On the Role of Contextual Information in the Organization of the Lexical Space

Flavia De Simone, Roberta Presta
Scienza Nuova Research Centre
Suor Orsola Benincasa University
Naples, Italy
Email: flavia.desimone@centroscienzanuova.it, roberta.presta@centroscienzanuova.it

Simona Collina
Suor Orsola Benincasa University
Naples, Italy
Email: simona.collina@unisob.na.it

Robert J. Hartsuiker
University of Ghent
Ghent, Belgium
Email: robert.hartsuiker@ugent.be

***Abstract*—A hotly debated topic in Psycholinguistics concerns the mental representation of words. The current theories about mental lexicon agree on the idea that contextual information plays a crucial role in the organization of lexical knowledge in mind. This paper presents the results of a study we conducted onto two large scale corpora, an Italian one and a Dutch one, aiming at the evaluation of the power of words context in language learning and processing. To this aim, we leverage an outstanding computational model resembling the basic aspects of the internal language formation process. The experiment outcomes show that, starting from the contextual information, it is possible to gain knowledge of even language-specific characteristics.The results corroborate the language-independence of the model we used. We motivated the representativeness of the model also in the light of the current psychological theories.**

Keywords–Distributed Semantic Representation; Contextual Information; Self-Organizing Map; Semantic Map.

I. INTRODUCTION

There is a general agreement that the degree of semantic similarity between two linguistic expressions depends on the similarity of the linguistic contexts in which they appear. While contextual information analysis represents a valuable quantitative method for semantic analysis and lexical resource induction, from a cognitive perspective, it is also supposed to play a causal role in forming general lexical representations.

In distributed models [1], word meaning is typically represented as a vector in a high dimensional space. Semantically similar words tend to cluster in such a space. The more related they are, the closer they are placed. There are two main approaches to generate semantic distributed models: (i) *feature-based* approach and (ii) *corpus-based* approach. To generate a feature-based model, the first step is to ask human subjects to choose a fixed number of words ("features") to describe the considered target words, these words representing the "context". As one of the main drawbacks, they do not work well with closed class of words (such as, for example, determiners and prepositions) and abstract words. On the other hand, corpus-based models are generated precisely in order to start from large scale corpora of words. The hypothesis is that meaning should be constructed based on the statistical co-occurrences of target words in the corpora. As the power of a computational model depends on the capability to capture the mental property of language, a common issue for both the approaches is the limited number of words they considered in their grammatical features. Specifically, both the approaches cluster nouns/objects and verbs/actions but lack in considering the variability through which a speaker describes reality by means for example of action words like "destruction", so far syntactically a noun but semantically describing an event. This is a very important matter of fact, giving that computational models are often taken as a simulation of the cognitive processes involved in using language [2] [3].

In this study, by using a distributed semantics corpus-based approach, we aim at analyzing the lexical-semantic space of words, including action words, in order to specify how these are represented compared to nouns/objects-verbs/actions dimensions and to better specify how the chosen approach is suitable to model language. We consider a distributed semantics corpus-based approach based on the well-known Contextual Self-Organizing Map (SOM) algorithm [4] and analyze the maps resulting from the processing of two large scale corpora, an Italian corpus and a Dutch corpus.

The paper is structured as follows. Related work needed to place this study is presented in Section II. Section III illustrates the algorithm we have leveraged to produce the semantic maps. In Section IV, we delve into the details of the experimental part of this work and comment the obtained results. Conclusion and future work are finally discussed in Section V.

II. RELATED WORK

According to Jackendoff in [5], a word is a long-term memory trace of phonological, syntactic, and semantic information. Particularly, he suggested that this trace *"lists a small chunk of phonology, a small chunk of syntax, and a small chunk of semantics"*. Over the years, this view of the mental lexicon has been enriched by the idea that contextual information plays a crucial role in the organization of lexical knowledge in mind [6]. This hypothesis is supported by empirical evidence: a series of priming experiments showed that verbs [7] and nouns of events [8] prime agents and objects, suggesting that the mental lexicon encodes event-based relations. As suggested by Elman in [9], the assumption that the meaning of a word is never out-of-context is the insight that underlies computational models which derive words representations from statistical

 ISBN: 978-1-61208-390-2

co-occurrences in large-scale corpora. In order to test the richness of contextual information in deriving lexical-semantic representations of words, a computational model of this sort has been tested. The work in [4] has been considered as a reference guide for the practical methodology we have applied in our study. The cited work realizes the corpus-based analysis of an English and of a Chinese corpus by means of the Contextual SOM algorithm (presented in the following) and illustrates as well the Matlab software package we have exploited in the experiments of this work.

III. Contextual Self-Organizing Maps

A self-organizing map [10] is an artificial neural network capable of unsupervised formation of topology-preserving spatial maps capturing input data characteristics. Input data are typically presented to the map in the form of N-dimensional normalized vectors. Each node of the network is characterized by its own coordinates in the $2-$dimensional grid and by a "weight vector" having the same dimension of the input vectors. The SOM is "trained" in order to let node weights progressively resembling, according to a specified similarity metric, the input data. After a sufficient number of training iterations ("epochs"), the node weights in the SOM will have approximated the distribution of the analyzed data by preserving their distance relationships: similar input will be mapped to neighboring nodes, where the mapping consists in the selection of the node having the most similar weight to the considered input. Consequently, thanks to the reduction of the problem dimensionality from N to 2, the map allows for a visual representation of the input distribution and clusters of similar data can be identified by looking ad the corresponding node regions.

A self-organizing *semantic* map is a self-organized map aimed at representing the semantic space of words on a two-dimensional surface [11]. In order to deal with symbolic input, such as words and their contextual information, an ad-hoc pre-processing phase needs to be addressed. As a result of that phase, a distinct N-dimensional unit-length vector will be assigned to each word and the map will be trained on such dataset. The procedure to build such an input dataset is fully detailed and motivated in [11]. We herein report only the main concepts needed to understand the basic rationale behind the performed corpus elaboration. Each input vector is made by two parts: a symbol part, representing a numerical index uniquely associated with the target word, and an attribute part, named the "average context vector" of the target word. As a preliminary step, to each word is assigned a distinct random D-dimensional vector of unit length. For each target word, it is considered its context, i.e. all the words preceding and succeeding the target word in the corpus, together with their co-occurrence values. Then, two D-dimensional vectors are calculated: (i) the weighted average of the random vectors associated to the predecessors, and (ii) the weighted average of the random vectors associated to the successors. The average context vector of the target word is then built as the sequence of the aforementioned vectors, by obtaining this way a $2D$ dimensional vector.

After the training phase, the map is stimulated with the vectors representing the target words: they are built by concatenating the symbol part associated with each word followed by a null attribute part. The "best matching units" (i.e., the nodes with the most similar weight vectors) are then identified and labeled. This process results in the construction of the graphic semantic map, where it is possible to visually observe "similar" words mapped into clustered areas.

IV. Experiments

To the aim of evaluating the power of the contextual information, we have run two experiments onto two different large-scale corpora, an Italian corpus and a Dutch corpus respectively. We have preprocessed such digital corpora and passed them as input to a Contextual SOM in order to analyze the resulting semantic maps and discuss the represented lexical categories. We have leveraged the Matlab software package documented in [4] to run the Contextual SOM algorithm. In the following, details about the analyzed corpora and the adopted procedure are provided. Finally, we discuss the experiments outcomes.

A. Materials

Two large scale corpora have been analyzed. The first corpus is an Italian corpus, extracted by the CoLFIS database [12] including 194624 word tokens with 7065 unique word types. Such a corpus is made by articles from several Italian journals. The second one is a Dutch corpus, precisely an extract of the SUBTLEX-NL corpus [13], consisting of 1047467 tokens with 33962 types. The Dutch corpus is composed by different movie subtitles. Romance and Germanic languages differ in the quantity of syntactic information carried by single words and in the syntactic structure of phrases.

B. Procedure

The experimental process takes different steps, as described in the following:

1) *pre-processing of the corpus;* to run the algorithm, it has been necessary to pre-process the digital corpus by producing two files: a frequency file, in which word types are listed according to their frequency in the corpus, and a second file, in which each word of the corpus is translated into a numerical index corresponding to the number of the word in the frequency list. In the following, only words having an occurrence greater than 5 in the corpus are considered.
2) *vectorization;* in this phase, to each word is associated a normalized $100-$dimensional random vector, as in the preliminary step of the method described in Section III.
3) *generation of the co-occurrence matrix;* the co-occurrence matrix counts the number of times that word *i* precedes or succeeds word *j* in the corpus. The computation of such matrix is needed in order to build for each word the appropriate average context vector part to be used to train the map.
4) *computation of the input vectors and training of the map;* the input dataset is composed as depicted in Section III. We have used a 50x60 map, by inheriting the same parameters settings adopted in [4] for similar analysis. The network has been trained for 200 epochs.
5) *generation of the semantic map;* we have produced a $300-$word map for the Italian corpus and a

 ISBN: 978-1-61208-390-2

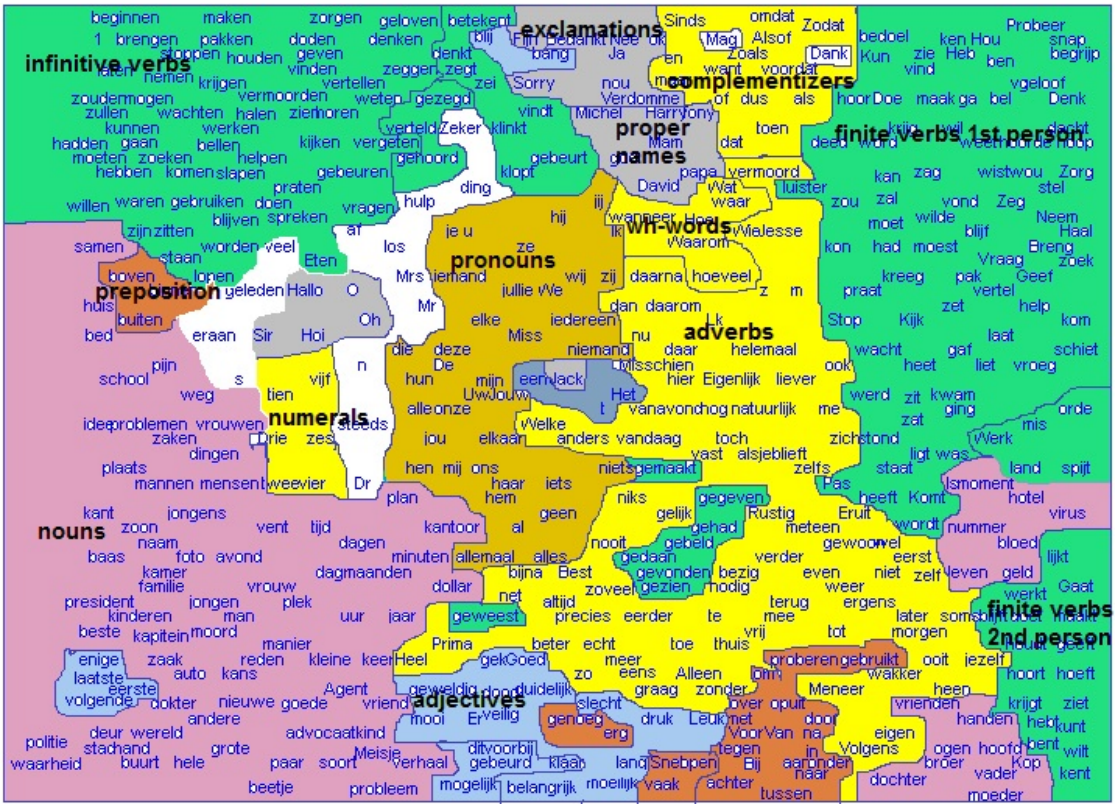

Figure 1. Self-organized map generated from the Dutch corpus

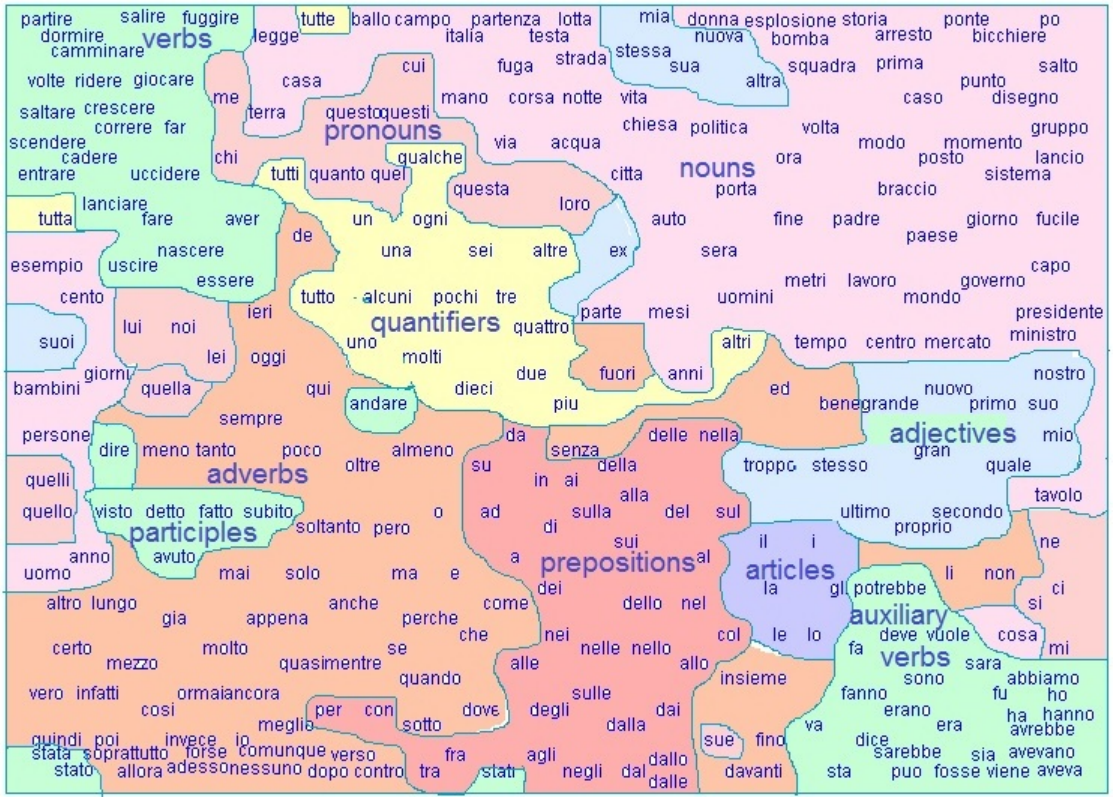

Figure 2. Self-organized map generated from the Italian corpus

500−word map for the Dutch corpus by submitting to the map, respectively, the 300 and 500 most frequent words of the two corpora for the sake of obtaining readable maps.

C. Results

Figure 1 and Figure 2 are the graphic representations of the outcoming semantic maps. We manually draw the boundaries between the semantic clusters in order to make them more clearly visible and to highlight the results. As already mentioned, in the Dutch map just the 500 words with the highest frequencies are presented.

With respect to Figure 1, the first evidence is that the model clusters the major lexical classes, as nouns and verbs, together with closed classes, function words classes as pronouns, preposition and the so-called "wh-words" (why, what, when, where). As suggested in [14], "the closed classes represent a more restricted range of meanings, and the meanings of closed-class words tend to be less detailed and less referential than open-class words." The class of verbs is distributed according to tense (finite vs. infinite), person (1−st and 2−nd), and mood (i.e, all the verbs beginning in capital letters are imperative or interrogative and they are collocated at the margins of the clusters). The class of nouns seems to be organized for gender, with "de-words" in the left part of the map separated from "het-words" in the right part. So far, in the distribution of words in space the lexical-syntactic dimension seems to be more pregnant than the semantic dimension. As to the case of the Italian map, also here we can see that the map clusters the major lexical categories: among the 300 words, it is indeed possible to identify 80 nouns, 24 verbs in the infinitive form, 26 auxiliary verbs, 8 past participles. Unlike in a feature-based model, also closed class as determiners, adverbs, prepositions, and abstract words as "manner" ("modo"), "case" ("caso"), "time" ("tempo") are clustered. Moreover, plural and singular nouns have been clustered separately. All plural nouns, as "men" ("uomini"), "months" ("mesi"), "years" ("anni") are kept together, while all around we can see singular nouns as "center" ("centro"), "work" ("lavoro"), "father" ("padre"). So far, the model is sensitive to semantic and conceptual properties of words. But the network captures also grammatical relations as gender. In the right part of the noun cluster, indeed, there are all masculine words while in the left part there are just feminine words. Also words close in the meaning as "day" ("giorno") and "night" ("notte") are far positioned from each other because they do not share the same grammatical gender. Gender is not only a syntactic property of a word but above all is an arbitrary property. In addition, the map clusters action nouns as "throw" ("lancio"), "jump" ("salto"), "arrest" ("arresto"), "explosion" ("esplosione") with nouns and far from verbs, even if there is a sub-cluster that put them at the margins of network.

V. Conclusion and future work

The resulting maps capture the semantic properties of words: semantically similar words are mapped to spatially close positions. More surprisingly, despite the syntactic differences between the two languages, both maps capture syntactic properties as well: grammatical class, mood and tense for verbs, gender for nouns appear as clusters in the visual representations of the networks. Action nouns (like "throw" ("lancio"), "jump" ("salto"), "arrest" ("arresto"), "explosion" ("esplosione")) are clustered with nouns and far from verbs, even if there is a sub-cluster that put them at the margins of network.

These data corroborate the idea that words are not only vectors of semantic information, but also syntactically rich entities, in line with psycholinguistic evidences. The results obtained add a piece of evidence in the long debate on the lexical organization of words and they support a grammatical class distinction that is independent from semantic [15].

In closing, words seem to carry more information than suggested in [5], by this way posing questions about how this further knowledge is stored in mind. The findings of the experimental campaign herein presented fit the view recently expressed in [16] and [9], which challenged the traditional idea of the mental lexicon as a dictionary: *"Rather than putting word knowledge into a passive storage ..., words might be thought of in the same way that one thinks of other kinds of sensory stimuli: they act directly on mental states ..., it is in the precise nature of their causal effects that the specific properties of words phonological, syntactic, semantic, pragmatic, and so forth are revealed"*.

Further data analysis will be necessary to test the potentiality of the algorithm in the simulation of real categorization processes and to discard the hypothesis that the lexical structure in terms of the order of words in the phrases is the only responsible for the organization of the lexical categories in the map. To do so, we will consider the opportunity to use corpora from other languages, like Hebrew for example, where the phrase structure is more flexible and the meaning of a phrase is independent from words order.

Acknowledgment

This research has been performed with support from the EU ARTEMIS JU project HoliDes (http://www.holides.eu/) SP−8, GA No.: 332933. Any contents herein reflect only the authors' views. The ARTEMIS JU is not liable for any use that may be made of the information contained herein. Finally, the authors would like to special thank Professor Marc Brysbaert for his precious support.

References

[1] A. Lenci, "Distributional Semantics in Linguistic and Cognitive Research," Italian journal of linguistics, vol. 20, no. 1, 2008, pp. 1–31.

[2] P. Tabossi, S. Collina, F. Pizzioli, A. Caporali, and A. Basso, "Speaking of Actions: the Case of CM," Cognitive Neuropsychology, vol. 27(2), 2010, pp. 152–180.

[3] S. Collina, P. Marangolo, and P. Tabossi, "The Role of Argument Structure in the Production of Nouns and Verbs ," Neuropsychologia, vol. 39, no. 11, 2001, pp. 1125 – 1137.

[4] X. Zhao, P. Li, and T. Kohonen, "Contextual Self-Organizing Map: Software for Constructing Semantic Representations," Behavior research methods, vol. 43, no. 1, 2011, pp. 77–88.

[5] R. Jackendoff, Foundations of Language: Brain, Meaning, Grammar, Evolution. Oxford University Press, USA, 2002.

[6] J. J. Van Berkum, C. M. Brown, P. Zwitserlood, V. Kooijman, and P. Hagoort, "Anticipating Upcoming Words in Discourse: Evidence from ERPs and Reading Times," Journal of Experimental Psychology: Learning, Memory, and Cognition, vol. 31, no. 3, 2005, p. 443.

[7] T. R. Ferretti, K. McRae, and A. Hatherell, "Integrating Verbs, Situation Schemas, and Thematic Role Concepts," in Journal of Memory and Language, 2001, pp. 516–547.

 ISBN: 978-1-61208-390-2

[8] M. Hare, M. Jones, C. Thomson, S. Kelly, and K. McRae, "Activating Event Knowledge," Cognition, vol. 111, no. 2, 2009, pp. 151–167.

[9] J. L. Elman, "Lexical Knowledge without a Lexicon," in Mental Lexicon, 2011, pp. 1–33.

[10] T. Kohonen, Ed., Self-organizing Maps. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1997.

[11] H. Ritter and T. Kohonen, "Self-organizing Semantic Maps," Biological Cybernetics, vol. 61, no. 4, 1989, pp. 241–254.

[12] P. M. Bertinetto, C. Burani, A. Laudanna, L. Marconi, D. Ratti, C. Rolando, and A. M. Thornton, "Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)," 2005, http://linguistica.sns.it/CoLFIS/Home.htm, [accessed: 2015-02-12].

[13] E. Keuleers, M. Brysbaert, and B. New, "SUBTLEX-NL: A New Measure for Dutch Word Frequency Based on Film Subtitles," Behavior research methods, vol. 42, no. 3, 2010, pp. 643–650.

[14] M. L. Murphy, Lexical Meaning. Cambridge University Press, 2010.

[15] F. De Simone and S. Collina, "The Picture-Word Interference Paradigm: Grammatical Class Effects in Lexical Production," Journal of Psycholinguistic Research, 2015, submitted.

[16] J. L. Elman, "An Alternative View of the Mental Lexicon," in Trends in Cognitive Sciences, 2004, pp. 301–306.

 ISBN: 978-1-61208-390-2