Learning properties of Noun Phrases: from data to functions

Valeria Quochi, Basilio Calderone

Istituto di Linguistica Computazionale CNR, Scuola Normale Superiore
Via Moruzzi 1 Pisa -Italy, Piazza dei Cavalieri 7 Pisa -Italy
valeria.quochi@ilc.cnr.it, b.calderone@sns.it

Abstract

The paper presents two experiments of unsupervised classification of Italian noun phrases. The goal of the experiments is to identify the most prominent contextual properties that allow for a functional classification of noun phrases. For this purpose, we used a Self Organizing Map is trained with syntactically-annotated contexts containing noun phrases. The contexts are defined by means of a set of features representing morpho-syntactic properties of both nouns and their wider contexts. Two types of experiments have been run: one based on noun types and the other based on noun tokens. The results of the type simulation show that when frequency is the most prominent classification factor, the network isolates idiomatic or fixed phrases. The results of the token simulation experiment, instead, show that, of the 36 attributes represented in the original input matrix, only a few of them are prominent in the re-organization of the map. In particular, key features in the emergent macro-classification are the type of determiner and the grammatical number of the noun. An additional but not less interesting result is an organization into semantic/pragmatic micro-classes. In conclusions, our result confirm the relative prominence of determiner type and grammatical number in the task of noun (phrase) categorization.

1. Introduction

We describe here an exploratory study on the acquisition of functional properties of nouns in language use. This work models contextual and morpho-syntactic information in order to discover fundamental properties of Noun Phrases (NPs henceforth) in Italian[1]. Context analysis is crucial in our investigation: we assume in fact that nouns *per se* have no semantic/functional property other than the default referential one. However, depending on the wider context in which they occur, nouns or better noun phrases, may be used in different ways: to predicate, to refer to specific, individuated entities or they can be be more generally type referring (Crof and Cruse, 2004).

Our aim in this work is to see whether, given a large set of (psychologically plausible) morpho-syntactic contextual features and an unsupervised learning method, (functional) similarities of nouns emerge from language use. We set up two simulation experiments using a Self-Organizing Map learning protocol (section 3.1.). For the present purposes, we analyze the final organization of a SOM trained with morphosyntactically-defined contexts of noun phrases in order to investigate the prominence of the various morpho-syntactic properties, i.e. the relevant dimensions on the basis of which the map self-organizes and the correlation to linguistic functional properties of noun phrases.

The present paper is organized as follows: first we briefly mention some related works on the acquisition of deep lexical properties of nouns in languages other than Italian. Section 3. presents the methodology adopted: the learning system, the dataset and the feature extraction and representation process. Section 4. describes the experiments based on noun types and noun tokens and briefly discusses the outcomes. Finally a discussion of the result and the future work is given in Figure 5.

[1]The term Noun Phrase (NP) will be used here as a theory-independent general label for various kinds of nominal chunks (noun, determiner+noun, adjective+noun, ...).

1.1. Linguistic Background

The standard function of nouns is to name portions of reality, to label entities. A noun typically denotes the kind of thing that its referent belongs to. Naming is therefore a kind of categorization. Assuming this, we will say that the primary cognitive function of nouns is to form a classification system of things in the world that we use in referring to them (Dryer, 2004, 50).

Nouns, however, are seldom used in isolation; noun phrases (or more generally nominal chunks) may have different, contextual functions. Functions of noun phrases are to signal the countability, new vs. given status, generic or individuated character of the entity referred to, and its degree of referentiality (Crof and Cruse, 2004; Delfitto, 2002).

In many languages, the type of determiner present in the NP and the number of the noun are the linguistic cues that are generally held responsible for signaling the function in context (countability, givenness and specificity in particular). However, there is considerable variation both among and within languages. In some theories, determiners are acknowledged great importance, they are even considered the heads of noun phrases (i.e. Sugayama and Hudson (2005). In Cognitive Linguistics, instead, they are assigned a fundamental property, they signal the "grounding" of a noun phrase (its contextual identification within the speech event, (Langacker, 2004, 77-85)).

Countability is considered responsible for the construal of an entity as an individuated unit. This difference corresponds to the bound/unbound structural schematization in Cognitive Linguistics (Langacker, 1987). Countability may also construe an entity as of a specific type, e.g. *chair* vs. *furniture* (Crof and Cruse, 2004).

Assuming that naming is categorizing and that categories are not neat, but have fuzzy boundaries, the meaning and function of nouns cannot be totally pre-established, but must be construed dynamically in context. Therefore, the structure of the noun phrase and its surrounding context should reveal the specific construal of the noun. Put in

another way, if nouns conceptualize categories then their functions and denotations should emerge from their actual use in language.

1.2. Related Works

Works automatic acquisition of so-called *deep lexical (or grammatical) acquisition*, in particular of countability (and specificity), exist for English and for Spanish. All of them however make use of supervised categorization approaches in which the possible categories are set *a priori*.

Baldwin and Bond (2003) for example describe a classification system for the acquisition of countability preferences for English nouns from corpora based on 4 countability categories. Such categories were determined by theoretical insights/preconcepts about the grammar and praxis in the computational linguistics community (i.e. they looked at the classifications given in COMLEX and in ALT-J/E, dictionaries for a machine translation system.

Peng and Araki (2005) developed a system for discovering countability of English compound nouns from web data.

Williams (2003) reports on a series of experiments both on human subjects and using a neural network system for the acquisition of gender information.

Bel et al. (2007) is an interesting experiment on the acquisition of countability of Spanish nouns from corpus data using a Decision Tree learner.

Classification systems in general set *a priori* the number and types of classes into which they want to classify the inputs; therefore, a theory of the plausible classes must be presupposed. Our aim instead is to observe if and what kind of categorial organization emerges from a morpho-syntactically described set of nominal contexts and what are the linguistic features that allow for an organization of the input. An interesting observations coming from previous related works is that distributional information of features is a good representation for the task of learning deep properties of lexical items. Bel et al. in particular adopt a representation format of feature similar to ours: they encode the features occurring in the contexts in terms of presence/absence, i.e. of binary values. This seems to work fine also for unsupervised approaches as the one described here.

2. The goal

The main goal of the set of experiments presented here is to study the 'contextual representations' of NP constructions based on their morpho-syntactic properties and those of the contexts in which they appear, in order to investigate to what extent the cognitive-semantic properties of noun phrases, as identified in the literature, are actually emergent from the language use, and therefore can be learnt from texts. More specifically, our research question here is what kind of and how much morpho-syntactic information is necessary to obtain a functional classification of noun usages?

The main aim of this exploratory study on the acquisition of Italian deep lexical properties is to observe the behavior of nouns in post verbal positions in an unsupervised, auto-organizing system. For this reason, we tried to represent as much distributional morphosyntactic information as possible for the target noun contexts, in order to provide the network with many possible linguistic cues and observe if it managed to come up with some categorization and which linguistic cues emerged as most relevant. This is also a means to find support/disconfirms to theoretical assumptions on the functional properties of nouns.

3. Methodology

For the investigation on the emergence of functional properties of nouns phrases we adopt an unsupervised connectionist approach using Self-Organizing Maps (Kohonen, 2001).

3.1. Self-Organizing Maps

The Self-Organizing Map (SOM) (Kohonen, 2001) is an unsupervised neural network algorithm that is able to arrange complex and high-dimensional data space into low-dimensional space so that similar inputs are, in general, found near each other on the map. The mapping is performed in such a way that the topological relationship in the n-dimensional input space is maintained when mapped to the SOM. The final organization of the SOM reflects internal similarities and frequency distribution of the data in the training set.

The location of input signals tend to become spatially ordered as if some meaningful coordinate system for different input features were being created over the map. In this perspective the location or coordinates of a neuron in the map correspond to a particular domain of the input patterns. A SOM, in this sense, is characterized by the formation of a topographic map of the input patterns in which the spatial location (the coordinates) of a neuron in the map are indicative of intrinsic features exhibited by the input (Fig. 1).

In the experiments described below, we used a standard SOM (20x20) learning protocol. The learning protocol proceeds by first initializing the synaptic strengths (or connection weights) of the neurons in the map by assigning values picked from a random or uniform distribution. This point is crucial because no *a priori* order or knowledge is imposed onto the map. After the initialization, the self-organization process involves two essential steps:

- **Step 1**: the input data vector x is compared to the weight vectors m_i and the Best Match Unit (BMU) m_c is located.
- **Step 2**: the neurons within the neighborhood h_{ci} of c are tuned to the input vector x.

These steps are repeated for the entire training corpus.

In Step 1, the BMU to the input vector is found. The BMU is determined using the smallest Euclidian distance function, defined as $\| x - m_i \|$. The BMU, m_c, is found using the following equation:

$$\| x - m_c \| = min\{\| x - m_i \|\} \quad (1)$$

Once the BMU is found, Step 2 initiates. This is the learning step in which the map surrounding neuron c is adjusted towards the input data vector. Neurons within a specified geometric distance, h_{ci}, will activate each other and learn something from the same input vector x. This will have

a smoothing effect on the weight vectors in its neighborhood. The number of neurons affected depends upon the neighborhood function. The learning algorithm is defined as:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \qquad (2)$$

where $t = 0, 1, 2, \ldots$ is the discrete-time coordinate. The function $h_{ci}(t)$ is the neighborhood of the winning neuron c, and acts as the so–called *neighborhood function* that is a smoothing kernel defined over the lattice points. The function $h_{ci}(t)$ is usually defined as a Gaussian function:

$$h_{ci} = \alpha(t) \cdot exp\left(-\frac{\parallel r_c - r_i \parallel^2}{2\sigma^2(t)}\right) \qquad (3)$$

where $\alpha(t)$ is a learning rate and the parameter $\sigma(t)$ defines the radius of $N_c(t)$. In the original algorithm (Kohonen, 2001), both $\alpha(t)$ and $\sigma(t)$ are monotonically decreasing functions of time.

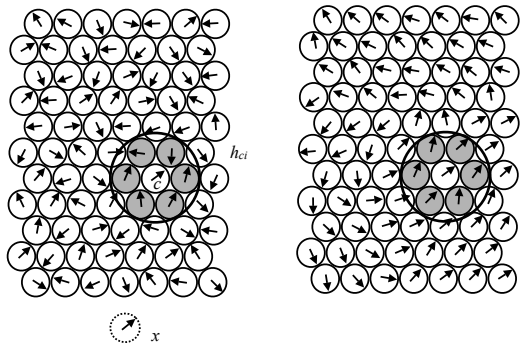

Figure 1: A randomly initialized SOM after one learning step (left panel) and a fully trained SOM (right panel).

The training process is illustrated in Figure 1. First, the weight vectors are mapped randomly onto a two–dimensional grid and are represented by arrows pointing in random direction (left panel of the figure). In the random SOM the closest match to input data vector x has been found in the neuron c (Step 1). The neuron within the neighborhood h_{ci} learn from neuron c (Step 2). The size of the neighborhood h_{ci} is determined by the parameter $N_c(t)$, which is the neighborhood radius. The weight vectors within the neighborhood h_{ci} tune to, or learn from, the input data vector x. How much the vectors learn depends upon the learning rate $\alpha(t)$. In Figure 1 (right panel) fully trained map is displayed. In a fully trained map, a number of groups should emerge, with the weight vectors between the groups 'flowing' smoothly into the different groups. Generally the SOM is trained in two phases. The first phase is a rough training of the map in which the system is allowed to learn a lot from each data vector. Therefore, learning rate and radius are high in this first phase. The second phase is a fine-tuning phase, in which the SOM learns less at a time, but data input are introduced to the map more times. Thus, learning rate and radius are lower than in the first phase, but the training length is much higher.

3.2. U-matrix

The distance between the neighboring codebook vectors highlights different cluster regions in the map, which is thus a useful visualization tool. The distance for each neuron is the average distance between neighboring codebook vectors. Neurons at the edges of the map have fewer neighbors. The average of the distance to the nearest neighbors is called unified distance and the matrix of these values for all neurons is called *U-matrix* (Ultsch and Siemon, 1990).

In a **U-matrix** representation, the distance between adjacent neurons is calculated and presented with different colorings between adjacent positions on the map. Dark colorings highlight areas of the map whose units react consistently to the same stimuli. White coloring between output units, on the other hand, corresponds to a large distance (a gap) between their corresponding prototype vectors[2]. In short, dark areas can be viewed as clusters, and white areas as chaotically reacting cluster separators.

In NLP, SOMs have been previously used to model continuous and multidimensional semantic/pragmatic spaces (Ritter and Kohonen, 1989; Honkela et al., 1995) as well as morphology acquisition in a given language (Calderone et al., 2007). Li and colleagues (Li et al., 2004), moreover, have exploited SOMs for simulating the early lexical acquisition by children.

3.3. The Dataset: Feature Extraction and Representation

As input training data we used a collection of Italian verb-noun contexts automatically extracted and analyzed (at chunking level) from a 3 million corpus of contemporary Italian newspaper article (the PAROLE subcorpus (Bindi et al., 2000)[3]). The training data consist of 847 noun types and 2321 tokens. Each noun token has been extracted from the corpus with its surrounding context. In order to normalize the context type and size for each noun, we selected nouns occurring in post verbal position as potential direct objects of the verb *fare* 'do/make' together with the verb chunk and two chunks on the right of the noun.

Ex. <verb chunk: head:*fare*> <nominal chunk> <chunk 1> <chunk 2>

Only contexts of *fare* have been chosen because it is the most general purpose Italian verb governing a variety of noun types and it is most often used as a light verb or and a causative. Therefore, we can (simplistically) assumed *fare* to have little impact on the semantic functions of object noun phrases[4].

[2]Unfortunately, the black and white pictures on the paper here do not allow a proper appreciation of the map.

[3]For text chunking we used the *Italian NLP Tools* developed at ILC-CNR (Ita,) and (Lenci et al., 2003), in particular for details on the chunker.

[4]Except that it will occur in many support verb constructions. However, this should not pose problems to our results, rather it will be interesting to see whether they are classified separately.

3.3.1. Selected Features

The extracted contexts are represented as vectors of 36 features representing the morpho-syntactic properties of their elements. Given our goals, we did not pick out features that, in the literature, are considered to be good cues for noun functions, rather we represent most morpho-syntactic properties of the entire noun contexts, as resulting from automatic text chunking. Specifically, for each noun item in the corpus, we represent as binary features: verb finiteness, mood, number and person, presence/absence of a causative, noun gender, number and person, determiner type (zero, definite or indefinite), type of preposition in the chunk following the noun[5].

In the first experiment, the dataset is organized around noun types and features, therefore, encode frequency indexes for each variable (i.e. feature). The second experiment, instead, considers noun tokens and features are therefore encoded as binary values simply representing the presence/ absence of each feature in the contexts of each occurrence of the nouns in the corpus.

Running our SOM on the dataset described above we obtained a semantic-functional clustering of Italian noun phrases governed by the verb *fare* 'do/make' and we identify the relevant features. We applied the SOM algorithm for simultaneous clustering and visualization of the data. The visualization provided a means for understanding and qualitatively evaluating the resulting clustering and the feature selection.

4. Experiments

4.1. Experiment I: Type Simulation

In Experiment I we performed a 'global' distributional analysis of NP contexts, i.e. an analysis based on noun types.

Let N be the set of 847 noun types, namely $N = \{n_i \mid i = 1, \ldots, 847\}$. We take into account the 36 set of contextual/morpho-syntactic features. It is natural to represent each NP n_i as a vector $X^{(i)} = X_1^{(i)}, \ldots, X_{36}^{(i)}$ where each component $X_j^{(i)}$ represents the frequency of the j-th contextual/morpho-syntactic features for that particular NP n_i.

4.2. Experiment II: Token Simulation

In Experiment II we performed 'local' distributional analysis of NP contexts: contexts are considered on the basis of noun tokens.

Let N be the set of 2321 NP token, namely $N = \{n_i \mid i = 1, \ldots, 2321\}$. Again, we take into account the set of 36 contextual/morpho-syntactic features. It is natural to represent each NP n_i as a vector where each component $X_j^{(i)}$ represents the absence/presence (in binary encoding) of the j-th contextual/morpho-syntactic features for that particular NP n_i.

Figure 3 presents the resulting U-matrix map. For readability purposes, the map in Figure 3 has been labeled indicating the most frequent noun (context) for each neuron. Neurons in reality represent clusters of noun contexts behaving in a similar way according to the map.

[5]The translation of these features into binary variables yielded 36 components.

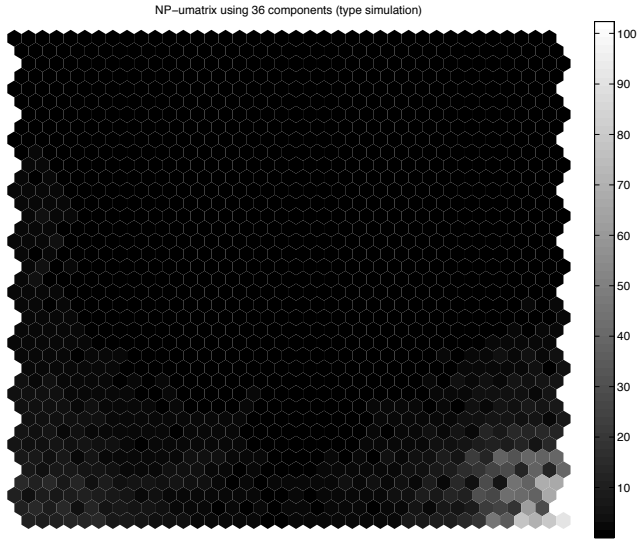

Figure 2: Type Simulation: U-matrix with 36 components.

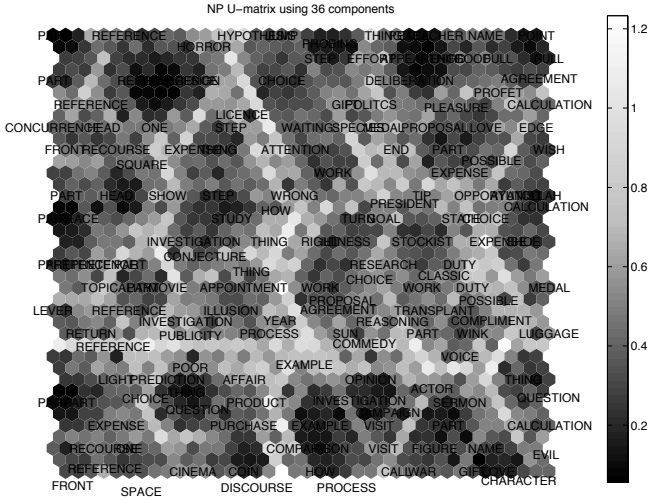

Figure 3: Token Simulation: U-matrix with 36 components.

5. Results and Discussion

5.1. Experiment I

Experiment I uses contexts organized by noun types. The main "categorizing" cue exploited by the SOM appears to be simply frequency. No morpho-syntactic component represented in the vectors appear to be prominent. As a result, we see on the map that the system learns mostly highly fixed, lexicalized or idiomatic expressions (isolated in the area at the bottom right corner of the map in 2.

Ex.

far parte (di) 'lit. make part of', 'be part of, belong to'
far riferimento (a) 'lit. make reference to', 'refer to'
far capo (a) 'lit. make head to', 'refer to'
far leva (su) 'lit. make lever on', 'to play on'
far fronte (a) 'lit. make front to', 'to face'

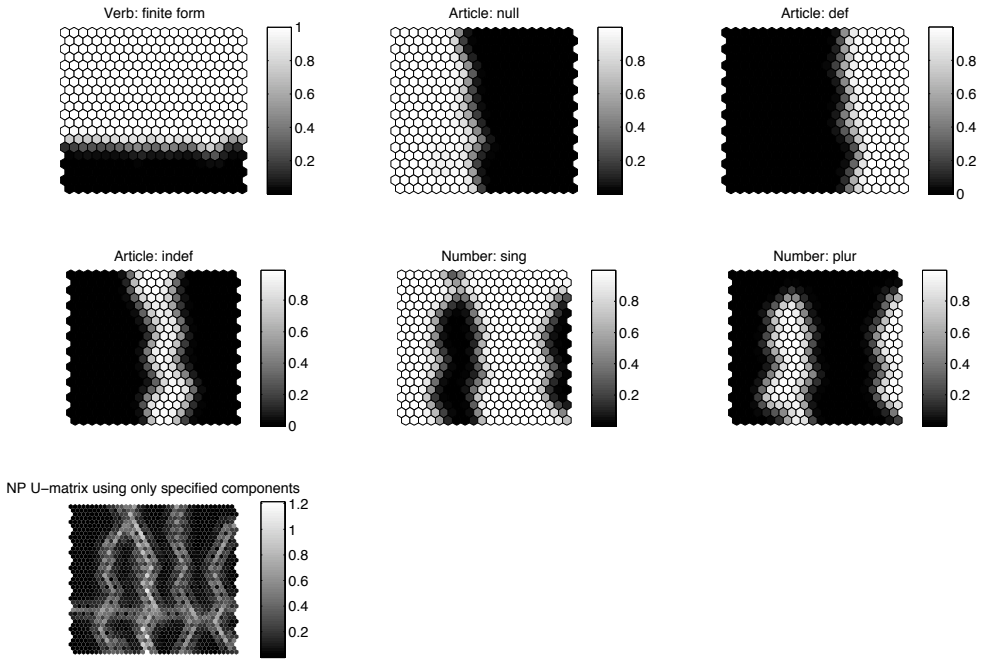

Figure 4: NP U-matrix using 5 components of the input data.

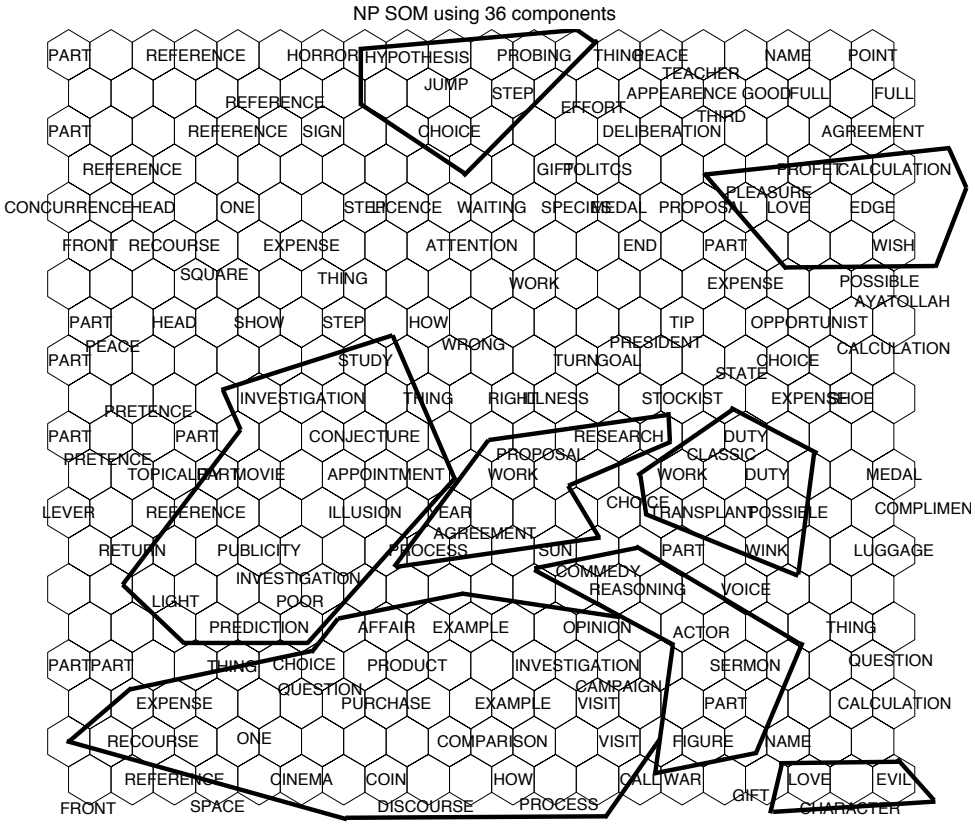

Figure 5: Pragmatic/Semantic similarities. Only the label with most instances is kept

More precisely, in Experiment I, where nouns are considered by type (i.e. lemma) and therefore their vectors represent frequency weights of the morphosyntactic properties of their contexts, the network learns the most typical contexts of nouns.

Given the not very big size of the training corpus, it is natural that only a few, most fixed and common phrases are isolated (and in fact it is an interesting result in itself). It would be interesting to observe the behavior of the SOM within a larger corpus, both in terms of noun type and token dimension. With such conditions, we would expect a larger number of lexicalized phrases to be isolated.

For the goal of the present investigation, however, the results of this experiment are not of particularly interest, since they do not show any prominent role of linguistic cues.

5.2. Experiment II

Experiment II, based on NP tokens, instead, gives interesting results. In this case, we clearly observe an organization of the map into macro classes (Figure 5), where boundaries are quite neat. The most evident classes are three vertical ones, that appear to be determined by the distribution of determiners in corpus. Two macro classes (at the top and bottom of the map) are determined by the distribution of the form of the verb: i.e. finite and infinite. And two other macro classes are based on the distribution of grammatical number of the noun. Of the 36 features used in the last set of experiments, only 6 morpho-syntactic contextual features appear to be relevant for the system to cluster the noun phrases, namely the type of determiner, the grammatical number of the head noun, and the form of the verb. As we discussed in the background section, type of determiner and grammatical number are, in fact, cues that are held responsible for the expression of the cognitive functions of noun phrases in the literature (i.e. determiner phrase as a basic functional category).

In order to test the impact of the the most prominent morpho-syntactic contextual features over the self-organization of the data set (in token simulation) we isolated the activation of 6 morpho-syntactic features that seem to influence the SOM topology more than the others. The activation of each of the 6 features is displayed on the SOM grid (Figure 4) according to the activation values. Also the U-matrix calculated only from these components is reported (Figure 4). A topological overlapping of this new U-matrix with the U-matrix using 36 features (Figure 2) confirms, in visualization terms, the prominence of the features selected.

Secondly, we also observe a learning of semantic/pragmatic categories: the network organizes the contexts into bundles of semantically/pragmatically similar nouns, thus providing another empirical support to the distributional hypothesis. Figure 5 shows some of the most evident semantic bundles learned by the system.

This result is even more interesting, in that no proper semantic feature has been represented explicitly in the vectors (a part from features like grammatical gender and number, which can be considered as morpho-semantic features).

6. Future Work

The work described in the present paper can be seen as a usage-based modeling of noun "contextual representations", where the context acceptation drives the pragmatic(/semantic) function of the NP itself. Figure 5 presents only the labels for the higher frequent NPs. At the current stage of the experiments, it is not possible to analyze in detail the clusters of contexts represented by each neuron. However, we expect that a more detailed analysis of the NP-context groupings, especially in the token simulation SOM, would highlight not only a more consistent number of contextually similar interrelated NPs, but also sub-areas of "contextual spaces" in which the NPs receive similar treatment in pragmatic/functional terms. This qualitative investigation will be the subject of future experiments.

7. Acknowledgments

This work originates from a collaboration between the two authors which started out of curiosity during their PhD research. We would like to thank our institutions that have indirectly permitted and supported such activity. We would also like to thank the three anonymous reviewers for they valuable comments and suggestions.

8. References

T. Baldwin and F. Bond. 2003. Learning the countability of english nouns from corpus data. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, July 2003*, pages 463–470.

N. Bel, S. Espeja, and Monserrat Marimon. 2007. Automatic acquistion of grammatical types for nouns. In *Proceedings of NAACL HLT, April 2007, Companion Volume*, pages 5–8, Rochester, NY. ACL.

Remo Bindi, Paola Baroni, Monica Monachini, and Elisabetta Gola. 2000. PAROLE-sottoinsieme. Internal report, ILC-CNR, Pisa.

B. Calderone, I. Herreros, and V. Pirrelli. 2007. Learning inflection: The importance of starting big. *Lingue & Linguaggio*, (2):175–200.

W. Crof and A. Cruse. 2004. *Cognitive Linguistics*. CUP, Cambridge.

D. Delfitto. 2002. *Genericity in language. Issues of syntax, logical form and interpretation*. Dell'Orso, Alessandria.

Matthew S. Dryer. 2004. Noun phrases without nouns. *Functions of Language*, 11:43–76(34).

T. Honkela, V. Pulkki, and T. Kohonen. 1995. Contextual relations of words in Grimm tales analyzed by self-organizing map. In F. Fogelman-Soulié and P. Gallinari, editors, *Proceedings of International Conference on Artificial Neural Networks, ICANN'95*, volume II, pages 3–7, Nanterre, France. EC2.

Italian NLP tools. http://foxdrake.ilc.cnr.it/webtools/.

T. Kohonen. 2001. *Self Organizing Maps*. Springer Verlag, Berlin.

R. Langacker. 1987. *Foundations of Cognitive Grammar I: Theoretical prerequisites*. Stanford University., Stanford.

R. W. Langacker. 2004. Remarks on nominal grounding. *Functions of Language*, 11:77–113(37).

A. Lenci, S. Montemagni, and V. Pirrelli. 2003. *Chunk-it*. an italian shallow parser for robust syntactic annotation. In *Computational Linguistics in Pisa - Linguistica Computazionale a Pisa*, volume XVI-XVII. IEPI, Pisa-Roma.

P. Li, I. Farkas, and B. MacWhinney. 2004. Early lexical development in a self-organizing neural network. *Neural Networks*, 8–9(17):1345–1362.

J. Peng and K. Araki. 2005. Detecting the countability of english compound nouns using web-based models. In *Second International Joint Conference on Natural Language Processing, Jeju Island, Korea, October 11-13, 200. Companion Volume to the Proceedings of Conference*, pages 103–107. Asian Federation of Natural Language Processing.

H. Ritter and T. Kohonen. 1989. Self-organizing semantic maps. *Biological Cybernetics*, (61):241–254.

K. Sugayama and R. Hudson, editors. 2005. *Word Grammar. New Perspectives on a Theory of Language Structure*. Continuum, Kobe.

A. Ultsch and H. P. Siemon. 1990. Kohonen's self organizing feature maps for exploratory data analysis. In *Proceedings of International Neural Network Conference,*, pages 305–308, Dordrecht, The Netherlands.

J. N. Williams. 2003. Inducing abstract linguistic representations: Human and connectionist learning of noun classes. In R. van Hout, A Hulk, F. Kuiken, and R. Towell, editors, *The Lexicon-Syntax Interface in Second Language Acquisition*, pages 151–174. John Benjamins, Amsterdam.