In the end, we could obtain a hierarchy of candidates to be considered as different types of constructions, ranging from the most abstract syntactico-semantic constructions combining different semantic classes (12-13) to the most concrete lexico-syntactic constructions (i.e., lemma combinations) (8-11).

## 3.4 Evaluation

In this section we evaluate the quality of the results obtained through the DISCOver methodology: the clusters obtained (see Section 3.4.1) and the lexico-syntactic patterns (see Section 3.4.2).

### 3.4.1 Clustering Evaluation

DISCOver is a methodology for discovering lexico-syntactic patterns.The clusters of semantically related words are a by-product that we obtain as part of the process. Since the focus of this work is the methodology used and the patterns obtained, the evaluation of all possible representation and clustering algorithms is outside the scope of this article. Nevertheless, we prepared a cluster evaluation experiment in order to justify our choice and show that the quality of the obtained vectors and clusters is at least comparable with other state-of-the-art methods. As a baseline, we use standard Word2Vec [Mikolov et al., 2013c], representations with the recommended built-in k-means clustering algorithm. We evaluated the resulting clusters with respect to two criteria: a) the POS purity of each cluster, calculated automatically; and b) the semantic coherence of the lemmas in each cluster, evaluated manually by experts. The criterium applied to determine the coherence of cluster was to check if the words within the cluster held one of the following semantic relations: synonymy, hypernymy or hyponymy.

CLUTO obtained much higher results in terms of both evaluation criteria. The POS coherence of the obtained clusters was 98%, compared to 70% obtained by Word2Vec. Manual evaluation shows that 99% of the clusters obtained by CLUTO were more semantically coherent than the corresponding ones obtained by Word2Vec. These results justify the representations and parameters as adequate for the task and as comparable with the state of the art. Kovatchev et al. [2016] present a more in-depth comparison of the clustering algorithms using corpora of different sizes.

### 3.4.2 Pattern Evaluation

Obtaining high quality lexico-syntactic patterns is the main objective of the DISCOver methodology. In this section, we present two different evaluations of the

obtained patterns: (1) an automatic evaluation, applying statistical association measures; and (2) a manual evaluation by expert linguists[35]. For these evaluations, we used the sum of the patterns of both the 15,000 and 10,000 word configurations.

First, we evaluated the patterns automatically using statistical association measures and a different, much larger, corpus (Diana-Araknion++). In Section 3.3.1, we define two main properties of constructions: 1) Syntactic-semantic coherence and 2) Generalizability. "Syntactic-semantic coherence" entails that the words in each pattern need to be syntactically and semantically related. The "syntactic coherence" of the patterns is not evaluated explicitly, as it is considered to be a by-product of the methodology: all linked clusters from which the patterns are derived have a plausible syntactic relationship and a high connectivity score (see Section 3.3.5.1). However, we need to evaluate the semantic coherence of the patterns, that is, whether there is a semantic relation between the two lemmas. Defining and evaluating "semantic relatedness" is a non-trivial task, which often requires the use of external resources, such as WordNet and BabelNet [Navigli and Ponzetto, 2012]. However, these resources are built considering the paradigmatic relationship between words (such as synonymy, hypernymy, and hyponymy), while we are interested in evaluating syntagmatic relationships.

Evert [2008] and Pecina [2010] discuss the use of association measures for identifying collocations. They define collocations as "the empirical concept of recurrent and predictable word combinations, which are a directly observable property of natural language". In the context of distributional semantics, this definition corresponds to "semantic coherence".

In the DISCOver process, we obtained two qualitatively different types of candidates-to-be-constructions: Attested-Patterns, which are observed in the corpus and Unattested-Patterns, which are obtained as a result of a generalization process that includes clustering, linking and filtering. In order to evaluate the quality of these candidates-to-be-constructions, we formulate two hypotheses and disprove their corresponding null hypotheses.

- **Hypothesis 1**: *The two lemmas in each construction are semantically related.*

  Null hypothesis 1 (henceforth $H_0 1$): The degree of statistical association between the two lemmas in each of the Attested-Patterns, measured in a corpus other than the one they were extracted from, is equal to statistical chance.

[35] An extrinsic evaluation has also been carried out in a text classification task (See Section 3.5).

- **Hypothesis 2**: *Constructions can be generalized and/or derived from other constructions through generalization*. Unattested-Patterns (derived through a generalization process) should be possible language expressions and have the property of semantic coherence.

  Null hypothesis 2.1 (henceforth $H_0$2.1): Unattested-Patterns are not possible language expressions. They cannot appear in a corpus.

  Null hypothesis 2.2 (henceforth $H_0$2.2): If Unattested-Patterns appear in a corpus, they will not have the property of semantic coherence. That is, they will have association scores equal to statistical chance.

In order to prove the two main hypotheses we needed to disprove the three null hypotheses.

For a baseline of $H_0$1, we extracted a list of all bigrams (BI-Patterns) from the original Diana-Araknion corpus. Each bigram contains at least one of the 15,000 most frequent words. We removed all bigrams containing non-content words. All of the Attested-Patterns and the BI-Patterns were found and extracted from the Diana-Araknion 100M token corpus.

For a baseline of $H_0$2.1, we generated patterns by combining frequent lemmas (FL-Patterns): FL-Patterns-15 contain all combinations of the most frequent 15,000 lemmas found in the Diana-Arakion corpus; FL-Patterns-30 contain all combinations in which one lemma is among the 15,000 most frequent lemmas and the other among the 30,000 most frequent ones; FL-Patterns-all contain all word combinations which contain at least one of the 15,000 most frequent lemmas[36].

We use two different statistical methods [Evert, 2008]: simple Mutual Information (MI), which is an effect size measure, and the Z-score (Z-sc), which is an evidence-based measure. Effect-size measures and evidence-based measures are qualitatively different, and for evaluation can be used complementarily. Our final experimental setup includes the following:

- Attested-Patterns, in five different test groups, based on their observed frequency in the Diana-Araknion corpus:
  - Att-Patterns-all with an original frequency of 1 or more
  - Att-Patterns-2 with an original frequency of 2 or more
  - Att-Patterns-3 with an original frequency of 3 or more

[36]The total number of lemmas used in the FL-Patterns (all) is 422,000.