Variance of the markers in the CHILDES corpus.

The variance of the markers we used to classify nouns in the Brown's section of the CHILDES corpus was calculated across its 1551 nouns, and the four markers with highest variance were used, a posteriori, to characterize the three most significant dimensions of mass/count variability, as independently generated by multi-dimensional scaling.

Finally, we contrast mass/count entropy values extracted from the corpus from those measured from the informant responses. To obtain entropy estimates from the CHILDES corpus, which can be used for the comparison, we first binarize the corpus co-occurrence frequency table, such that if a marker was found at least once with a noun, it was assigned the value of 1, and 0 otherwise. With this method, the total entropy of the corpus data was calculated to be 3.75 bits, as compared to the English informant entropy, which is 3.92. Since 522 nouns (151 abstract and 371 concrete) are common to the corpus and informant usage tables, we calculated the entropy on the MC dimension for them, too.

Informant entropy for concrete nouns on MC dimension	2.16
Corpus entropy for concrete nouns on the MC dimension	1.37
Informant entropy for abstract nouns on MC dimension	2.46
Corpus entropy for abstract nouns on the MC dimension	1.35

Table 10: The entropy values for nouns in both the database and the CHILDES corpus.

The entropy of the corpus on the MC dimension is lower than that of informants, perhaps due to the restricted contexts in which sentences can occur in a corpus, as opposed to the freedom of choice to the informants. The normalized mutual information, including all dimensions, between binarized corpus and informant data after sampling correction is 0.051 for concrete nouns and 0.001 for abstract nouns.

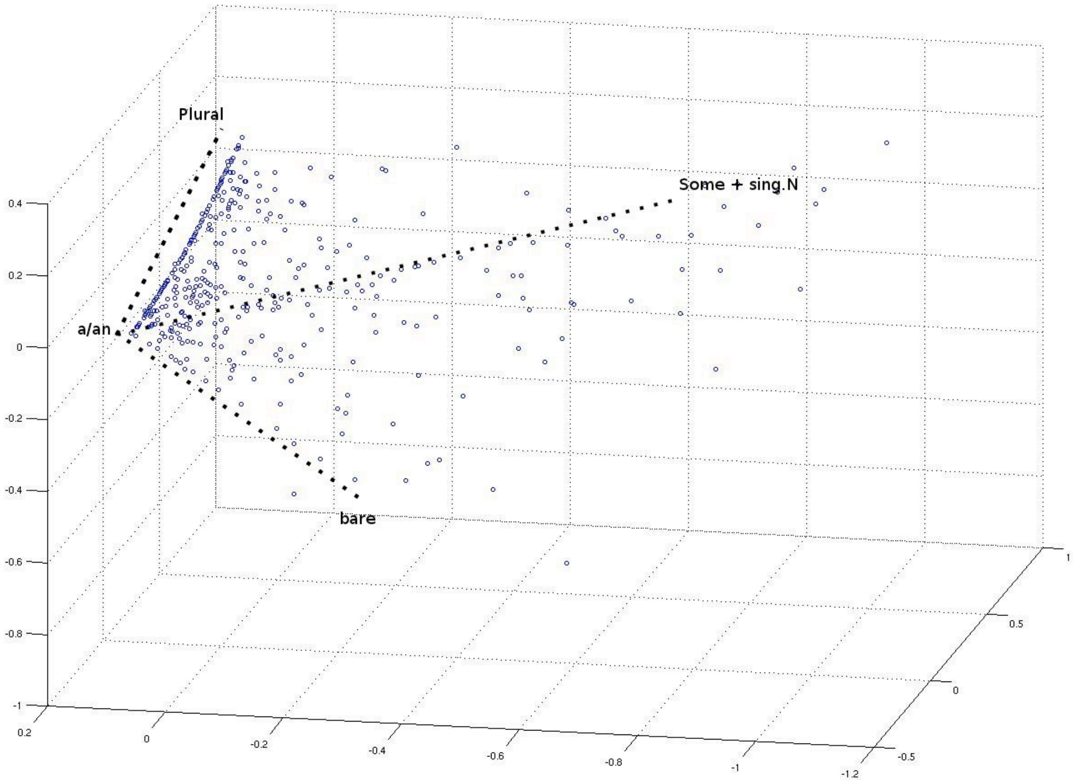

Figure 14: Visualization of the nouns in the Brown's section of CHILDES corpus in three dimensions, from multi-dimensional scaling.

4. Discussion

This is to our knowledge the first wide scale examination of cross-linguistic variation in the expression of the mass count distinction, which attempts to investigate the question of the degree to which the distinction is driven by perceptual-semantic attributes. Previous discussions in terms of data have stayed more or less at the level of the anecdotal. Our major contributions to the discussion are to show that the relation between such universal perceptual-semantic attributes and syntactic usage in specific languages is very weak; as is the relation between languages: There is a core group of count nouns where semantic atomicity corresponds directly with count syntax, but beyond this there is indeed widespread cross-linguistic variation in whether or not a concept is expressed via count syntax. In our sample of 1,434 nouns, in the five languages excluding Marathi, approximately 50% were what we would call 'robustly count', however only 392 were robustly count cross linguistically. We have little to say about core mass nouns, of which there were few or none in our sample.

This might conceivably be because of the way in which we chose our data base, rather than because of the inherently lower number of mass nouns in the languages. We leave it to other studies to identify a significant core group of mass nouns, cross-linguistically.

We have made a number of observations which are relevant to the discussion of the mass/count distinction:

I. Semantic or 'real world' attributes do not lead in a straightforward manner to individual syntactic rules in the mass/count domain, hence we have to probe a potential mapping, for any given natural language, between semantic attributes and a constellation of multiple syntactic rules. The obvious alternation i.e. atomic vs. homogeneous does not predict mass vs. count morphosyntax. This provides solid statistical support for the theoretical discussion in Gillon (1992), Chierchia (1998), Rothstein (2010), and many others.

II. When probing this domain with multiple syntactic usage alternatives, the distribution of 1,434 frequently occurring nouns in six natural languages is typically very far from binary. The largest single class of nouns in five of the six languages was the pure count prototype, i.e. the nouns classed 'count' by all syntactic probes. The rest are distributed in a graded fashion, with fewer and fewer nouns having more usage properties opposite to those of pure count nouns. Out of the 1,434 nouns, on average 873 were 'pure' count in a single language, range [693–1058], when excluding Marathi (where the figure was 255), but only 392 were 'pure count' in all other five ('typical') languages.

III. Outside of the pure count nouns, the correspondence between languages is weak, even when considering a single matching usage marker in each of the five non-exceptional 'typical' languages in the sample. In other words, learning what is a pure count noun, in any of these five languages, gave no significant clues as to the content of the pure count class in any of the other languages, beyond the 392 nouns which were pure count in all languages.

IV. Marathi differs from the other 'typical' languages in having a substantial fraction of nouns close to a pure mass prototype, particularly among abstract nouns, and a distribution closer to bimodal.

V. The semantic attributes that may be at the origin of the syntactic usage properties are distributed similarly, across concrete nouns, to the typical syntactic distribution, with most concrete nouns having 'count-like' attributes, and gradually decreasing proportions showing progressively more mass-like attributes.

VI. Despite the overall similarity between distributions, of semantic attributes and of syntactic usage properties (in all languages tested except Marathi) the correspondence in position along the main mass/count dimensions between semantics and syntax is very weak, even for concrete nouns. Quantitatively, in terms of variance it is midway between fully matching and random, and in terms of mutual information it is close to random. The different range reflects the non-linearity of the MI measure, but both measures point at the weakness of the observed correlation.

VII. Similarly, the correspondence between languages is weak, whatever measure is used.

VIII. Taking into account the detailed attributes and syntactic rules, rather than only the main mass/count dimension, the correspondence remains weak.

IX. There is considerable variability also among informants of the same language; part of which may be due to the testing paradigm.

X. A similar distribution along the main mass/count dimension can be gauged from 1,551 nouns extracted from the adult section of the English-language CHILDES database, after a different analysis, namely in terms of graded rather than binary syntactic usage frequencies. The three main dimensions of syntactic variability of nouns in the CHILDES database describe an asymmetrically loaded pyramid: most nouns are countable, and simply vary in their plurality at each instance; many fewer nouns span the other two dimensions, characterized by an increasing frequency of use in bare form, and of use with *some*+singular form, both mass-like attributes.

The results that we have reported have been purely statistical, that is to say, we have reported numbers with no discussion of any of the 1,434 items that make up of data base (where an item is a token from a particular language plus its particular feature values). An analysis of patterns within the data base is obviously the next stage in a linguistic analysis. This analysis will involve investigating whether there are recognizable patterns within the variation which are open to interpretation, whether there are lexical classes of nouns which function as classes cross linguistically, and if so how to characterize them. For example, *advice, information,* and *evidence* are strongly count in Hebrew and Italian, and mass in English. Do they behave as a class in other languages too? However, the results that we have so far already have theoretical implications relevant for continued research into the semantics and grammatical aspects of the mass/count distinction, and we conclude by specifying three of them.

First, we have provided solid empirical evidence that count syntax is not a direct reflection of atomicity in the denotation. Our initial aim to quantify the correlation (which we had presumed strong) between non-homogeneous nouns and count syntax could not reach beyond a core group of 392 nouns which pattern as pure count in all languages checked, excluding Marathi. This indicates a weak correspondence between perceptual/semantic and grammatical or morphosyntactic properties. Note that the 392 cross-linguistically count nouns included approximately 27% abstract nouns (284 concrete and 108 abstract), thus it is not even possible to argue that count syntax correlates directly with concrete atomic entities. Beyond this group of 392 nouns, the low level of mutual information between any two languages indicates language-specific grammaticalization of the distinction. This means that it is no longer possible to assume a general correlation between atomicity and count, and homogeneity and non-count. A preliminary examination of the 284 items in the pure count group which are concrete rather than abstract indicates a high number of [+animate] nouns, in particular individuals of a certain profession (*scientist, nurse, preacher, slave, spectator*), nouns denoting buildings with a particular function (*library, bank, apothecary*) and nouns denoting artifacts that individuals stand in an one-to-one relation with (*wallet, watch, handkerchief*). All these predicates are atomic in an absolute sense, since they come in inherently individuable units, but they also frequently occur in contexts in which a particular instantiation of the predicate is perceptually salient. Thus it is plausible to posit that atomicity may be a necessary condition of a non-abstract noun being robustly count cross-linguistically. However, beyond this there are no straightforward generalizations.

The fact that these 284 nouns constitute between a third and a half of the robustly count nouns, in any particular language, indicate that beyond this weak generalization the grammaticalization of the correlation between atomicity and count differs from language to language. Furthermore, there are 108 abstract nouns in the pure count group, where the criteria for atomicity are by definition not well defined (since 'atomicity' is usually taken to express non-overlapping properties of matter). One can at this stage hypothesize potential criteria, for example, individuation via events: *nightmare, appointment*, and *crash* are all robustly count and non-concrete and atomic instantiations can be potentially be individuated via temporally located events. But this requires a notion of event individuation, itself problematic (see, e.g., Parsons 1990) and even then, leaves open the question of which event-types are 'inherently atomic' and which not. The conclusion for the linguist is that exploration of the basis of the mass/count distinction must be language particular, and will involve semantic features far beyond the homogeneous/atomic distinction.

We can draw a second theoretical implication from our results. We have seen that, for each language (again excluding Marathi), the approximately 50% of nouns which are not purely or robustly count in almost all cases cannot be characterized as 'pure mass'. These nouns are located at varying distances from the pure count class, depending on how many non-count features they have. This could be taken as support for the view that mass/count syntax is imposed on a neutral root, that it is appropriate to talk of mass or count 'usage', and that essentially, noun roots are flexible and can appear in either context. This is the view taken in Borer (2005), who claims that 'being a count noun' is an exoskeletal phenomenon, the result of count syntax being imposed on a neutral syntactic root. Our data, however, show that approximately 50% of the nouns in each language do show a consistent count pattern, and furthermore, as stressed above, beyond the first 392 nouns, the choice of which nouns are used consistently as counts is specified within a language (subject to some idiolectal variation) and not across languages. This suggests that count syntax is a lexical specification, and that beyond a core group, it is specified independently for each language.

A third point is that our data reveals cross-linguistically (again excluding Marathi) a large group of pure count nouns, and no comparable group of mass nouns. This may be taken to support the widely accepted view (e.g., Chierchia 1998, Borer 2005, Rothstein 2010) that mass syntax is the default case, and that count nouns are derived from mass nouns via some form of operation, which results in their sharing common properties. The degree to which Marathi differs from the other languages studied also forces us to realize that languages with a mass/count contrast may differ quite radically in how they implement it, and that the division of languages into those which have a count/mass distinction and those which do not tells us little about typological variation.

The overall conclusion is that the questions that linguists have been asking should be reformulated: Instead of looking for a general semantic characterization of the mass/count distinction which will explain the grammatical distribution cross-linguistically, linguists should be looking for language-specific patterns or generalizations, indicating that in a particular language, certain lexical classes are or are not grammaticalized as count. (For example, a cursory examination of the data indicates that Marathi is very restricted in allowing count syntax for abstract nouns.) If there are cross-linguistic generalizations, we might expect for them to have an implicational structure in the sense of Greenberg (1963), i.e. we could look for patterns of the form: If lexical class C_1 is pure count, then lexical class C_2 is also pure count. But it is an open question whether we would find them at any significant level. We should avoid classifying nouns as 'count', 'mass' or 'flexible'. In particular, our data show that non-robustly count nouns are flexible in different ways and to different degrees. What these ways and degrees are is still to be investigated.

If there is a general characterization of the mass/count distinction, then it probably is in terms of how the denotations of count (or mass) predicates are represented in the language, rather than in terms of any real-world feature. For example, Rothstein (2010) suggests that count nouns denote entities which are indexed for the context in which they count as atomic. This leaves place for particular languages to rank features which contribute to contextual salience, or to give them different weights, which might then influence patterns in classifying nouns as count. Features which weigh heavily in their contribution to count syntax in all languages would result in the set of pure count nouns cross-linguistically. In any case, the set of robust count nouns and the lack of a set of robust mass nouns indicate that we are more likely to find a general semantic characterization of count nouns than of mass nouns.

At a deeper epistemological level, not only is mass count syntax largely left undetermined by semantic attributes, it is also mistaken to regard it as a binary or quasi-binary structure. The distribution of syntactic usage properties is very far from bimodal in five out of the six languages tested, in fact it has nothing to do with bimodality. One is led to think of this grammaticalization as a graded self-organization process, operating within languages and to some extent within individual speakers, and driven only to a limited extent by universal attributes, and plausibly governed or at least constrained by language specific principles. However, at this stage we cannot tell to what degree the grammaticalization is governed, beyond the universal semantic or perceptual principles that we have attempted to quantify, by language-specific principles of different nature, such as cultural factors, historical accidents, individual language acquisition history, even context dependence within individual speakers. What is already clear, however, is that a domain of grammar, that to the non-specialist may seem rather straightforward, in fact opens new vistas on the character of what are improperly called language 'rules'.

Appendix

The following tables are the equivalent of Table 1, for languages other than English.

No.	Syntactic Questions
1.	Can the noun be used with ‘a(n)’? (անորոշ գոյական +’մի’ հոդ)
2.	Number distinction: Can the noun be used with plural form? (հոգնակի թիվ)
3.	Can it be used in combination with numerals? (համադրում թվականների հետ)
4.	In combination with classifiers or measure phrases that manipulate number? (համադրում դասակարգիչների հետ)
5.	Can the noun be used with ‘every’/’each’? (Ամեն/յուրաքանչյուր)
6.	Can it be used with ‘(a) little’? (մի քիչ)
7.	Can it be used with ‘(a) few’? (մի քանի)
8.	In combination of ‘many’ + plural form of noun? (շատ + գոյականի հոգնակի թիվ)
9.	In combination of ‘much’ + singular form of noun? (շատ + գոյականի եզակի թիվ)

Table A1: List of questions used in Armenian to compile the usage table.

No.	Syntactic Questions
1.	Can the noun appear in the singular?
2.	Can the basic form appear with *af* (as in *af yeled lo ‘ana*, ‘not a single boy answered’)?
3.	Is there a plural form?
4.	Can the plural form of the noun appear with a number?
5.	Can the singular form of the noun appear after *kol* ‘every’?
6.	Can the singular form appear with *kzat, me’at, harbe* (‘a little, a little, a lot’)?
7.	Can the noun appear with *tipa* (literally ‘a drop’)?
8.	Can the noun appear with a classifier?
9.	Is it possible to say 10 + the singular form of the noun?

Table A2: List of questions used in Hebrew to compile the usage table.

No.	Syntactic Questions
1.	Can it be used with ‘many’/‘few’?
2.	Can it be pluralized?
3.	Can it be used with ‘every’?
4.	Can it be used with numerals?
5.	Can it be used ‘with a lot of’?

Table A3: List of questions used in Hindi to compile the usage table.

No.	Syntactic Questions
1.	Can the noun be in singular form with the indefinite article (*un/o/a*)?
2.	Can it appear (suitably pluralized) with a numeral (*due, tre*)?
3.	Can the noun appear with at least one singular indeterminate quantifier (*molto/molta/un po' di*)? Note: *non molto* should not be considered.
4.	Can the singular form be preceded by indefinite quantifier *qualche*?
5.	Can the singular form be preceded by exact quantifiers (*chili di, litri di*)?
6.	Can the singular form be preceded by *non molto* (‘not much’)?
7.	Can it have a plural form with a definite article (*i, gli, le*)?
8.	Can the plural form be preceded by exact quantifiers (*chili di, litri di*)?

Table A4: List of questions used in Italian.

No.	Syntactic Questions
1.	Can it appear with a numeral?
2.	Can it be used in combination with an exact quantifier (kilo, liter)?
3.	Can it be used with the article *ek* (‘a’)?
4.	Can it be pluralized?
5.	Does the morphology change when pluralized?

Table A5: List of questions used in Marathi.

Note: The questions were posed to the informants in their respective languages, not in the English translation.

References

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

Bale, Alan C. & David Barner. 2011. Mass-count distinction. *Oxford Bibliographies Online*, http://ladlab.ucsd.edu/pdfs/BB@Oxford.pdf.

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

Borer, Hagit. 2005. *In Name Only*, vol. I: *Structuring Sense*. Oxford: Oxford University Press.

Chierchia, Gennaro. 1998. Plurality of mass nouns and the notion of 'semantic parameter'. In Susan Rothstein (ed.), *Events and Grammar*, 53–103. Dordrecht: Kluwer.

Chierchia, Gennaro. 2010. Mass nouns, vagueness and semantic variation. *Synthese* 174, 99–149.

Gillon, Brendan S. 1992. Toward a common semantics for English count and mass nouns. *Linguistics and Philosophy* 15, 597–640.

Greenberg, Joseph. 1963. *Universals of Language*. Cambridge, MA: MIT Press.

Grimm, Scott & Beth Levin. 2011. Furniture and other functional aggregates: More and less countable than mass nouns. Paper presented at *Sinn und Bedeutung 16*, University of Utrecht. [6–8 September 2011]

Hacohen, Aviya. 2010. On the (changing?) status of the mass/count distinction in Hebrew: Evidence from acquisition. *Proceedings of the 25th Annual Meeting of the Israel Associations for Theoretical Linguistics*, doi: http://linguistics.huji.ac.il/IATL/25/Hacohen.pdf.

Jespersen, Otto. 1924. *The Philosophy of Grammar*. London: Allen and Unwin.

Koptjevskaja-Tamm, Maria. 2004. Mass and collection. In Geert Booji, Christian Lehmann and Joachim Mugdan (eds.), *Morphology: A Handbook on Inflection and Word Formation*, vol. 2, 1016–1031. Berlin: Walter de Gruyter.

Landman, Fred. 2010. Count nouns, mass nouns, neat nouns, mess nouns. In Barbara H. Partee, Michael Glanzberg & Jurgis Skilters (eds.), *The Baltic International Yearbook of Cognition, Logic and Communication*, vol. 6, 1–67. Manhattan, KS: New Prairie Press.

Landmann, Fred & Susan Rothstein. 2012. The felicity of aspectual *for*-phrases – Part 1: Homogeneity. *Language and Linguistics Compass* 6, 85–96.

Link, Godehard. 1983. The logical analysis of plurals and mass terms: A lattice-theoretical approach. In Reiner B. Bäuerle, Christoph Schwarze & Arnim von Stechow (eds.) *Meaning, Use and Interpretation*, 303–323. Berlin: Mouton de Gruyter. [Reprinted in Paul Portner & Barbara Partee (eds.). 2002. *Formal Semantics: The Essential Readings*, 127–146. Oxford: Blackwell.]

Markman, Ellen M. 1985. Why superordinate category terms can be mass nouns. *Cognition* 19, 31–53.

MacWhinney, Brian. 1995. *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, NJ: Erlbaum.

Nicolas, David A. 2010. Towards a semantics for mass expression derived from gradable expressions. *Recherches Linguistiques de Vincennes* 39, 163–198.

Panzeri, Stefano & Alessandro Treves. 1996. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems* 7, 87–107.

Parsons, Terence. 1990. *Events in the Semantics of English*. Cambridge, MA: MIT Press.

Pelletier, Francis J. 2010. Descriptive metaphysics, natural language metaphysics, Sapir–Whorf, and all that stuff: Evidence from the mass-count distinction. In Barbara H. Partee, Michael Glanzberg & Jurgis Skilters (eds.), *The Baltic International Yearbook of Cognition, Logic and Communication*, vol. 6, 1–46. Manhattan, KS: New Prairie Press.

Pires de Oliveira, Roberta & Susan Rothstein. 2011. Bare singular noun phrases are mass in Brazilian Portuguese. *Lingua* 121, 2153–2175.

Prasada, Sandeep, Krag Ferenz & Todd Haskell. 2002. Conceiving of entities as objects and stuff. *Cognition* 83, 141–165.

Raymond, William D., Alice F. Healy & Samantha J. McDonnel. 2011. Pairing words with syntactic frames: Syntax, semantics, and count-mass usage. *Journal of Psycholinguistic Research* 40, 327–349.

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

Soja, Nancy N., Susan E. Carey & Elizabeth S. Spelke. 1991. Ontological categories guide young children's inductions of word meanings: Object terms and substance terms. *Cognition* 38, 179–211.

Taler, Vanessa, Gonia Jarema & Daniel Saumier. 2005. Semantic and Syntactic aspects of the mass/count distinction: A case study of semantic dementia. *Brain and Cognition* 57, 222–225.

Wierzbicka, Anna. 1988. *Oats and Wheat: The Semantics of Grammar*. Amsterdam: John Benjamins.

Ritwik Kulkarni
SISSA
Neuroscience
via Bonomea 265
Trieste 34136
Italy
rkulkarn@sissa.it

Susan Rothstein
Bar Ilan University
Gonda Brain Research Center
Ramat Gan 52900
Israel
susan.rothstein1@gmail.com

Alessandro Treves
SISSA
Neuroscience
via Bonomea 265
Trieste 34136
Italy
ale@sissa.it