ISSN 2348-3004
Volume 6 Number2
Journal of Advances in Linguistics

A Neural Network Perspective on the Syntactic-Semantic Association between Mass and Count Nouns

ABSTRACT

Analysing aspects of how our brain processes language may provide, even before the language faculty is really understood, useful insights into higher order cognitive functions. We take a small exploratory step in this direction with an attempt to test the ability of a standard, biologically plausible self-organising neural network to learn the association between syntax and semantics around the mass-count distinction. The mass-count distinction relates to the countability or un-countability of nouns, both in terms of their syntactic usage and of their semantic perception. A previous statistical study has shown that the mass-count distinction is not bimodal and exhibits complex fuzzy relations between syntax and semantics. A neural network that expresses competition amongst output neurons with lateral inhibition is shown to identify the basic classes of mass and count in the syntactic markers and to produce a graded distribution of the nouns along the mass-count spectrum. The network however fails to successfully map the semantic classes of the nouns to their syntactic usage, thus corroborating the hypothesis that the syntactic usage of nouns in the mass-count domain is not simply predicted by the semantics of the noun.

Keywords

Mass-Count distinction, Syntax-semantics interaction, Self-organisation, Neural networks

Academic Discipline And Sub-Disciplines

Cognitive Linguistics, Computational Neuroscience

SUBJECT CLASSIFICATION

Linguistics, Neuroscience

TYPE (METHOD/APPROACH)

Information Theory, Neural network simulations

1. INTRODUCTION

The question of how the brain acquires language can be posed in terms of its ability to discover, from exposure to a corpus, the syntactic structure of a specific natural language and its relation with universal semantics. This has been a subject of study and of intense debate for the past few decades [1]. Natural language acquisition appears to presuppose certain cognitive abilities like rule recognition, generalisation and compositionality. These high-level abstract capabilities should be realized in the language domain and in specific sub-domains by general-purpose neural processing machinery, since there is no evidence for dedicated circuitry of a distinct type for each sub-domain nor, for that matter, for languages a whole. How can rule recognition and generalization be implemented in standard, vanilla neural networks? To explore this issue, we focus our attention on a sub-domain of the syntax/semantics interface, the mass-count distinction. Following up on our previous study, where we investigated statistical aspects of the mass-count distinction in 6 languages, with relation to its cross-linguistic syntactic and semantic properties, we now aim to study the learnability of those syntactic properties by a basic neural network model, with the distant goal of eventually understanding how such processes might be implemented in the brain.

The intuitively plausible assumption is that mass nouns denote substance or ‘stuff' and do not denote individuated objects, whereas count nouns denote atomic entities that can be easily counted [2][3][4][5][6]. This semantic difference seems to be reflected in the syntactic usage of the nouns in many natural languages, since mass nouns and count nouns are associated with a different array of syntactic properties, For example, in English, mass nouns are associated with quantifiers like ‘some', ‘much' and require a measure classifier (kilos, boxes) when used with numerals, on the other hand count nouns are associated with determiners like ‘a/an', quantifiers like ‘many/few' and can be used with numerals without a measure classifier. The traditional approach to the semantics of the mass-count distinction is that it can be expressed through properties of atomicity, cumulativity and homogeneity [7] . Count nouns are said to be atomic. A noun is atomic when its denotation includes distinguishable smallest elements which cannot be further divided into objects which are also in the noun denotation. So *chair* is count since it includes minimal chair-objects in its denotation which cannot be subdivided into smaller chairs. Mass nouns are said to be cumulative and homogeneous. A noun is cumulative if the sum of two separate entitiesin the noun denotation is still in the denotation of the noun. For example if A is water and B is water then A and B together are water. A noun is homogeneous if an entity its denotation can be subdivided into parts which are also in its denotation. For example, any part of something which is water is water. So mass nouns are non-atomic and exhibit properties of being homogeneous and cumulative, whereas count nouns have opposite properties. However, as many linguists have pointed out, a simplistic mapping between homogeneity and mass syntax and/or atomicity and count syntax on the other would imply that the expressions in different languages denoting the same real world objects would be consistently count or mass cross-linguistically. This is not the case. As we showed in [8], words with a similar interpretation may be associated with very different arrays of syntactic properties cross-linguistically. A noun which is associated with a count array in one language may not be associated with a count array in a different language. Furthermore, over a sample of 6 different languages we saw that there is no binary divide into mass/count nouns, but rather a continuum with a group

of nouns which are count with respect to all properties, and then a range of nouns which are more of less count depending on how many count properties they display. This places the mass-count distinction at an interesting interface between the semantic properties of nouns and the syntax, since it raises the question of (i) what semantic properties are associated with count and mass syntax respectively, (ii) why there is variation in the noun categorization as mass or count cross-linguistically and (iii) how the knowledge of what is mass and count in a particular language is acquired.

1.1 Statistical analysis of cross-linguistic distribution of mass and count nouns

1.1.1 Data collection and distance distribution from pure count nouns

In our previous study 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. A set of yes/no questions was prepared, in each language, to probe the usage of the nouns in the mass/count domain. The questions probed whether a noun from the list could be associated with a particular morphological or syntactic marker relevant in distinguishing mass/count properties. A similar set of questions probed the semantic usage of the nouns using questions regarding the semantics properties of the nouns relevant for the mass-count distinction. Thus each nouns was associated with a binary string of 1 (Yes) and 0 (No), indicating how that particular noun is used in the mass-count space by an informant. Since the data thus obtained is high dimensional in principle, as a first approximation, we consider the hypothesis that most of the information is contained on a single dimension of 'mass' and 'count'. We collapse the high dimensional data onto a single dimension (named as the MC dimension) 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. 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.

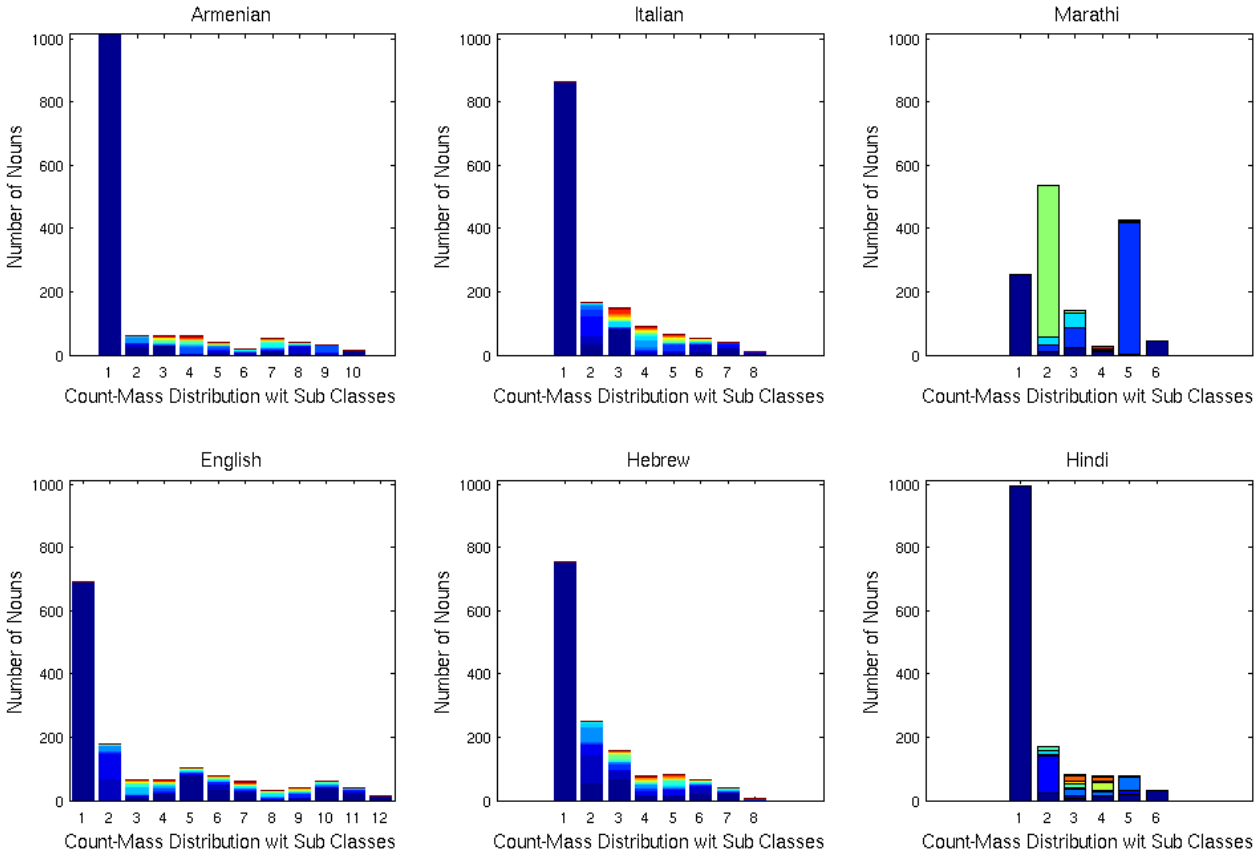

Figure 1: [8] Distribution of nouns along the main mass/count dimension. Each histogram reports the frequency of nouns in the database, for a particular language, at increasing distances from pure count usage (1) and towards pure mass usage (N+1), where N is the number of syntactic question for the language. Shades in the bars indicate the proportion of nouns in each of the syntactic classes occurring at the same Hamming distance from the pure count.

1.1.2 Entropy and Mutual Information measures

A more detailed comparison between the languages, which preserves the multidimensional information, can be obtained by measuring the mutual information between languages and the entropy of a language. Entropy quantifies the amount of 'information' contained in a system (here the amount of information contained in a language in terms of how nouns are

clustered according their usage defined by the binary strings), whereas mutual information quantifies how much of this information is shared between a pair of languages, indicating the extent up to which clustering is similar between the pair of languages. Higher entropy means that a language has a high number of significant clusters thus pointing towards a rich classification of nounsin that language, whereas high mutual information would imply that two languages agree to a high degree on how nouns should be classified in the mass-count domain.

Table 1: [8] Language–entropy relations. Entropy values in the six languages and in 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.

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 1 shows that the entropies of the languages are in the range of 2–4 bits, which indicates the presence of the entropic equivalent of 2^2–2^4 equi-populated classes of nouns (from slightly above 4 for Hindi to just below 16 for English). They provide a quantitative estimate of the diversity in the mass-count classification beyond a dichotomous categorisation, which would have resulted in entropy values around or below 1 bit.

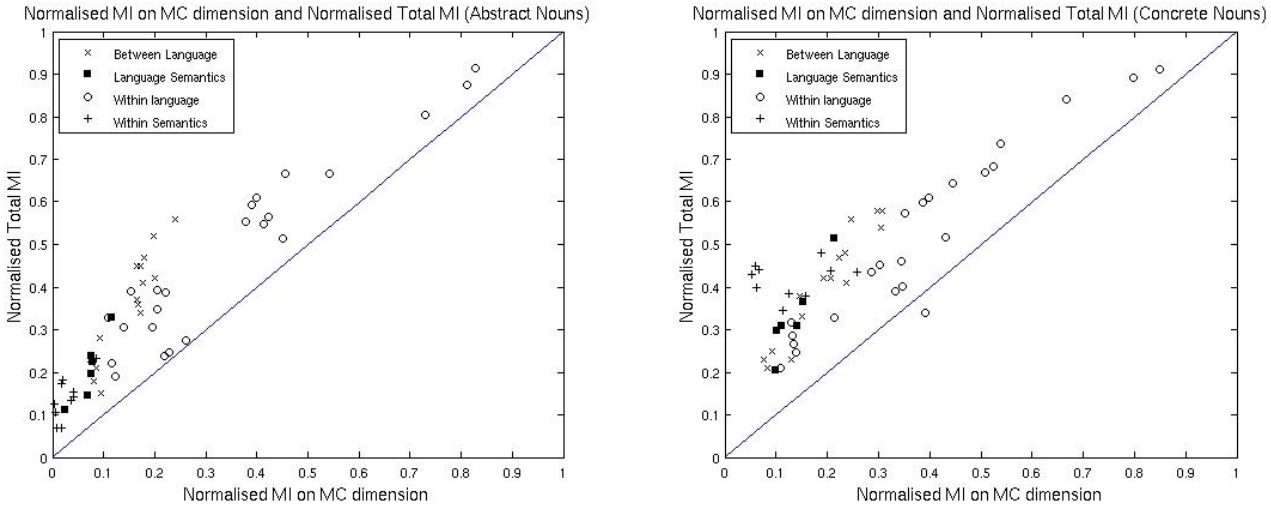

Figure 2: [8] Scatter plots comparing Mutual Information along the main MC dimension with total Mutual Information values.

In general, while the graded distribution is similar across languages, syntactic classes do not map onto each other, nor do they reflect, beyond weak correlations, semantic attributes of the concepts as seen from the low values of Mutual Information on both the MC dimension as well as the total mutual information (Figure 2). These findings are in line with the hypothesis that much of the mass/count syntax emerges from language-specific processes.

1.2 Network modelling

Our goal in the current study is to assess the learnability of syntactic and semantic features of the mass-count distinction using simple neural networks. Artificial neural networks have a long history as a method for neurally plausible cognitive modelling [9][10], and can be endowedwith properties including feature extraction, memory storage, pattern recognition, generalisation, fault tolerance. Understanding how humans might acquire the capacity for handling syntax in a specific sub-domain might start from encoding syntactic/semantic knowledge into a neural network, which self-organizes with a prescribed learning algorithm to recode that information in a neurally plausible format. That way one may draw parallels about governing principles in the brain that bring about the acquisition of syntax. Taking cues from biological neurons, most artificial neural networks employ 'Hebbian' plasticity rules, wherein the synaptic connection between two units is strengthened if they are activated nearly simultaneously, thus leading to associative learning of the conjunction or sequence of activations. Here we consider a competitive network, a simple self-organising network which through 'unsupervised' learning may produce a useful form of recoding. A competitive network, under the right conditions, is able to discover patterns and clusters in a stimulus space and to train itself to correctly identify and group inputs that share a close resemblance to each other. A competitive network is particularly interesting in our case since much of linguistic information during language acquisition is rather 'discovered' than explicitly taught. Moreover, mass and count nouns have been shown to exhibit differential evoked potential responses, both with a syntactic and with a semantic stimulus [11] . We aim to study the performance of a simple competitive network in view of understanding how well can syntactic and semantic features of the nouns in our mass-count database be accommodated within a single network, thus exploring if the network can indeed achieve some rule-recognition that will allow it to successfully categorise nouns in the syntactic mass-count space.

2. METHODS:

2.1 A standard network

Our network consists of a single input and a single output layer. At the input layer each unit represents a syntactic feature ('numeral', 'a/an' etc) in case of the syntactic network or a semantic feature ('fixed shape', 'fluidity' etc) for the semantic network. The input layer is binary, and for each noun given as input a given unit can be active (activation value 1) to indicate that the feature can be attributed to the noun, or inactive (value 0) to indicate that it cannot. Thus a single learning event for the network includes the application of a binary input string containing the syntactic or semantic information pertaining to a single noun, activity propagation to the output units, and modification of the synaptic weights according to the prescribed learning rule. In a variant to be considered later, instead of self-organizing an output representation of nouns, we explore the self-organization of syntactic features ('markers'); in that variant, rather than an input noun with the features as components, we apply as input a single feature/marker, with the nouns as components, i.e. there are a few very long input string instead of many short strings. On the output side, the number of units is variable, determined by the simulation requirements. Unlike the input units, outputs units are graded, taking continuous values in the range of 0 to 1. A competition amongst the output units based upon their activation levels decides the final output level of each unit.

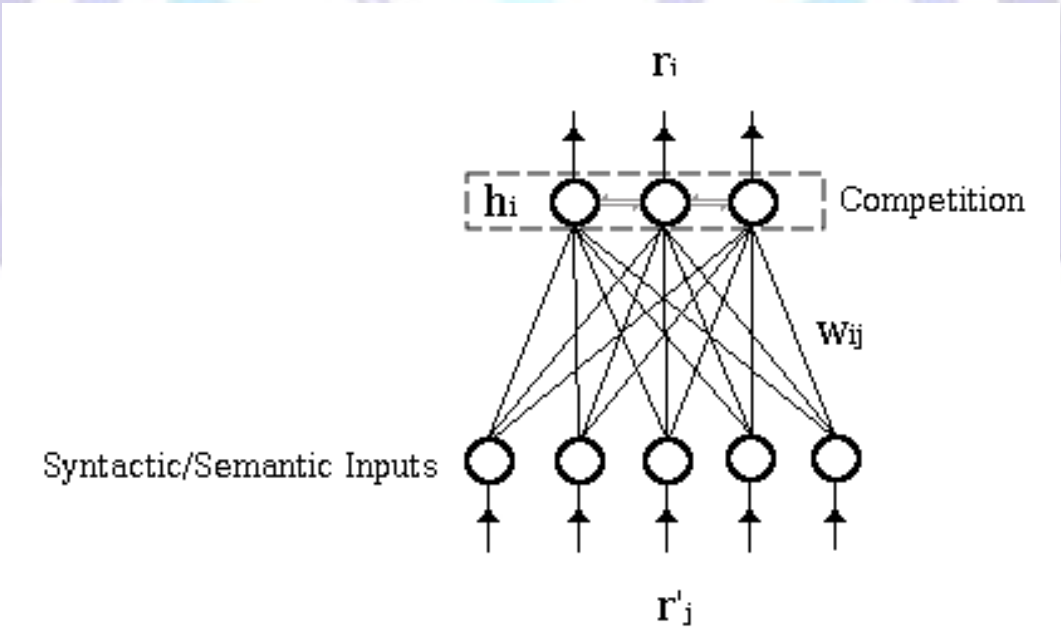

Figure 3: Schematic diagram of the artificial neural network, showing an input layer where units are binary strings containing syntactic/semantic information of nouns and an output layer where units compete with each other to produce graded firing rates based on connection weights and on competition.

We use a fully connected network, where each input unit *j* is connected to each output unit *i* with a synapse whose connection strength is given by w_{ij}. The training sequence is executed as follows:

i. An input is presented to the network and the activation h_i of each output unit *i* is calculated as

$$h_i = \sum_{j=1}^{N} r_j' w_{ij}$$

where N is the number of input units and r' is the input vector. The w_{ij}'s are initially set at random values, which randomly causes certain output units to have a higher activation and lower activation in others.

ii. The final output firing of each unit *r*, is decided after setting up the competition between output units as

$$r_i = \frac{e^{\left(\frac{h_i}{T}\right)}}{\sum_i e^{\left(\frac{h_i}{T}\right)}}$$

Here T governs the strength of the competition, lowering T makes the competition stronger and as T approaches 0 it becomes 'winner take all', a case where only the unit which wins has a maximum firing rate while all other units are suppressed to be inactive; while the competition becomes softer as T is raised higher, allowing more graded output firing rates. Firing rates are automatically normalised in the range of 0 to 1 by this form of the output function (thus also allowing a probabilistic interpretation of the firing rates of the units).

iii. In the next step we adjust the weights w_{ij} according to the Hebbian rule, taking into account the input and output firing rates of the units obtained in the previous step. The learning rule here is slightly modified from the standard Hebbian rule to incorporate normalisation of the weights during learning in a biologically plausible way. Normalising weights is important since it prevents a small fraction of connections becoming too strong and resulting in the same units winning the competition each time. The weights are adjusted at each presentation of an input as

$$\delta w_{ij} = k r_i \left(r_j' - w_{ij}\right)$$

k in the above equation controls the learning rate, i.e. the size of increments in the weights as new input-output pairs are presented to the network. The change in weight is proportional to the input and output firing rates, however the second term restricts a monotonous increase in weights by causing a decay proportional to the activity of the output unit and to its existing synaptic strength.

One training iteration includes presenting each noun in the list once and the above process is repeated for the desired number of iterations.

2.2 Information measures

As mentioned previously we use mutual information as a measure to analyse the correspondence between two representations, encoded either in a syntactic network trained on input information about marker usage for the nouns in a particular language, or in a semantic network trained on information about the semantic properties of the nouns [8] . Here we focus on systems that have undergone a slow process of self-organisation to categorise their inputs.

We first begin by calculating the entropy of the output of the network and the information it contains about the clustering of nouns in the input. Nouns are clustered together if they have exactly the same output firing rates. So in effect the output, labelled as O, contains *n* clusters 1...*i*...*n*, where each cluster contains nouns that are classified as identical by the network. Entropy is then defined as

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

Where *p(i)* is the 'probability' obtained as the relative frequency of the nouns in the cluster *i.*

To calculate mutual information between two representations X and Y we first obtain equivalence classes, i.e. groups where a particular set of nouns has been clustered into the same class in both X and Y. For example, if nouns like 'man' and 'dog' fall in the same cluster in X and are also found in the same cluster in Y, they belong to the same equivalence class. The joint entropy of X and Y, *H(X,Y)* is then used to calculate the mutual information as

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

Equivalently, one may consider a joint probability distribution table of nouns in X and Y from the equivalence classes, and expressmutual information as

$$I(X;Y) = \sum_{i,j} p(i,j) log_2 \left(\frac{p(i,j)}{p(i)p(j)} \right)$$

Mutual information co-varies with the relevant individual entropies, so to facilitate comparisons we use normalised mutual information. When X and Y show completely unrelated clusters, *p(i,j) = p(i)p(j)* and mutual information is 0, giving it a strict lower bound. It has a natural upper bound in the sense that the mutual information between X and Y can never be greater than the lower of the two entropies $H(X) \wedge H(Y)$. Thus for comparison purposes we define normalised mutual information as $I(X;Y)/min\left(H(X),H(Y)\right)$, which lies in the range of 0 to 1.

3. RESULTS

3.1 Classification of Markers

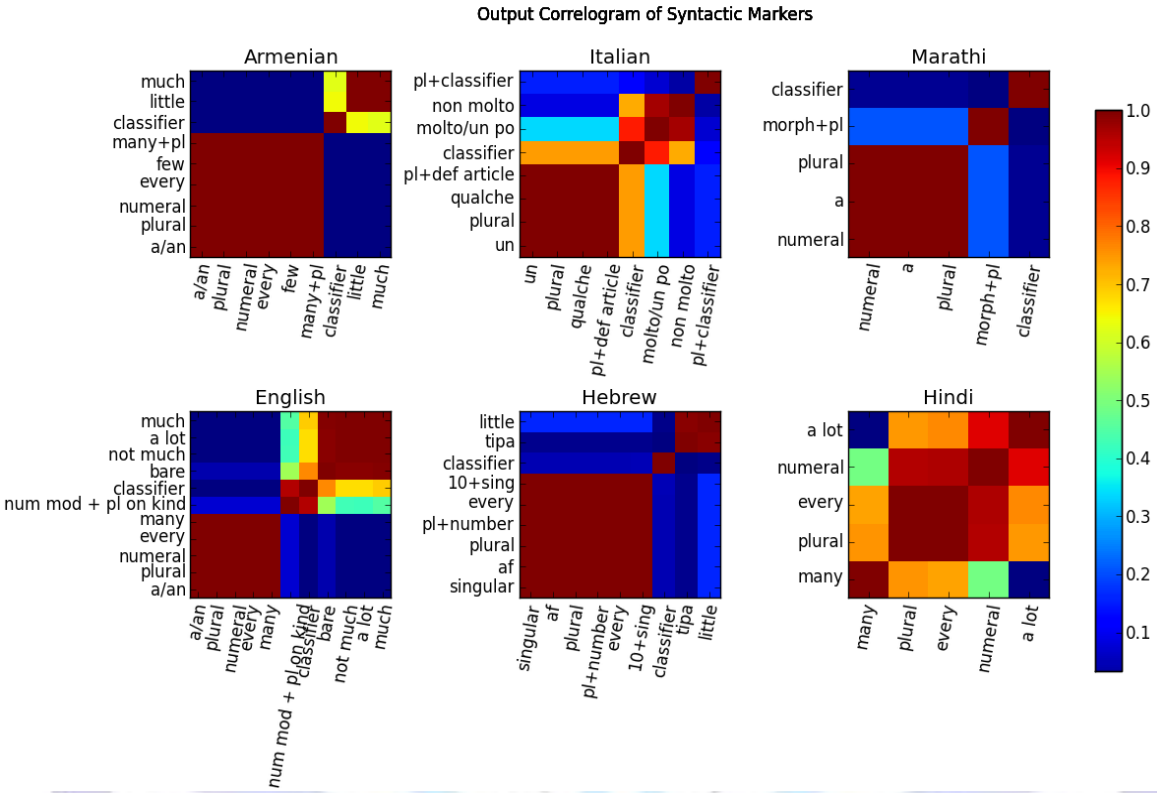

Figure 4. Correlograms for 784 concrete nouns in each of the 6 languages in our study. Dark blue regions represent complete lack of correlation (orthogonal vectors) while dark red regions represent congruent vectors.

First, in what we earlier called a variant of the standard network approach, we present as input the syntactic markers used in the classification of the nouns. Here an input vector is comprised of *n* units, where *n* is the number of nouns (784 in case of concrete nouns and 650 in case of abstract nouns), for each of the syntactic markers. Thus an input includes information on how that particular marker is used over all the nouns. Each input vector is presented once in one iteration, for 50 such iterations, which is also when the synaptic weight matrix is observed not to change with further iterations. We use output units with $T = 0.1$ and a learning rate $k = 0.1$. After obtaining the output firing rates for each input marker at the end of the iterations, we calculate the correlogram, representing how correlated the output vectors are with each other, hence giving information about marker categorization. We show the mean correlograms over 50 distinct network simulations.

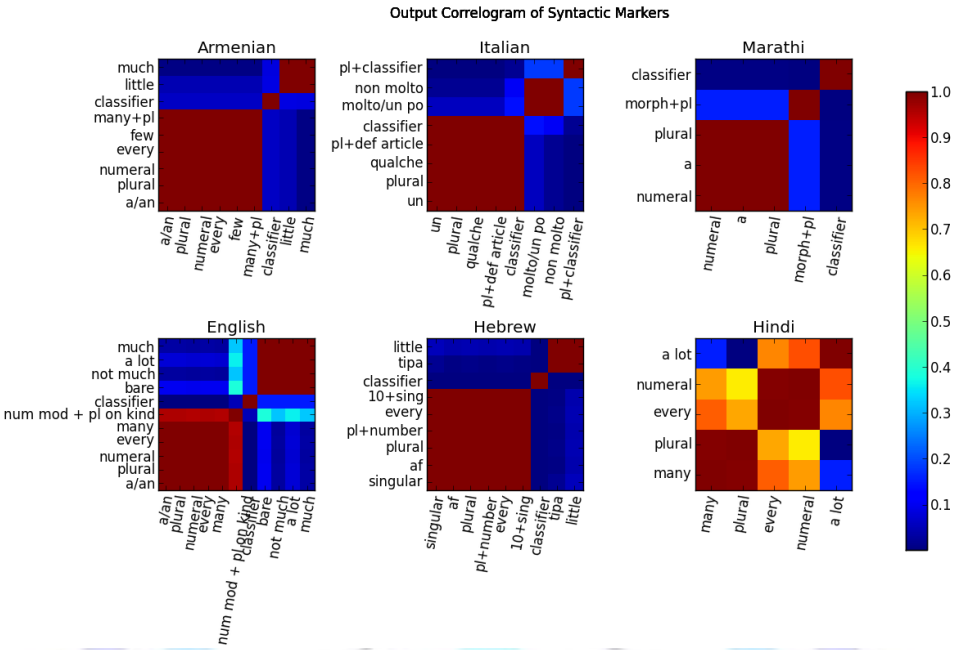

Figure 5. Correlograms, same as in Figure 4, but for 650 abstract nouns. Note that markers are ordered in the same way as in Fig.4

The correlograms in Figure 4 allow s to visually indentify markers that fall in the same category, as self-organized in the output of the network. High levels of correlation between two markers signify close proximity in the firing rates of the output units for that pair of markers, and are represented by warm shades towards brown. For concrete nouns in Armenian, markers like 'a/an', 'plural', 'numeral', 'few', 'every' and 'many+plural' have a correlation of 1, thus occupying the same position in the output space of the network. These are markers that can be applied to count nouns and not to mass nouns. Instead, the typical mass markers of 'measure classifier' form an independent representation, whereas 'little' and 'much' share the same position in output space but distant from the count markers. Italian, Marathi, English and Hebrew follow the same Armenian line of grouping count markers together and having separate but nearby representation for mass markers, distant from the count markers. Hindi is different, as 4 of the 5 markers that were chosen appear to be 'count' in nature, but all show gradation within the broad count category.

Results are similar for abstract nouns except for Italian having fewer graded categorisation than for concrete nouns (Figure 5).

The competitive network can be similarly tested on semantic features based on what value each feature assumes over all the nouns. As seen in figure 6, semantic features are neatly divided into mass and count features. Count features like 'single unit', 'boundary', 'stable shape' and 'degradation' all have a correlation of 1 with each other and 0 with mass features like 'free flow', 'container shape' and 'mixing'. While 'free flow' forms a separate representation, 'container shape' and 'mixing' have the same output activation.

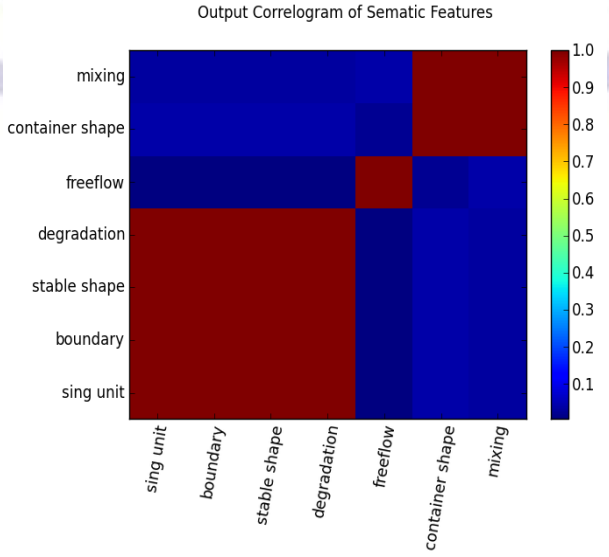

Figure 6. Correlogram of semantic markers for concrete nouns.

3.2 The categorization of nouns

Similar to the procedure in section 3.1, we now present nouns as input to the network and analyze the activation of the output units. The input vector here consists of *n* units for each noun, equalling the number of markers for a language, hence containing information on how the noun is used over all the mass-count markers for that language. The parameters used are N=3, T = 1, k = 0.01 and we run the simulation over 10 iterations. Figure 7 shows the position of the nouns in the 3-D output space, where each axis represents an output unit. Axes are selected such that x, y and z, respectively, represent units in descending order of variance over the values of output activation they span. The shade of each point signifies where that noun (or cluster of nouns, since nouns classified as identical are co-incident) lies on the MC dimension as defined by the Hamming distance from the pure count string (see section 1.1 A). Black indicates a distance of 0, thus pure count, while white indicates a distance of 1, representing a 'mass noun'.

Nouns are seen to approximately fall along a single line for all languages (a predominantly linear structure for English), barring an outlier at 0 which represents inputs, all of which are inactive for a noun. Moreover we can see a gradient from black to white, which implies that nouns, even though not completely faithful, to a great extent lie along a gradient from 'count' to 'mass'. We further visualise the distribution of nouns on this line, so as to assess the frequency of nouns in each cluster. The axis with maximum variance is selected and a histogram of the number of nouns in each cluster along this axis is plotted.

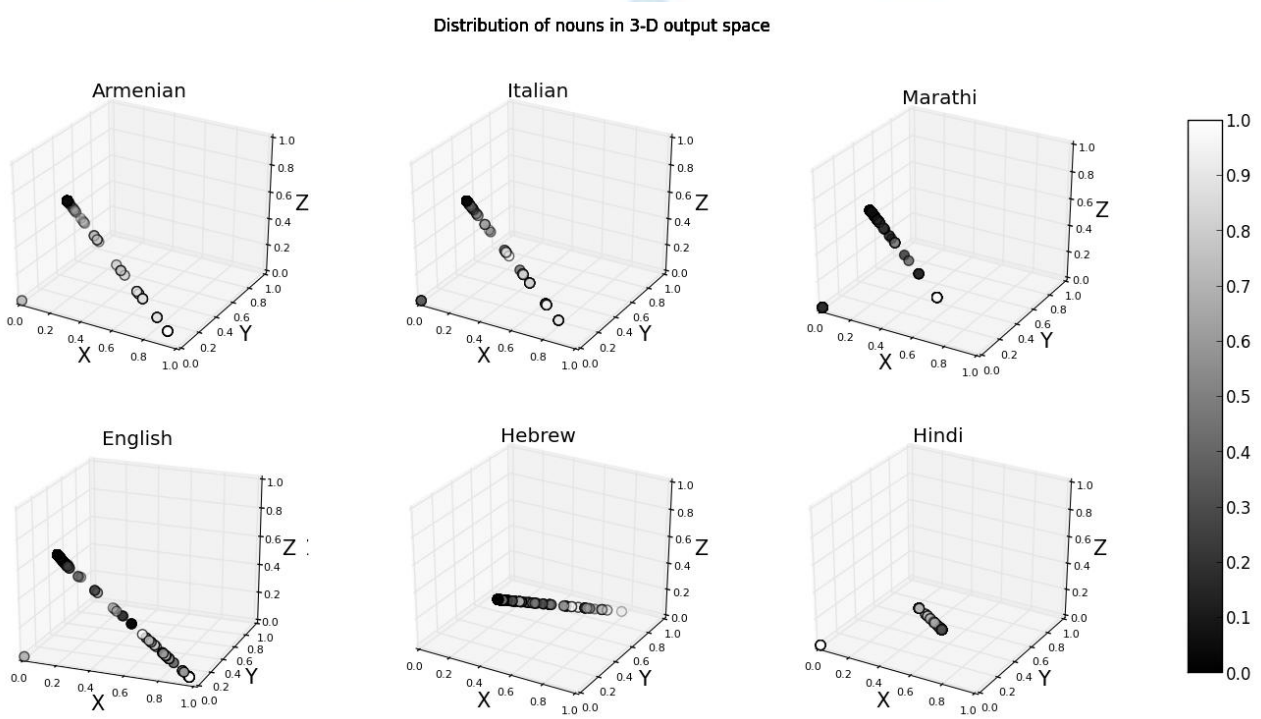

Figure 7. Position of 784 concrete nouns in the output space as defined by 3 output units in 6 languages. The gray scale indicates the Hamming distance of the noun on the MC dimension, from black = 'pure count' to white = 'pure mass'.

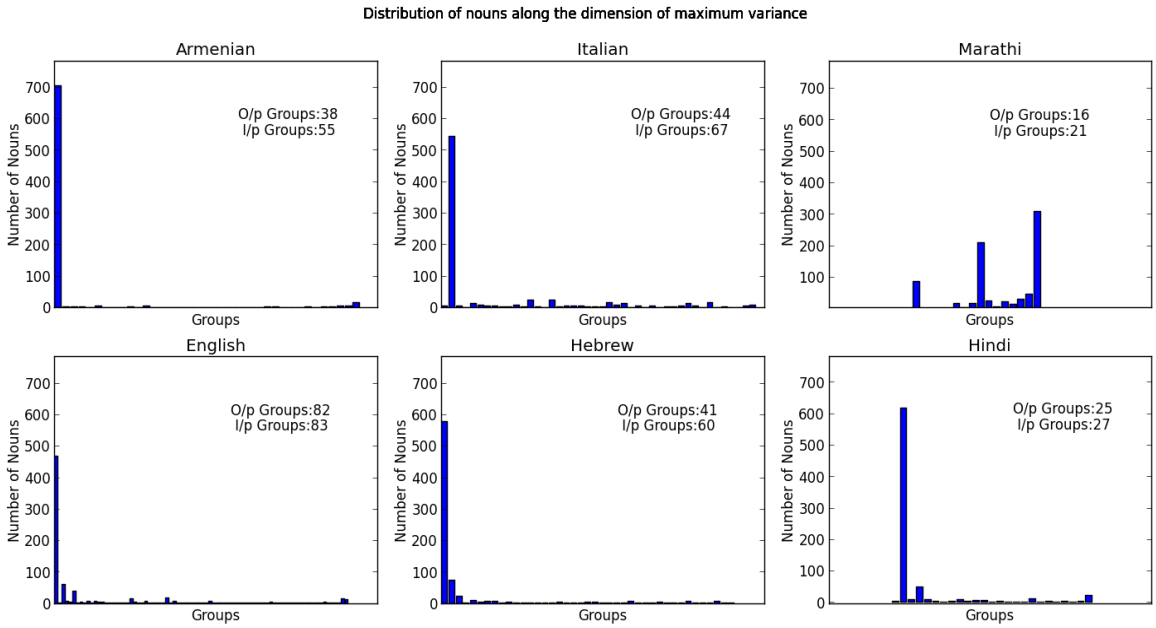

Figure 8. Histogram of nouns in reference to figure 7, along the axis of maximum variance.

Looking at the histogram of the linear alignment of clusters in the output space, we find that the cluster near the count end has the highest number of nouns, which is followed by smaller bars towards the mass end. Marathi is different, in having a significant cluster towards the mass end too.

Is it interesting to note that a dimensionally reduced, entropy preserving representation of the mass-count nouns has a notional similarity to the concept of the MC dimension as in section 1.1A, Figure 1. The MC dimension was introduced as a concept to better understand the mass-count division in terms of the 'pure count' string, but a competitive network with the appropriate parameters is able to bring about a roughly similar distribution without needing a prior definition ad hoc.

Below are shown the same plots as in figure 7 but for 650 abstract nouns. Differences are seen in English and Marathi: in English the group of clusters of black count nouns is more stretched in a separate direction, essentially having two separate lines for mass (white) and count (black) nouns; whereas for Marathi, clusters occupy a narrow space with a small separating distance. Histograms on the dimension of maximum variance show a similar gradation from count to mass and the difference in Marathi with two significant peaks at the count and mass ends each.

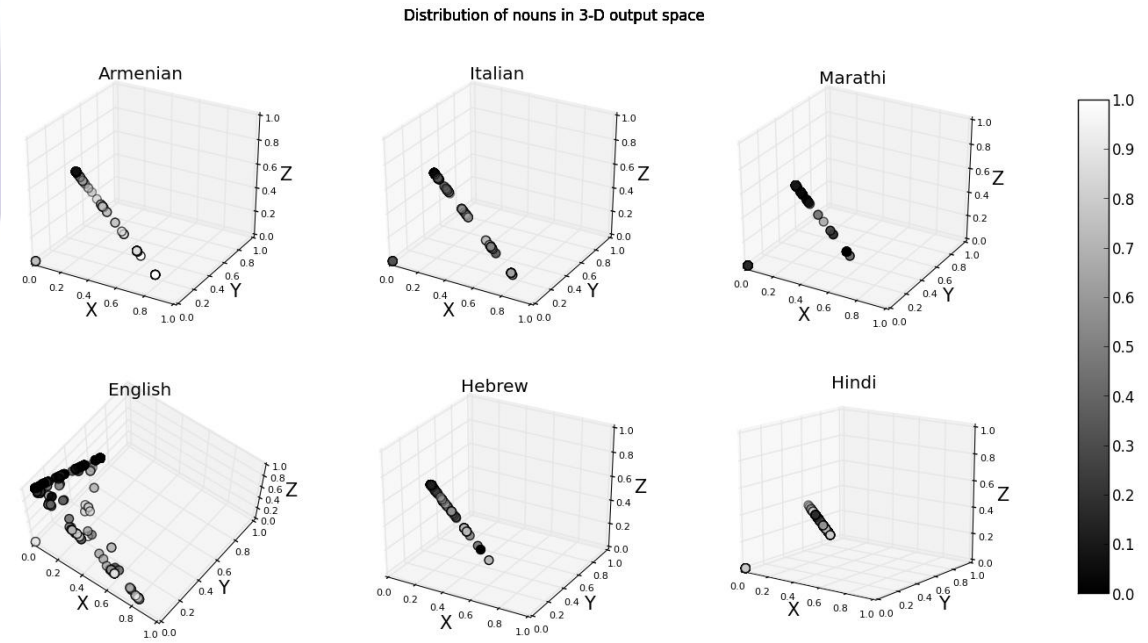

Figure 9.Position of 650 abstract nouns in the output space as defined by 3 output units in 6 languages. The gray scale indicates the Hamming distance of the noun on the MC dimension, from black = 'pure count' to white = 'pure mass'.

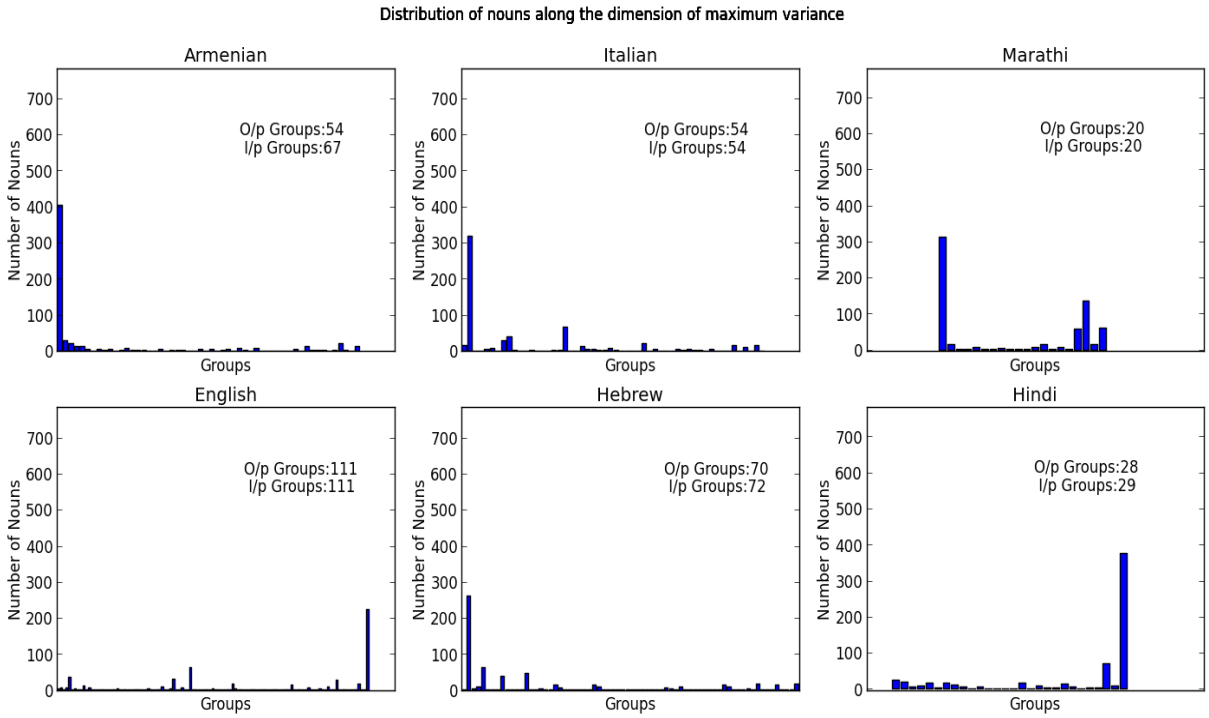

Figure 10.Histogram of nouns in reference to figure 9, along the axis of maximum variance.

3.3 THE SYNTAX-SEMANTICS INTERACTION

As we saw from the information theoretical analysis, the syntax and semantics of the mass-count distinction share only a weak direct link, in the core structure of the count class [8]. Thus acquiring the complete set of syntactic classes from semantic classes is not possible by any learning mechanism, due to a lack of a direct one-to-one correspondence. However it is improbable that syntax and semantics are independently learned without any mutual interaction during learning of mass-count concepts, and there is no evidence that either one is learnt before the other [12]. From the classification of markers above, we see that broad categories of mass and count can indeed be extracted out of the data, interestingly for both syntax and semantics, thus rendering some semantic sense to the syntactic distinction. Classes of nouns formed from these markers do not reflect, however, mass-count information in a straightforward manner between syntax and semantics. Hypothesising an underlying commonality of the mass-count divide between markers of syntax and semantics, we have tested the performance of the competitive network when syntax and semantics are simultaneously part of the input space duringthe learning phase, and test the correspondence between the syntactic and semantic classes after learning.

First, to compare with previous results, we have calculated the baseline mutual information between syntax and semantics by providing only semantic information to the network, with no syntactic information during the learning phase. The mutual information was calculated between syntactic data and the output of the semantic competitive network. When no syntactic information is present at the inputs, the resulting mutual information is about equal to the mutual information between the syntactic and semantic data, as calculated using the procedure in [8].

The competitive network brings about a dimensional reduction from a high dimensional input space to a lower dimensional output space defined by the number of output units. Furthermore the strength of the competition affects how different input clusters collapse onto each other, depending on the distance between them. These processes reduce the entropy from its input value, which critically affects the available mutual information between two data sets, when measured at the output. We observed that the network failed to consistently learn when the number of output units, N, was less than 3 and that there was no noticeable change in information measures for $N \geq 3$. Thus we set N =3 and varied the competition strength T to see the effect on the output entropy of the network. Figure 6 plots the ratio of the output entropy to input entropy of the network for various values of T. Since no information is added during processing, the output entropy is at most equal to the input entropy or less otherwise.

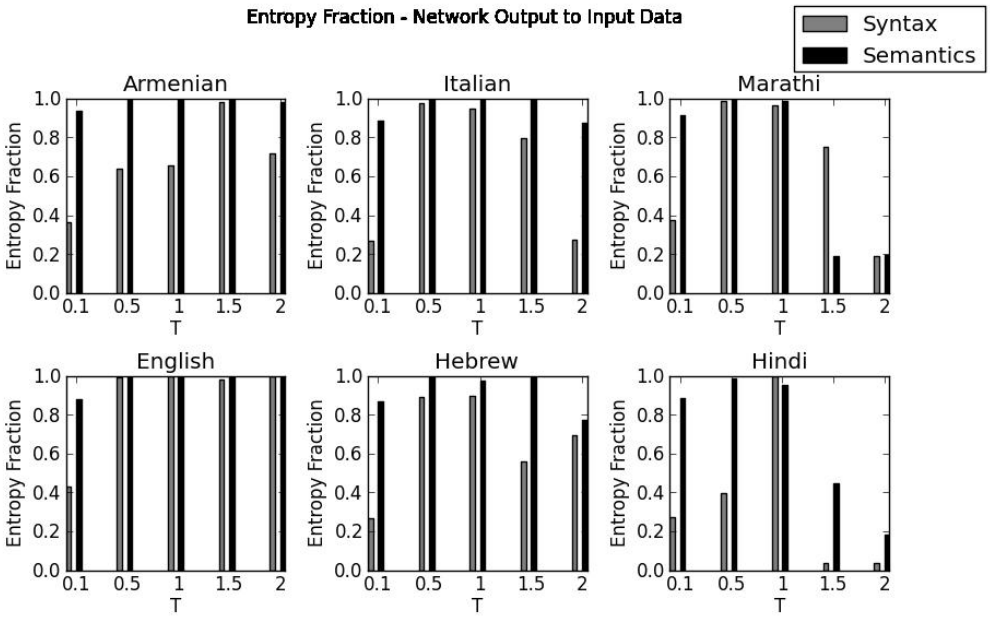

Figure 11. Variation of the output entropy to input entropy ratio for different values of competition strength, T.

Figure 11 shows that, around T = 1, input and output entropies are comparable for all languages (except for Armenian, where they are at T=1.5), and tends to drop on either side (especially for syntax), thus we select T = 1 as the optimal value of competition (when maximum information is retained after processing), to calculate mutual information between syntax and semantics.

Syntactic information is then provided to the network in a partial manner, in a proportion γ, which signifies the fraction of input units of the syntactic segment, of the input string, that are set to the activation levels of the syntactic string of a particular language. $\gamma = 0$ corresponds to when none of the syntactic input units are receiving any information and are set to 0; while $\gamma = 1$ implies that all of the syntactic information is present; for in-between cases a fraction $1 - \gamma$ units are randomly selected and set to 0. Thus we can vary the amount of syntactic information available to the network during learning and test the effect on the syntactic-semantic mutual information and weather the relevant syntactic and semantic classes are brought together in any systematic way. We train and test the network by providing the same proportion γ of syntactic inputs along with the semantic ones.

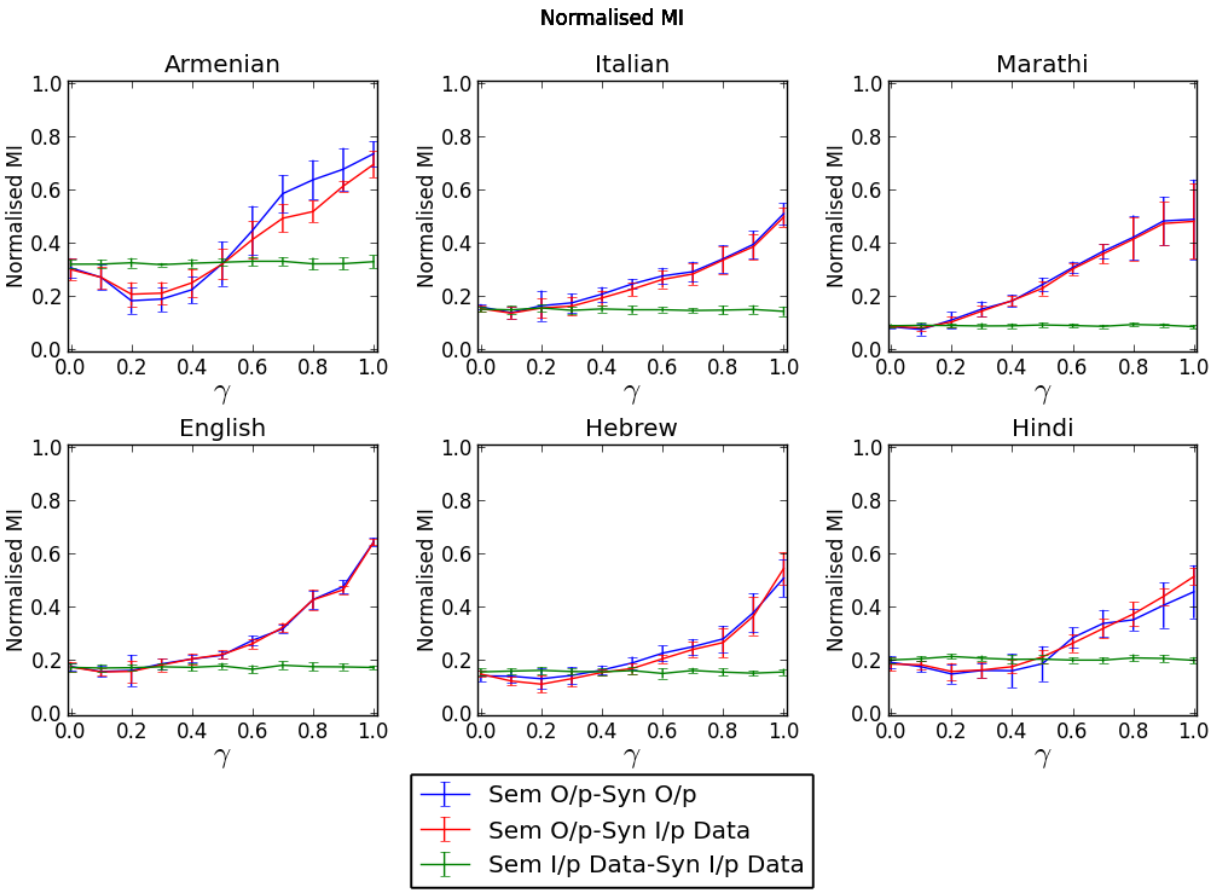

Figure 12. Mean normalised mutual information over 10 independent simulations at various γ for concrete nouns in 6 languages when $\gamma = 1$, *N* =3, 10 iterations.

Results are disappointing in the view that the combined syntactic-semantic inputs have only a limited influence on learning. Figure 12 depicts the performance of the network when it is tested on the semantic inputs while they are incrementally supplied with syntactic information. The green curves represent the mutual information between unprocessed semantic and syntactic input, this is the baseline mutual information between semantic and syntactic data which remains flat, i.e. independent of γ. Unprocessed inputs consist of 'binary strings' that form the raw data as mentioned in section 2.1. The small fluctuations seen in the baseline curve are a result of variation produced due to the sampling correction in calculating mutual information for each independent simulation [8]. The red curves represent the mutual information between the self-organised output of the network and unprocessed syntactic inputs while the blue curves plot the mutual information between self-organised semantic and syntactic outputs. The red and blue curves tend to follow each other closely, implying that the syntactic competitive network results in a dimensionally reduced faithful representation of the syntactic input data. The mutual information rises above the baseline as γ increases above the 0.4-0.5 region for Armenian, English, Hebrew and Hindi, and above the 0.2 region for Italian and Marathi.Thus semantic classes tend to gradually realign, with increasing γ, in such a manner that they correspond more to the syntactic classes as compared to the baseline. This reorganisation is however very limited: at $\gamma = 1$, the normalised mutual information for all languages is in the range of 0.5-0.6, which is around half way towards full agreement. Although interacting with syntax does help some reorganisation of the semantic classes, the divide between syntax and semantics is clear and almost half of the semantic information cannot be shared with syntax at $\gamma = 1$.

The performance of the network is further limited by the fact that it cannot be driven by semantic units only, with no syntactic information during testing. A 'syntactic context' is necessary at the inputs for the network to result in a mutual information performance above baseline. When tested without syntactic information, or with only a partial amount, the drop in the normalised mutual information is significant, with only a tiny trace of learning shown by the network.

4. CONCLUSIONS

This is clearly a first attempt at exploring the learnability of this specific subdomain with a simple neural network. The results lead to a number of inferences.

1. In most languages, syntactic markers tend to categorize 'spontaneously' between mass and count markers, lending validity to the intuitive perception of a quasi-binary distinction. This is not fully true, however, and particularly in Hindi the markers chosen show a graded distribution of mutual correlations.
2. Nouns, instead, tend in most languages to distribute quite closely along a line which coincides with the main

mass/count dimension introduced in our previous study [8]. Along this line, nouns are very crowded at the count end, and scattered all along towards the mass end. Their distribution is therefore graded rather than binary, with no emergence of a single 'mass' class, but rather of several non-exclusive but distinct ways for a noun to be different from pure count. For example, in Armenian (Figure 8) nouns like 'bird' and 'ship' belong to the 'pure count' class while 'troop 'and 'lunch' are in the 9th class away from the count class. On the mass end, nouns like 'cotton' and 'milk' are at the extreme mass end of the spectrum while 'coffee' and 'wheat' are more mass-like nouns but not at the pure mass end. The exception is English, where there are at least two clear non-equivalent dimensions of non-countability.

Both the above observations are interesting because the mass-count information in the categorisation arises on its own. The markers, in some cases, very cleanly segregate themselves into mass and count. The nouns are reduced to a one dimensional representation along a mass-count spectrum. Even though the network fails to associate specific syntactic markers with specific nouns based on the semantics, the network does develop a 'concept', if we may say, of what the mass-count classification is. The diversity and richness of this classification however, prevents a simple network to learn specific associations. Which brings us to the third observation,

3. Finally, the lack of a significant mutual information between semantics and syntax implies, as we have verified, that the latter cannot be extracted solely from the former. Further, when allowing the competitive network to self-organize on the basis of full semantics and partial syntactic inputs, and testing it with the full syntactic inputs, the mutual information obtained with the full syntactic usage distribution is only at most about half the corresponding entropy value. This occurs in fact only when the full syntax is given in the input also at training, and it indicates that giving also semantics information affects negatively rather than positively the performance of the network.

Overall, these observations do not clarify how mass count syntax may be acquired by humans with neurally plausible mechanisms, on the basis of matching semantic information and syntactic properties. They reinforce the conclusions of our earlier study, that mass count syntax is far from a rigid binary contrast. It appears as the flexible, language-specific and even, when within language, speaker-specific usage of a variety of binary markers to a quantitatively and qualitatively graded repertoire of nouns, where being non-count can be expressed in many ways. This supports recent work by Rothstein (in press) arguing that the count/mass contrast is not a reflection of a contrast between atomicity and homogeneityor between objects and stuff. Instead, it reflects a perspectival contrasts between entities presented grammatically as countable and those presented as contextually non-countable.Taking the properties of the referents into account is of limited use in generating grammatical generalizations.

REFERENCES

[1] Pinker S. 1995. The language instinct: The new science of language and mind (Vol. 7529). Penguin UK.

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

[3] Pelletier, F.J., 2010. Descriptive metaphysics, natural language metaphysics, Sapir-Whorf, and all that stuff: Evidence from the mass-count distinction. *Baltic International Yearbook of Cognition, Logic and Communication*, *6*(1), p.7.

[4] Bale A.C. & Barner D. 2009. The interpretation of functional heads: Using comparatives to explore the mass/count distinction. Journal of Semantics 26,217–252.

[5] Chierchia G. 2010. Mass nouns, vagueness and semantic variation. Synthese 174, 99–149.

[6] Prasada S., KragF.& Haskell T. 2002. Conceiving of entities as objects and stuff. Cognition 83, 141–165.

[7] Link G. 1983. The logical analysis of plurals and mass terms: A lattice-theoretical approach. In R.B. Bäuerle, C.Schwarze & A.von Stechow (eds.) Meaning, Use and Interpretation, 303–323. Berlin: Moutonde Gruyter. [Reprinted in P.Portner & B. Partee (eds.). 2002. Formal Semantics: The Essential Readings, 127–146. Oxford: Blackwell.]

[8] "Author" 2013.

[9] Elman J. L. 1991. Distributed representations, simple recurrent networks, and grammatical structure. Machine Learning, 7:195-224.

[10] Nyamapfene A. 2009. Computational investigation of early child language acquisition using multimodal neural networks: a review of three models. Artificial Intelligence Review 31 (1-4): 35-44.

[11] Chiarelli, V., El Yagoubi, R., Mondini, S., Bisiacchi, P. and Semenza, C., 2011. The syntactic and semantic processing of mass and count nouns: An ERP study. *PloS one*, *6*(10), p.e25885.

[12] Nicolas D.A. 2010. Towards a semantics for mass expression derived from gradable expressions. Recherches Linguistiques de Vincennes 39, 163–198