Patterns-15. Three experts were asked to classify each pattern as a correct or incorrect construction. The instructions given to them were: a) evaluate whether the pattern is a possible Spanish pattern in your judgement as a native speaker; b) in case of doubt, consult the Google Search engine to check whether it is used by users. Our research questions in this evaluation were: 1) How do the experts evaluate the patterns obtained by DISCOver?; 2) Are experts more likely to accept patterns obtained by DISCOver than random patterns of frequent words?

The average percentage of agreement between the three annotators was 81.67% (see Table 3.10), which is considered high for a semantic evaluation task. The corresponding Fleiss Kappa score is 0.602 with expected agreement of 0.539, which is statistically significant.

Table 3.10 Interannotator agreement test

Annotators (A)	**%Agreement**
A1 and A2	85%
A1 and A3	80.17%
A2 and A3	79.83%
A1, A2 and A3	81.67%

The results of the evaluation are shown in Table 3.11. We use three pattern quality categories. “Strict Positive” includes patterns that were annotated as positive by all three annotators, “Positive” includes patterns that were annotated as positive by at least two annotators and “Negative” groups together patterns that were annotated as positive by one or none of the annotators. The experts accepted the majority of the DISCOver patterns as constructions. At the same time they rejected the majority of the FL-Patterns. We also want to highlight that the percentage of “Strict Positive" patterns is very similar to the percentage of patterns that obtain a high association score. These findings confirm the results that we obtained in the automatic evaluation (See Tables 3.6 and 3.9).

Table 3.11 Expert evaluation

	DISCOver	**FL-Patterns**
Strict Positive	84%	14%
Positive	93%	38%
Negative	7%	62%

3.5 Conclusions and Future Work

This article describes DISCOver, an unsupervised methodology for automatically identifying lexico-syntactic patterns to be considered as constructions. We based this methodology on the pattern-construction hypothesis, which states that the linguistic contexts that are relevant for defining a cluster of semantically related words tend to be (part of) a lexico-syntactic construction.

Following this assumption, we developed a bottom-up language independent methodology to discover lexico-syntactic patterns in corpora. The DSM developed allows us to model the contexts of words (lemmas) taking into account their dependency directions and dependency labels. We applied a clustering process to the resulting matrix to obtain clusters of semantically related lemmas. Then we linked all the clusters that were strongly semantically related and we used them as a source of information for deriving lexico-syntactic patterns, obtaining a total number of 220,732 candidates to be constructions. We evaluated the DISCOver methodology by applying different evaluations. First, the patterns were automatically evaluated using statistical association measures and a different, much larger, corpus. We evaluated whether the patterns we generated obtained a significantly higher association score than statistical chance. We also compared the asociation scores of the DISCOver patterns with a baseline of bigrams. DISCOver obtained better results with respect to both baselines. The patterns obtained by generalization were additionaly evaluated against a baseline of randomly generated patterns. DISCOver significantly outperforms these baselines. Second, the patterns were manually evaluated by expert linguists obtaining good results (89.33%).

This methodology only requires having at one's disposal a medium-sized corpus automatically annotated with POS tags and syntactic dependencies. Therefore, our methodology can be easily replicated with other corpora and other languages. For instance, the DISCOver patterns were also used in a text classification task [Franco-Salvador et al., 2015]. The patterns obtained using our methodology have been compared to other representations (i.e., tf-idf, tf-idf *n*-grams, and enriched graph). The use of these patterns results in an accuracy of 91.69%, which outperfoms the representations based on tf-idf (25.26%), tf-idf *n*-grams (79.26%) and an enriched graph (43.98%), proving to be the best option to represent the content of the corpus.

Furthermore, our methodology increases the descriptive power of the source corpus. First, the lexico-syntactic patterns generated constitute a structured and formalized semantic representation of the corpus. Second, the linking process enlarges the content of the initial data with new relationships not directly present in the corpus (i.e., a total of 167,443 Unattested-Patterns).

The Diana-Araknion-KB[40] can be used as a source of information to derive relevant linguistic information, such as the selection restrictions of verbs, nouns and adjectives; to disambiguate syntactic analysis in order to discard candidate parse trees; to provide a knowledge base of related words with a high degree of association measures for psycholinguistic research; and, to allow for a fine-grained corpus comparision.

The methodology presented and the results obtained, which are available in the Diana-Araknion-KB, open several lines of future research.

First, the Diana-Araknion-KB can be used as a source of information for the development of patterns at different levels of abstraction, in such a way as to obtain a hierarchy of patterns with components belonging to different levels of linguistic knowledge, that is, combining lexical, morpho-syntactic and semantic information. Second, since the same semantic category can be shared by more than one cluster, we could group them into metaclusters containing all the clusters with the same semantic category. Third, a further cluster linking process could be carried out allowing all members of a metacluster to combine with all the target clusters that are related with at least one of the members of the metacluster. Fourth, constructions could be linked in terms of transitivity to obtain larger structures. That is, if cluster A combines with cluster B, and B combines with cluster C, we have the candidate construction: A+B+C. Fifth, the methodology can be used to extract and study patterns in corpora from a specific area, such as the Biomedical domain.

To sum up, we consider that this methodology for discovering constructions outperforms the results of other proposals in the sense that it is fully automatic, language independent, and easily replicable in other corpora and languages. The quality of the results obtained and their wide range of possible applications confirm the DISCOver methodology as a promising line of research and DSMs as a good choice for discovering linguistic knowledge.

[40] Available at `http://clic.ub.edu/corpus/`