In the comparison of the two methodologies, the results show that CLUTO outperforms Word2Vec with respect to grouping, using corpora of medium size (20M - 40M). However, the quality of the results does depend on the size of the corpus. At 40M CLUTO already obtains very high quality results (98% PoS coherence and 3.4/4 strength of semantic relationships in the evaluation of the experts) so further increase of the corpus is not likely to show large improvement. On the contrary at 40M Word2Vec still has room for improvement and we expect to narrow the difference between the two methodologies using much larger corpora (1B and above).

In the comparison of the different preprocessing corpora (i.e., raw, lemma, and PoS) in Word2Vec, the results show that lemmatization and PoS tagging largely improve the quality of the groups in both CBOW and Skip-Gram algorithms. This observation is consistent throughout all of the experiments and with respect to all of the evaluation criteria.

The presented comparison opens several lines of future research. First, the evaluation can be extended to bigger corpora, bigger number of vectors, and other languages. Second, the information provided and the suggested criteria for evaluation can be applied to other approaches to DSM and grouping. Finally, the different methodologies and preprocessing options can be evaluated in as part of more complex systems.

Chapter 3

DISCOver: DIStributional Approach Based on Syntactic Dependencies for Discovering COnstructions

M. Antònia Martí, Mariona Taulé, Venelin Kovatchev, and Maria Salamó
University of Barcelona

Published at
Corpus Linguistics and Linguistic Theory, 2019

Abstract One of the goals in Cognitive Linguistics is the automatic identification and analysis of constructions, since they are fundamental linguistic units for understanding language. This article presents DISCOver, an unsupervised methodology for the automatic discovery of lexico-syntactic patterns that can be considered as candidates for constructions. This methodology follows a distributional semantic approach. Concretely, it is based on our proposed pattern-construction hypothesis: those contexts that are relevant to the definition of a cluster of semantically related words tend to be (part of) lexico-syntactic constructions. Our proposal uses Distributional Semantic Models (DSM) for modeling the context taking into account syntactic dependencies. After a clustering process, we linked all those clusters with strong relationships and we use them as a source of information for deriving lexico-syntactic patterns, obtaining a total number of 220,732 candidates from a 100 million token corpus of Spanish. We evaluated the patterns obtained intrinsically, applying statistical association measures and they