DISCOver: DIStributional approach based on syntactic dependencies for discovering COnstructions

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 were also evaluated qualitativaly by experts. Our results were superior to the baseline in both quality and quantity in all cases. While our experiments have been carried out using a Spanish corpus, this methodology is language independent and only requires a large corpus annotated with the parts of speech and dependencies to be applied.

Keywords: Constructions, Semantics, Distributional Semantic Models

1 Introduction

In cognitive models of language [Croft and Cruse, 2004], a construction is a conventional symbolic unit that involves a pairing of form and meaning that occurs with a certain frequency. Constructions can be of different types depending on their complexity –morphemes, words, compound words, collocates, idioms and more schematic patterns [Goldberg, 1995, 2006]. Cognitive Linguistics assumes the hypothesis that these constructions are learned from usage and stored in

the human memory [Tomasello, 2000], where they are accessed during both the production and comprehension of language. Therefore, constructions are fundamental linguistic units for inferring the structure of language and their identification is crucial for understanding language.

Although a broad range of these linguistic structures have been subjected to linguistic analysis [Nunberg et al., 1994, Wray and Perkins, 2000, Fillmore et al., 2012], we assume that there exist a huge number of constructions that are as yet undiscovered. There are very different approaches to the task of identifying and discovering them, depending on the type of construction we are looking for or dealing with. This fact allows for the use of a wide range of methods and approaches aiming at the treatment of this kind of linguistic units. We distinguish between two different approaches, those that have been guided by previously gathered empirical data[1], and those approaches that apply methods oriented to discovering new constructions from scratch (see Section 2).

Following the latter approach, this article presents DISCOver, an unsupervised methodology for the automatic identification and extraction of lexico-syntactic patterns that are candidates for consideration as constructions (see Section 3). It is based on the Harris distributional hypothesis [Harris, 1954] [2], which states that semantically related words (or other linguistic units) will share the same context.[3] We propose the pattern-construction hypothesis, which states that those contexts that are relevant to the definition of a cluster of semantically related words tend to be (part of) lexico-syntactic constructions. What is new in our hypothesis is that we consider all the contexts that are relevant to define a cluster of semantically related words to be part of a construction. In these approaches, Distributional Space Models (DSMs) are used to represent the semantics of words on the basis of the contexts they share. This is in line with the idea proposed by Landauer et al. [2007], who states that DSMs are plausible models of some aspects of human cognition [Baroni and Lenci, 2010].

In our methodology, the DSM consists of a frequency lemma-context matrix, in which the context is modeled taking into account syntactic dependency relations. Then, we build up clusters of semantically related words that share the same context and link them using the information present in their contexts.

1 See Goldberg [1995].

2 This idea was also developed by Firth [1957] and Wittgenstein [1953].

3 Related hypotheses, such as the extended distributional hypotheses, which states that "patterns that co-occur with similar pairs tend to have similar meanings" [Lin and Pantel, 2001], and latent relation hypotheses [Turney, 2008], which states that "pairs of words that co-occur in similar patterns tend to have similar relations" survived in Turney and Pantel [2010] have also influenced this work.

We automatically calculate a threshold in order to determine which clusters are more strongly related. We filter out those related clusters that do not reach the determined threshold and derive lexico-syntactic patterns that are candidates to be considered as constructions. These candidates are tuples involving two lexical items (lemmas) related both by a dependency direction and a dependency label (examples in (1))[4]:

1. a. accidente_n [>:mod:mortal_a]; accident_n[>:mod:mortal_a]
   b. aterrizar_v [>:dobj:avioneta_n]; to_land[>:dobj:small_plane_a]

The tuples correspond to different kinds of linguistic constructions, ranging from collocates (1a) to (parts of) verbal argument structures (1b). All the lexico-syntactic patterns obtained are instances of one of the syntactic dependencies present in the source corpus. We applied this methodology to the Diana-Araknion corpus, obtaining 220,732 patterns that are good candidates to be constructions[5].

Finally, we evaluated the quality of these patterns in two ways: applying statistical association measures and by manual revision by human experts. The results show significant improvement with respect to several baselines (see Section 4).

Although this method has been applied to the obtention of Spanish constructions, it is language independent and only requires a large corpus annotated with part-of-speech (POS) and syntactic dependencies.

The article is structured as follows. After presenting the related work in Section 2, the methodology applied for obtaining the constructions is described in Section 3. The evaluation of our methodology is presented in Section 4 and, finally, the conclusions and future work are drawn in Section 5.

2 Related Work

The boundaries of what a construction is are fuzzy: constructions can be lexical, syntactic, lexico-syntactic, morphological and can combine different levels of abstraction from concrete forms to abstract categories, including the possibility

4 The symbols '<' and '>' indicate the dependency direction and *mod*, *subj* and *dobj* are dependency labels (where *mod* stands for modifier, and *subj* and *dobj* stand for subject and direct object respectively).

5 All patterns obtained will be made available online.

of using variables, so they cover a wide range of linguistic constructs. For more examples, see Goldberg [2013].

As a consequence, there is no one accepted typology of this kind of linguistic units [Wray and Perkins, 2000]. There is, therefore, a broad field of research in which to explore the characteristics, the limits and the properties of constructions. In this context, an important task is to acquire the maximum amount of empirically grounded data concerning this kind of units. Thus, when approaching the task of attempting to identify the possible constructions that constitute the core of languages, it is difficult to decide what to look at or where to start [Sag et al., 2002]. For this reason, constructions are a challenge for Linguistics and Natural Language Processing (NLP), where we find statistical and symbolic approaches to deal with them.

Several linguistic traditions converge when we are trying to define the diverse form that a construction can take. From one side, there is an (almost total) overlapping between constructions and argument structure [Goldberg, 1995] and diatheses alternations [Levin, 1993]; from another side, in the lexicographic tradition, constructions also overlap with idioms and collocates. In the field of Computational Linguistics, these linguistic units tend to be grouped under the umbrella term MultiWord Expressions (MWE). Baldwin and Kim [2010] define MWE as those lexical items that are decomposable into multiple lexemes and present idiomatic behaviour at some level of linguistic analysis, as a consequence they should be considered as a unit at some level of computational processing. Also in the Computational Linguistics field, Stefanowitsch and Gries [2003] propose the term "collostruction" to refer to the wide range of complex linguistic units as defined in theoretical proposals of Cognitive Grammar. In our approach we consider as constructions those syntactic units consisting of two or more lexical items with internal semantic coherence. These constructions are compositional and appear with a frequency higher than expected.

From the NLP perspective, most approaches for dealing with constructions tend to apply methods that use previously defined empirical knowledge to find instances and variants of specific types of constructions in corpora. This approach allows us to obtain preidentified units and their variations at different degrees of complexity, but does not allow for the identification of as yet unidentified constructions. In order to discover new knowledge, we need an open and flexible method that give us usable and interpretable results. We organised this overview taking into consideration those approaches that try to find or discover constructions.

A frequent approach to gathering empirical data about constructions using NLP techniques is to look for well-known, highly conventionalized and previously defined constructions (see the works of Hwang et al. [2010], Muischnek and

Sajkan [2009], Kesselmeier et al. [2009], O'Donnell and Ellis [2010], Duffield et al. [2010]).

Very tied to Construction Grammar theory and in the framework of the methodologies based on statistical metrics, it is worth noting the works of Stefanowitsch and Gries [2003], Stefanowitsch and Gries [2008], and Gries et al. [2005]. Their research always focuses on specific types of constructions, on the analysis of their variants and on the degree of entrenchment between their elements. Gries and Ellis [2015] summarize different statistical measures applied to the analysis of constructions and evaluate their linguistic interpretation and impact.

From the perspective of methods oriented to the discovery of new constructions, we should distinguish between those approaches that include some kind of linguistic filtering of the type of constructions to be dealt with and those that do not apply any kind of restriction. All these methods are strongly grounded on statistical measures: in Evert [2008] and Pecina [2010] there is an exhaustive summary and criticism of statistical measures that calculate the degree of association between words.[6]

Looking for ways to identify potential collocations in corpora using statistical measures, Bartsch [2004] explores certain types of collocations involving verbs of verbal communication. Her approach is semiautomatic and involves a manual revision of the results. We also highlight the work of Pecina [2010], based on fully statistical methods. However, supervised machine learning requires annotated data, which creates a bottleneck in the absence of large corpora annotated for collocation extraction. A solution to this problem is presented by Dubremetz and Nivre [2014] who propose the use of the MWEtoolkit [Ramisch et al., 2010] to automatically extract candidates that fit a certain POS pattern. See also the work of Forsberg et al. [2014], Farahmand and Martins [2014], Tutubalina [2015].

From a different perspective, based on the calculation of n-grams, we also consider the results of the StringNet project [Wible and Tsao, 2010], a knowledge base (KB) which contains candidates to be constructions. In this case, no filters are applied to the lexico-syntactic patterns obtained. As a result, StringNet is a lexicogrammatical KB automatically extracted from the British National Corpus (BNC)[7] consisting of a massive archive of hybrid n-grams of co-occurring combinations of POS tags, lexemes and specific word forms.

6 The works referred to this section use the term collocate in a very weak sense, roughly equivalent to what is known as MWE in NLP.

7 www.natcorp.ox.ac.uk

We also want to highlight the approaches that use syntactic information for obtaining constructions, such as the work of Zuidema [2006], Sangati and van Cranenburgh [2015], based on the framework of Tree Substitution Grammar (TSG).

Harris distributional hypothesis has a great acceptance in the treatment of linguistic semantics to overcome traditional symbolic representations. Relying on this hypothesis, Gamallo et al. [2005] developed an unsupervised strategy to acquire syntactico-semantic restrictions for nouns, verbs and adjectives from partially parsed corpora. Although the resulting data could be used for deriving lexico-syntactic patterns their objective was to capture semantic generalizations, both for the predicates and their arguments.

Currently, there is an increasing interest in the use of distributional models for representing semantics, such as DSMs [Turney and Pantel, 2010, Baroni, 2013] or word embeddings [Mikolov et al., 2013]. These models derive word-representations in an unsupervised way from very large corpora. All of them rely on co-ocurrence patterns but differ in the way they reduce dimensionality. As pointed out in Murphy et al. [2012], the representations they derive from corpora are lacking in cognitive plausibility, with exceptions such as those defined in Baroni et al. [2010]. Our proposal shares with these authors the same semantic approach (distributional hypothesis), because we consider that these models are a good option in which to frame our methodology. In concrete, we used DSMs because they are highly linguistically interpretable and allow us to modelize the context, a key point in our methodology.

DSMs have been applied successfully in linguistic research [Shutova et al., 2010], in different NLP tasks and applications [Baroni and Lenci, 2010] and, especially, in tasks related with measuring different kinds of semantic similarity between words [Turney and Pantel, 2010]. Like us, Shutova et al. [2017] use distributional clustering techniques, though they use DSMs to investigate how to find metaphorical expressions. Recently, DSMs have been extended to phrases and sentences by means of composition operations deriving meaning representations for phrases and sentences from their parts (see Baroni [2013] and Mitchell and Lapata [2010] for an overview). Nevertheless, DSMs have rarely focused on the discovery of constructions. In this line, it is worth noting the papers presented in the shared task of the Workshop on Distributional Semantics and Compositionality [Biemann and Giesbrecht, 2011]. This workshop focused on the extraction of non-compositional phrases from large corpora by applying distributional models that assign a graded compositional score to a phrase. This score denotes the extent to which compositionality holds for a given expression. The participants applied a variety of approaches that can be classified into lexical association measures and Word Space Models. It is also worth noting

that approaches based on Word Space Models performed slightly better than methods relying solely on statistical association measures.

In the next section, we describe in depth the DISCOver methodology that we developed to discover lexico-syntactic constructions.

3 Methodology for discovering constructions

Following a distributional semantic approach, we developed an unsupervised bottom-up method for obtaining the lexico-syntactic patterns that can be considered candidates for constructions. This method uses a medium-sized corpus (100 million tokens) to obtain the distributional properties of words and to stablish similarity relations among them from their contexts. The representation of the contexts is based on syntactic dependencies.

Figure 1 depicts the five main steps involved in obtaining the lexico-syntactic patterns. Briefly, the first step is the linguistic processing of the Diana-Araknion corpus (See Section 3.2). In the next step, a DSM matrix is constructed with the frequencies of the lemmas in each one of the contexts (see Section 3.3). Step 3 focuses on clustering semantically related lemmas, that is, those lemmas that share a set of contexts (see Section 3.4). In the fourth step, we applied a generalization process by linking all clusters taking into account the information contained in the contexts and then filtering only those links that mantain the strongest relationships (See Section 3.5). Finally, we generate the lexico-syntactic patterns to be considered as candidates to be constructions from the related clusters selected in the previous step (See Section 3.6).

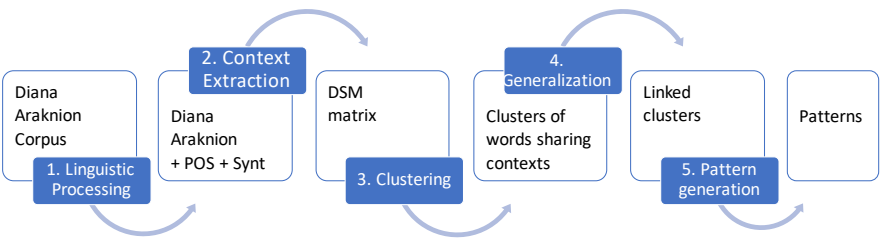

Fig. 1: Main steps in DISCOver methodology

3.1 Description of the task

Our methodology is based on the pattern-construction hypothesis, which states that those contexts that are relevant to the definition of a cluster of semantically related words tend to be (part of) lexico-syntactic constructions. In our experiments, "lexico-syntactic constructions" are patterns in the form of [*lemma, dependency_direction (dep_dir), dependency_label (dep_lab), context_lemma*] (for instance, [despeinar_v, >: dobj, cabellera_n][8]). *Dependency_label* is a type of syntactic relation between *lemma* and *context_lemma*, while *dependency_direction* is the direction of the *dependency_label*. To be considered candidates to be constructions patterns must have the following properties:

- *Syntactic-semantic coherence*: We expect the two lemmas in each pattern candidate to be syntactically and semantically related.
- *Generalizability*: The patterns can be generalized and/or derived from other patterns through generalization.

Based on these properties of constructions and the initial pattern-construction hypothesis, the main aims of the DISCOver methodology are the following:

1. To identify the contexts that are relevant for the definition of a cluster of semantically related words. Each of these contexts is part of a pattern candidate to be construction attested in the corpus (henceforth Attested-Patterns).
2. To use the previous contexts in a generalization process in order to identify unseen, but possible candidates to be constructions (henceforth Unattested-Patterns).

As a result we obtain two sets of qualitatively different patterns that are candidates to be constructions: attested and unattested patterns. We then proceed to evaluate the internal syntactic-semantic coherence of these patterns.

3.2 The Corpus

As shown in Figure 1, corpus creation is the first step in the process of obtaining lexico-syntactic patterns. Specifically, we built the Diana-Araknion[9] corpus, a

8 [to_tussle_v, >: dobj, one's_hair_n]

9 The Diana-Araknion corpus will be made freely available online.

Spanish corpus which consists of approximately 100 million tokens[10] (corresponding to 3 million sentences) gathered mainly from the Spanish Wikipedia (2009), literary works and texts from Spanish parliamentary discussions, news reports, news agency documents, and Spanish Royal Family speeches.

The corpus was automatically tokenized and linguistically processed with POS and lemma tagging, and syntactic dependency parsing. We used the Spanish analyzers available in the Freeling[11] open source language-processing library [Padró and Stanilovsky, 2012].

For the purpose of evaluation, we built Diana-Araknion++, a new corpus gathered from web-pages in Spanish. It includes Wikipedia 2015, articles from online newspapers, speeches from the European Parliament, university articles and sites from the Spanish webspace. This corpus was automatically tokenized and POS tagged and consists of 600M tokens.

3.3 Matrix

To generate the frequency matrix (see Step 2 in Figure 1), we used only the 15,000 most frequent lemmas extracted from the Diana-Araknion corpus including nouns (N), verbs (V), adjectives (A) and adverbs (R). We modeled the context in which the words occur giving rise to a *lemma-dep* matrix. This matrix corresponds to the type of *word-context* matrix defined in Turney and Pantel [2010] and in Baroni and Lenci [2010]. In the *lemma-dep* matrix, the context is based on parsed texts in which both dependency directions and dependency labels are taken into account. Each context is a triple of [*dependency_direction, dependency_label, context_lemma_POS*].

In what follows, we introduce how this lemma-context matrix is formally represented (see Section 3.3.1) and then we describe the matrix in more detail (see Section 3.3.2).

3.3.1 Formalization of the lemma-context matrix

Our DSM consists of a lemma-context PPMI matrix X with n_r rows and n_c columns. Note that each row vector i corresponds to a lemma, each column j corresponds to a co-occurrence context, and each cell in X has a numerical weighted value, x_{ij}. This weighted value is the result of applying Positive Pointwise Mutual

10 Concretely, the Diana-Araknion has 93,987,098 tokens and 1,321,174 types.
11 http://nlp.lsi.upc.edu/freeling.

Information (PPMI) [Niwa and Nitta, 1994] to a lemma-context frequency matrix F with size $n_r \times n_c$. Each element in this matrix, f_{ij}, is computed as the number of occurrences of lemma i in context j in the whole corpus. Lapesa and Evert [2014] perform a large-scale evaluation of different co-occurrence DSM models over various tasks. They show that term weighting through association scores significantly improves the performance of the DSM model.

3.3.2 *Lemma-dep* matrix

The matrix proposed in this work is a lemma-context matrix, hereafter *lemma-dep* matrix, based on syntactic dependencies[12]. In this matrix, the context j of a lemma i is a context word k (*context_lemma*) directly related by a dependency direction (*dep_dir*) and a dependency label (*dep_lab*) to the lemma i. The words of the lemma i belong to the following POS: *N*, *V*, *A* and *R*. Each lemma is assigned its corresponding POS. Therefore, in the matrix, context j contains three elements as defined in 1:

$$context = [dep\_dir : dep\_lab : context\_lemma] \quad (1)$$

where:

- *dep_dir*: has two possible values '<' or '>', indicating the direction of the dependency.
- *dep_lab*: indicates the dependency label of the lemma i and context_lemma k. The possible values are {*subj, dobj, iobj, creg, cpred, atr, cc, cag, spec, sp* and *mod*}. In the case of dependencies between a preposition and a noun, adjective or verb, the dependency label is labeled by the same preposition and its corresponding *dep_lab*, that is, *dobj*, *iobj*, *creg*, *cag*, *sp* or/and *cc*.
- *context_lemma* is the lemma of the context word k with its corresponding POS, which can be *N*, *V*, *A*, *R*, preposition*(P)*, number*(Z)* and date*(W)*. In the case of proper nouns, they are replaced by the *pn_n* (proper noun) POS.

Figure 2 shows an example of a dependency parsed sentence from which, for instance, three different contexts of the noun lemma *barba_n* [13] are generated:

12 We used the Spanish syntactico-semantic analyzer Treeler to analyse the Diana-Araknion corpus: http://devel.cpl.upc.edu/treeler .

13 'beard'

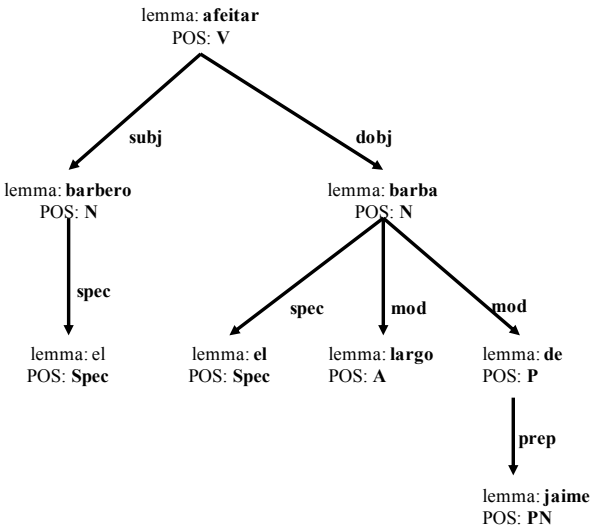

Fig. 2: Dependency parsed sentence: *El barbero afeita la larga barba de Jaime* ('The barber shaves off James's long beard')

[<:dobj:afeitar_v], [>:mod:largo_a] and [>:de_sp:pn_n][14]. These contexts are represented in the *lemma-dep* matrix.

In [<:dobj:afeitar_v], '<' indicates that the verb *afeitar_v* [15] maintains a parent dependency relation with *barba_n*, *dobj* indicates that *barba_n* is the direct object of *afeitar_v*, and *afeitar_v* is the context word (lemma k) related to *barba_n* (lemma i). In [>:mod:largo], *mod* indicates that the adjective *largo_a* [16] is a modifier of *barba_n*, and in [>:de_sp:pn_n] the proper noun (*Jaime* in Figure 2) is replaced by the *pn_n* POS tag[17].

For each context obtained from the dependency structure, three different dependency contexts are generated: one that makes all the elements of the context explicit, that is, the *dep_dir*, *dep_lab* and *context_lemma* (for example, [<:dobj:afeitar_v]); another in which the *dep_lab* is generalized by the

14 This context is the result of substituting the proper name "Jaime" by "pn_n".

15 'to shave off'

16 'long'

17 Since the POS tagger does not distinguish between subclasses of proper names (person, organization, place, etc.), the grouping of all with the *pn_n* tag gives better results. We used proper nouns in the *context_lemma* configuration, but not as words in the lemma i. Similarly, stopwords are not included in lemma i.

variable 'oth' (for example, [<:oth:afeitar_v])[18] and, finally, one context that generalizes the *context_lemma* by substituting it for the variable '*' (for example, [<:dobj:*_v])[19]. The three lemmas represented in example (2) do not share any context, therefore they could not be semantically related in our model. Instead, applying the generalization of contexts, we obtained a relationship between lemma_1 and lemma_2 in example (3), and between lemma_1 and lemma_3 in example (4). In example (3), the *dep_lab* is generalized, whereas in example (4) the *context_lemma* is generalized.

2. lemma_1 [<: $subj : robar\_v$[20]]
   lemma_2 [<: $dobj : robar\_v$]
   lemma_3 [<: $subj : hurtar\_v$[21]]

3. lemma_1 [<: $oth : robar\_v$]
   lemma_2 [<: $oth : robar\_v$]
   lemma_3 [<: $oth : hurtar\_v$]

4. lemma_1 [<: $subj : *\_v$]
   lemma_2 [<: $dobj : *\_v$]
   lemma_3 [<: $subj : *\_v$]

In this way, the generalization of contexts allows us to take into account contexts that are similar (they share two, but not all of the elements, of their context), but not identical. Therefore, we can distinguish between those lemmas that share the same or similar context, and those that have a completly different context. By adding these contexts that are similar but not identical we add new knowledge, that is, knowledge not directly present in the corpus. This new knowledge is used to generate the Unattested-Patterns.

3.4 Clustering

Once we described the X matrix, we proceeded to the third step detailed in Figure 1 that is devoted to the clustering of this matrix. The motivation of

18 The tag 'oth' (*other*) means that the dependency label is not specified.
19 The symbol '*_v' means that a verb occurs in this position, but we do not specify which one it is.
20 'to_rob'
21 'to_steal'

the clustering process is to find, for each lemma in the matrix, all semantically related words (lemmas). This will allow us to create new Unattested-Patterns after the linking and filtering cluster processes. To perform this clustering step, we used the CLUTO toolkit [Karypis, 2003][22], which is used to cluster a collection of objects (in our case, lemmas) into a predetermined number of clusters labeled k. We applied a methodology based on Caliński and Harabasz [1974] and using cosine similarity and CLUTO's $\mathcal{H}_2$ metric to estimate the optimal amount of clusters.

We experimented with a number of different clustering configurations. The variables we took into account were: a) the number of most frequent lemmas, with the 10,000 to 15,000 most frequent lemmas giving the best results; b) the inclusion of proper nouns or their substitution for their POS; and c) considering the lemmas with and without their POS.

We evaluated the results of these configurations manually and opted for 15,000 lemmas with proper nouns grouped according to their POS tag (pn_n) and with the POS tag assigned to the lemmas. This configuration gave an optimal k of 1,500 clusters applying the Caliński and Harabasz [1974] method and the $\mathcal{H}_2$ metric.

The inclusion of POS improves the internal consistency of the clusters. Since the POS tagger does not distinguish between subclasses of proper names (person, organization, place, etc.), grouping them according to the *pn_n* tag also gives better results. Regarding the number of lemmas, all results obtained using between 10,000 and 15,000 lemmas gave satisfactory results. The choice of the number of lemmas determines the number and the content of the clusters. In all cases, the quality of clusters obtained was acceptable. We consider a cluster as acceptable when all or almost all words contained in it share one of the following relations: synonymy, hypernymy, or hyponymy. This would allow for the use of one or more configurations for the obtention of the final lexico-syntactic patterns (see Section 3.6).

Using CLUTO with the selected configuration, we obtained a set of clusters $C = \{c_i : 1 \leq i \leq k\}$ from matrix X. Formally, the content of each cluster $c_i \in C$ is defined in 2, where le is a set of related lemmas and ctx is a set of contexts. Each lemma_pos only belongs to one cluster (i.e., it can only be defined in one le), whereas a context_lemma can be in several contexts (ctx) of different clusters.

22 We use VCLUSTER program provided in the toolkit, which computes the clustering using one of five different approaches. Four of these approaches are partitional, whereas the fifth approach is agglomerative.

$$c_i =< le, ctx > \quad (2)$$

Formally, a context (called *context_cluster*) in *ctx* is described as follows:

$$context\_cluster =< [dep\_dir : dep\_lab : context\_lemma], score > \quad (3)$$

where $dep\_dir, dep\_lab, context\_lemma$ corresponds to the definition of a context as shown in Section 3.3.2. The *score* is the sum of the different scores given by CLUTO[23].

For example, Table 1[24] describes the lemmas, *le*, and the most scored contexts, *ctx*, in cluster number 421_n (one of the clusters obtained in the corpus analyzed).

Tab. 1: Example of a real cluster (421_n) in the Diana-Araknion corpus in Spanish

Cluster: 421_n			
Lemmas (c_{421_le})	**barba_n, bigote_n, cabellera_n, cabello_n, ceja_n, crin_n, melena_n, mostacho_n, patilla_n, pelaje_n, pelo_n, perilla_n, vello_n**		
Contexts (c_{421_ctx})	**[< : dobj : erizar_v],11**	**[< : oth : erizar_v],11**	**[< : oth : rizar_v],10**
	[< : subj : erizar_v],10	**[> : mod : espeso_a],9**	**[> : oth : espeso_a],9**
	[> : mod : negro_a],7	**[< : oth : negro_a],5**	**[> : mod : gris_a],8**
	[< : dobj : rizar_v],8	**[> : oth : gris_a],7**	**[< : oth : pelo_n],6**
	[> : mod : rubio_a],7	**[> : mod : barba_n],7**	**[< : oth : atusar_v],7**
	[> : mod : largo_a],4	**[> : oth : rubio_a],6**	**[< : mod : pelo_n],2**
	[> : mod : rojizo_a],4	**[> : oth : rojizo_a],6**	**[> : oth : largo_a],3**
	[< : oth : bigote_n],3	**[> : mod : blanco_a],3**	**[> : mod : cano_a],5**
	[> : mod : hirsuto_a],5	**[> : oth : hirsuto_a],2**	**[> : oth : largo_a],3**
	[> : oth : negro_a],2	**[> : mod : rojizo_a],2**	

23 The sum of the twenty-five most descriptive and discriminative scores given automatically by CLUTO.

24 Examples, tables, translations to English as well as clusters will be made available online.

3.4.1 Results of the clustering process

Following our configuration, we obtained a total of 1,500 clusters in the clustering process (k=1500). It is worth noting that the clusters are highly morpho-syntactically and semantically cohesive.

The clusters contain lemmas belonging mostly to the same POS. It is worth mentioning that more than half of the clusters are nouns (54.20%), followed by verbs (25.80%) and adjectives (16.67%). Clusters of adverbs make up only 3.33% of the total.

Clusters contain relevant implicit information, in the sense that their lemmas belong to well-defined semantic categories, often at a very fine-grained level. For instance, we obtained clusters of adjectives with a *Positive Polarity* (5) and with a *Negative Polarity* (6)[24]. These results encourage us to tag all the clusters with one or more semantic labels. That will enrich the obtained patterns.

5. {c_{111}, *Positive_Polarity* adjectives: admirable_a, asombroso_a, genial_a...}[25]

6. {c_{38}, *Negative_Polarity* adjectives: atroz_a, aterrador_a, espantoso_a...}
[26]

3.5 Linking and filtering clusters

The process of linking clusters (see Step 4 in Figure 1) is based on the set of clusters and contexts obtained using Cluto. The processes of linking clusters and pattern generation detailed in Section 3.6 are the core steps of the DISCOver methodology. The process of linking clusters uses the set of the twenty-five highest scored contexts in each cluster. According to our pattern-construction hypothesis (see Section 3.1), the goal of the linking of clusters is to establish the relationships between clusters using their contexts, as defined in (3), obtaining as a result a matrix of all possible contextual relations between clusters (see Section 3.5.1). Next, we apply a filtering process in order to select strongly related links taking into account different criteria (see Section 3.5.2).

25 'admirable, amazing, great'
26 'atrocious, scary, frightening'

3.5.1 Linking clusters and building the matrix of related clusters

Basically, the aim of the cluster linking process is to establish the relationships between clusters and to store them in a matrix, *R_clusters*, with k rows and k columns. The k-value corresponds to the number of clusters obtained in the clustering step.

For building the matrix, for each origin cluster (x) each *dep_dir* and *dep_lab* of the *context_cluster* (defined in Equation 3) are converted into a *contextual_relation* (see Equation 4), while the *context_lemma* of the *context_cluster* is used to locate the cluster (y) in which it occurs. We obtain as a result a matrix, *R_clusters*, in which clusters are related according to a set of contextual relations stored in a *relation_set*. The sum of the scores of the *context_clusters* in 3 are added together in a matrix, *R_scores*. The *R_scores* matrix is later used in the process for determining filtering thresholds.

$$contextual\_relation =< dep\_dir, dep\_lab > \tag{4}$$

For the contextual relation, defined in 4, *dep_dir* and *dep_lab* are the dependency direction and the dependency label defined in a context of cluster i related to cluster j. Note that the *relation_set* of a cluster in itself is empty as $R\_clusters[i][i] = \emptyset$ and $R\_clusters[i][j] \neq R\_clusters[j][i]$.

Following the example of cluster 421_n, described in Table 1, the result of the cluster linking process for this particular cluster ($i = 421\_n$) is shown in Table 2[27]. The first column in this table shows the related clusters, j, the second column shows the relation_type that relates cluster 421_n to the related clusters j (i.e. STRONG, SEMI or WEAK, See 3.5.2), and finally the last column describes the lemmas in the related clusters.

3.5.2 Filtering related clusters

In the *R_clusters* matrix, not all contextual relationships between clusters are accepted since they have a low *R_scores*. For this reason, we established two criteria to automatically determine which relationships will be maintained and which ones are filtered out in the pattern generation process. For each criterion only those relations higher than a predetermined score value will be considered. The criteria are the following:

27 For the sake of simplicity, the contexts are not included in the Table 2 and we only show a relation of each type.

Tab. 2: Some examples of cluster linking process in cluster i=421_n (described in Table 1).

Related clusters(j)	Relation_ type	Lemmas ($c_{j.le}$, where c_j refers to the related cluster, j)
1223_a	strong	azabache_a, bermejo_a, cano_a, canoso_a, hirsuto_a, lacio_a, lustroso_a, ondulante_a, sedoso_a...
932_v	Semi	afeitar_v, atusar_v, cepillar_v, empolvar_v, enguantar_v, peinar_v, rasurar_v...
405_n	weak	contario_n, final_n, largo_n, menudo_n...

- **Criterion 1**: For each pair of clusters i and j, we take into account those relations that in each of their directions (i.e., $R\_scores[i][j]$ or $R\_scores[j][i]$) have a score above a minimum predetermined value, that is, $threshold_1$. This $threshold_1$ is automatically determined by finding a score value that allows for the grouping of 30% of the clusters. The relations that fulfill criterion 1 are called STRONG relations.
- **Criterion 2**: For each pair of clusters i and j, we take into account those relations in which the sum of scores in both directions (i.e., $R\_scores[i][j] + R\_scores[j][i]$) is higher than a predetermined value, that is, $threshold_2$, which is determined by finding a value that allows for the grouping of 50% of the clusters. The relations that fulfill criterion 2 are called SEMI relations.

Considering the example of cluster 421_n, the result of the filtering process is that, out of the three clusters linked to cluster 421_n in our example[24] (1223_a, 932_v, and 405_n), we will only select those with STRONG and SEMI relations, that is, 1223_a, and 932_v. Those labelled as WEAK (e.g., 405_n shown in Table 2) are filtered out because they do not reach the established thresholds.

3.6 Pattern generation

Once the process for automatically linking and filtering clusters was carried out, we proceeded to generate the lexico-syntactic patterns to be considered as candidates for constructions (see Step 5 in Figure 1). Each generated pattern is defined as follows:

$$pattern =< lemma_i, dep\_dir, dep\_lab, lemma_j > \quad (5)$$

where $lemma_i$ and $lemma_j$ are the lemmas contained in the related clusters (i and j), dep_dir and dep_lab are the dependency direction and the dependency label between the related clusters. So, there is a pattern for each $lemma_i$ and $lemma_j$ pair.

As we mentioned in Section 3.4, all possible configurations using between 10,000 and 15,000 lemmas gave acceptable related clusters. In order to increase the number of patterns generated we carried out the same process with a configuration using 10,000 lemmas. We combined the patterns obtained using the 10,000 and 15,000 lemmas together and removed those that were shared by both configurations. In Tables 3, 4 and 5, we show the number of resulting clusters and patterns, after removing the overlapping patterns, for the two configurations.

Tab. 3: Distribution of the number of related and unrelated clusters and their percentage

	10,000 lemmas	15,000 lemmas
Relation	**Clusters (%)**	**Clusters (%)**
Strong	**441 (31.50%)**	**461 (30.73%)**
Semi	**339 (24.21%)**	**396 (26.40%)**
Total	**780 (55.71%)**	**857 (57.13%)**
Weak	**589 (42.07%)**	**636 (42.40%)**
Unrelated	**31 (2.21%)**	**7 (0.47%)**

As shown in Table 3 (second and third columns), more than 55% of the linked clusters maintain STRONG and SEMI relationships, whereas only the 2.68% of the clusters remain unrelated. Table 4 (second and third columns) shows the distribution of linked clusters by POS in both configurations.

Tab. 4: Distribution of the number of related clusters and their percentage by POS

	10,000 lemmas	15,000 lemmas
POS	**Clusters (%)**	**Clusters (%)**
N	**415 (53.21%)**	**464 (54.14%)**
V	**197 (25.26%)**	**182 (12.24%)**
A	**142 (18.21%)**	**173 (20.19%)**
R	**26 (3.30%)**	**38 (4.43%)**
Total	**780 (100%)**	**857 (100%)**

The total number of lexico-syntactic patterns obtained from the two configurations of clusters (780 and 857 STRONG and SEMI related clusters) is 237,444. For the purpose of pattern generation, STRONG and SEMI clusters have been treated equally. From these patterns, we removed 16,712 patterns, those that were present in both sets of generated patterns, given as a result the total number of 220,732 patterns (See Table 5).

Tab. 5: Distribution of the generated patterns

Lemmas	Attested-Patterns	Unattested-Patterns	Total
10,000	**23,980**	**48,147**	**72,127**
15,000	**37,840**	**127,477**	**165,317**
10,000 + 15,000	**61,820**	**175,624**	**237,444**
Overlapping	**8,531**	**8,181**	**16,712**
Sum (no overlap)	**53,289**	**167,443**	**220,732**

The DISCOver methodology allows for the generation of patterns that actually occur in the corpus (Attested-Patterns), but also of lexico-syntactic patterns that are not present in the corpus but which are highly plausible in Spanish (Unattested-Patterns), since the components of the clusters are closely semantically related. As a result, we are able to enlarge the descriptive power of the source corpus. Among the patterns we generated, 61,820 were Attested-Patterns, that is, patterns that are present in the source corpus, and 175,624 were Unattested-Patterns, that is, new patterns (see Table 5).

Retaking the example of cluster 421_n and its related clusters we obtain patterns such as those shown in (7)[28]:

7. <bigote$_{c_421}$ <:dobj: cepillar$_{c_932_v}$>
   <melena$_{c_421}$ <:dobj: alisar$_{c_1267_v}$>
   <pelaje$_{c_421}$ >:mod: sedoso$_{c_1223_a}$>
   <perilla$_{c_421}$ >:mod: gris$_{c_149_a}$>

All of these patterns are Unattested-Patterns, that is, they do not occur in the Diana-Araknion corpus but are generated applying our methodology and are

28 <moustache$_{c_421}$ <:dobj: to_brush$_{c_932_v}$>; <mane$_{c_421}$ <:dobj: to_smooth$_{c_1267_v}$>; <fur$_{c_421}$ >:mod: silky$_{c_1223_a}$>; <goetee$_{c_421}$ >:mod: grey$_{c_149_a}$>

perfectly acceptable in Spanish. These patterns would not have been extracted using, for example, a *n*-gram based method or plain statistical methods.

It is worth noting the high degree of semantic cohesion between the lemmas of the same cluster and between the lemmas of the related clusters ((8)[29], (9)[30], (10)[31] and (11)[32]).

8. <accidente $_{c\_470}$ <:dobj causar$_{c_560}$>
   <fuego $_{c\_470}$ <:dobj evitar$_{c_560}$>
   <siniestro $_{c\_470}$ <:dobj producir$_{c_560}$>
9. <accidente $_{c\_470}$ <:subj desencadenar$_{c_560}$>
   <destrozo $_{c\_470}$ <:subj producir$_{c_560}$>
   <incendio $_{c\_470}$ <:subj originar$_{c_560}$>
10. <canciller $_{c\_70}$ >:mod argentino$_{c_1}$>
    <embajador $_{c\_70}$ >:mod belga$_{c_1}$>
    <mandatario $_{c\_70}$ >:mod chileno$_{c_1}$>
11. <cantante $_{c\_155}$ >:mod belga$_{c_1}$>
    <compositor $_{c\_155}$ >:mod canadiense$_{c_1}$>
    <pianista $_{c\_155}$ >:mod estadounidense$_{c_1}$>

This strong cohesion allows for a semantic annotation of the clusters to obtain more abstract syntactico-semantic constructions that combine semantic categories (12) and (13). The semantic labels associated with each cluster have been manually added, taking into account the WordNet upper ontologies.

12. <*Event*-n$_{c_470}$ <:dobj *Causative*-v $_{c\_560}$>
    <*Event*-n $_{c\_470}$ <:subj *Causative*-v $_{c\_560}$>
13. <*Person/Politician*-n $_{c\_70}$ >:mod *Nationality*-a $_{c\_1}$>
    <*Person/Musician*-n $_{c\_155}$ >:mod *Nationality*-a $_{c\_1}$>

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

29 <accident$_{c_470}$ <:dobj to_cause$_{c_560}$>; <fire$_{c_470}$ <:dobj to_avoid$_{c_560}$>; <sinister$_{c_470}$ <:dobj to_produce$_{c_560}$>.

30 <accident$_{c_470}$ <:subj to_trigger$_{c_560}$>, <ravage$_{c_470}$ <:subj to_produce$_{c_560}$>.

31 <chancellor$_{c_70}$ >:mod argentinian$_{c_1}$>; <ambassador$_{c_70}$ >:mod belgian$_{c_1}$>; <representative$_{c_70}$ >:mod chilian$_{c_1}$>

32 <singer $_{c\_155}$ >:mod belgian$_{c_1}$>; <song-writer$_{c_155}$ >:mod canadian$_{c_1}$>; <pianist$_{c_155}$ >:mod american$_{c_1}$>

concrete lexico-syntactic constructions (i.e., lemma combinations) (8-11), which are instances of the abstract constructions.

4 Evaluation

In this section we evaluate the quality of the results obtained through the DISCOver methodology: the clusters obtained (see Section 4.1) and the lexico-syntactic patterns (see Section 4.2).

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., 2013], representations with the recommended built-in k-means clustering algorithm. We evaluate 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 criteria applied had been to check if the words in a cluster hold 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.

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[33]. 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 3.1, 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 3.5.1). 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 [Miller, 1995] 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.*

33 An extrinsic evaluation has also been carried out in a text classification task (See Section 5).

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.

– **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[34].

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:

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

- 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
  - Att-Patterns-4 with an original frequency of 4 or more
  - Att-Patterns-5 with an original frequency of 5 or more
- BI-Patterns, with an original frequency of 5 or more[35]
- Unattested-Patterns
- FL-Patterns-15, FL-Patterns-30, FL-Patterns-all

Evaluating $H_0$1:

We calculated the MI and Z-sc association scores of the two words in each of the Attested-Patterns and BI-Patterns in the Diana-Araknion++ 600M token corpus. The association score was calculated based on the sentential co-occurrence of the two words. Patterns that co-occurred less than 5 times obtained a score of 0. First, we compared the obtained association with standard thresholds, representing statistical chance: 0, 0.5, and 1 for MI; 0, 1.96, and 3.29 for Z-sc. Second, we compared the average association score of the Attested Patterns with those of the BI-Patterns.

Table 6 shows what percentage of the Attested-Patterns in each group obtains scores higher than statistical chance. Overall, the majority of the Attested-Patterns outperform the statistical chance baseline. The results are consistent for both the measures and their thresholds, even though they measure the association in a qualitatively different manner. It is important to note that filtering out the Attested-Patterns with a frequency of 1 significantly improves the results. We believe this factor should be taken into consideration in future experiments.

Tab. 6: Association score of Attested-Patterns compared with statistical chance

	MI			Z-sc		
Patterns	**>0**	**>0.5**	**>1**	**>0**	**>1.96**	**>3.29**
Att-Patterns-5	**85%**	**83%**	**80%**	**85%**	**83%**	**82%**
Att-Patterns-4	**84%**	**82%**	**79%**	**84%**	**82%**	**80%**
Att-Patterns-3	**82%**	**80%**	**77%**	**82%**	**80%**	**78%**
Att-Patterns-2	**78%**	**76%**	**72%**	**78%**	**76%**	**73%**
Att-Patterns-all	**68%**	**66%**	**62%**	**68%**	**65%**	**62%**

35 5,285 of the BI-patterns coincide with Attested-Patterns.

As a complementary evaluation, we directly compared the association scores of the Attested-Patterns with those of the BI-Patterns. Table 7 shows the average association scores for the two types of patterns[36]. The Attested-Patterns have a much higher degree of association than the BI-Patterns. In the case of MI, the Attested-Patterns obtain scores more than two times higher than the BI-Patterns. In the case of Z-sc, the Attested-Patterns obtain scores between 30% and 100% higher than the BI-Patterns.

Tab. 7: Average association score of Attested-Patterns and BI-patterns

Patterns	Average MI	Average Z-sc
Attested-Patterns-5	**3.90**	**52**
Attested-Patterns-4	**3.86**	**49**
Attested-Patterns-3	**3.80**	**46**
Attested-Patterns-2	**3.70**	**42**
Attested-Patterns-all	**3.50**	**35**
BI-Patterns	**1.72**	**27**

The obtained results disprove $H_0 1$ and confirm Hypothesis 1. That is, we can conclude that the Attested-Patterns are semantically coherent.

Evaluating $H_0$2.1:

We checked how many of the Unattested-Patterns were present in Diana-Araknion++. As a baseline we used the FL-Patterns. Both Unattested-Patterns and FL-Patterns are not directly obtained, but are rather a result of generalization and generation using different methodologies. For each group, we calculated the percentage of the patterns that appear once and the percentage of the patterns that appear at least five times. Table 8 shows the results obtained.

Unattested-Patterns appear much more frequently than the patterns generated by simply combining frequent lemmas. 56% of the Unattested-Patterns were observed in Diana-Araknion++. This is more than double the observance rate of the FL-Patterns-15 and five times higher than for FL-Patterns-30. 24% of the Unattested-Patterns appear in Diana-Araknion++ with a frequency of 5 or more. This is almost three times higher than FL-Patterns-15 and six times higher than FL-Patterns-30. The results of FL-Patterns-all are much lower, showing that unfiltered pattern generation is not effective. Unattested-Patterns are linguistic patterns given that they appear in a corpus with a much higher probability than

36 The average is calculated as a simple average of all patterns of the corresponding type.

Tab. 8: Occurrence of Unttested-Patterns and FL-Patterns

Patterns	Occurred Once	Occurred Five Times
Unattested-Patterns	**54%**	**24%**
FL-Patterns-15	**24%**	**9%**
FL-Patterns-30	**11%**	**4%**
FL-Patterns-all	**4%**	**0.6%**

patterns generated using a simpler frequency based methodology. These results disprove $H_0$2.1.

Evaluating $H_0$2.2:

We calculated the association score (MI and Z-sc) between the lemmas in each of the Unattested-Patterns that occurred at least 5 times[37] in Diana-Araknion++. We compared the scores with the same thresholds we used when evaluating $H_0$1. Table 9 shows the percentage of patterns with a score higher than the statistical chance thresholds.

Tab. 9: Association scores of Unttested-Patterns

	MI			Z-sc		
Patterns	**>0**	**>0.5**	**>1**	**>0**	**>1.96**	**>3.29**
Unattested-Patterns	**93%**	**86%**	**76%**	**93%**	**80%**	**70%**

The observed degree of association is very high. Over 90% of the observed Unattested-Patterns obtained a positive association score with respect to both measures. When comparing them with the statistical chance thresholds, the obtained results are similar to those obtained by Attested-Patterns in $H_0$1. The Unattested-Patterns, when observed in a different corpus, are semantically coherent. This disproves $H_0$2.2.

In conclusion, the automated statistical evaluation of the patterns obtained by DISCOver shows that: (1) Attested-Patterns are semantically coherent, as they outperform two baselines: statistical chance thresholds and BI-Patterns. These results disprove $H_0$1.; (2) A significant percentage (56%) of the Unattested-Patterns can be found in Diana-Araknion++, which is much higher than the occurrence

37 Calculating this score for patterns with lower frequency is unreliable due to the low-frequency bias in some of the measures.

of FL-Patterns. These results disprove $H_0 2.1$; (3) Whenever Unattested-Patterns occur in Diana-Araknion++, the statistical association between the lemmas in the patterns is much higher than the statistical chance baseline. This disproves $H_0 2.2$.

As we have disproved all 3 of the null hypotheses, we can conclude that the patterns obtained by the DISCOver methodology have both properties of constructions: syntactic and semantic coherence and generalizability. Therefore they are good candidates-to-be-constructions.

We also performed a manual evaluation of the lexico-syntactic patterns. This complementary validation reinforces the results obtained in the two statistical evaluations. We prepared a dataset of 600 patterns for the manual evaluation: 300 patterns obtained by applying the DISCOver methodology (the patterns were randomly selected from all Attested and Unattested Patterns) and 300 of the FL-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 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.

Tab. 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 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 6 and 9).

Tab. 11: Expert evaluation

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

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 association 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[38] 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,

38 Available online.

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.

References

Timothy Baldwin and Su Nam Kim. Multiword Expressions. *Handbook of natural language processing*, 2:267–92, 2010.

Marco Baroni. Composition in distributional semantics. *Language and Linguistics Compass*, 7 (10):511–22, 2013.

Marco Baroni and Alessandro Lenci. Distributional Memory: A General Framework for Corpus-based Semantics. *Computational Linguistics*, 36(4):673–721, 2010. ISSN 0891-2017.

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2):222–54, 2010.

S. Bartsch. *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunter Narr Verlag, 2004.

Chris Biemann and Eugenie Giesbrecht. Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the workshop on distributional semantics and compositionality*, pages 21–8. Association for Computational Linguistics, 2011.

T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation*, 3(1):1–27, 1974.

W. Croft and D.A. Cruse. *Cognitive Linguistics*. Cambridge Textbooks in Linguistics. Cambridge University Press, 2004. ISBN 9780521667708.

Marie Dubremetz and Joakim Nivre. Extraction of Nominal Multiword Expressions in French. *EACL 2014*, page 72, 2014.

Cecily Jill Duffield, Jena D Hwang, and Laura A Michaelis. Identifying assertions in text and discourse: the presentational relative clause construction. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 17–24. Association for Computational Linguistics, 2010.

Stefan Evert. Corpora and collocations. *Corpus Linguistics. An International Handbook*, 2: 223–33, 2008.

Meghdad Farahmand and Ronaldo Martins. A Supervised Model for Extraction of Multiword Expressions Based on Statistical Context Features. *EACL 2014*, page 10, 2014.

Charles J Fillmore, Russell Lee-Goldman, and Russell Rhodes. The Framenet constructicon. *Sign-based Construction Grammar. CSLI, Stanford, CA*, 2012.

Markus Forsberg, Richard Johansson, Linnéa Bäckström, Lars Borin, Benjamin Lyngfelt, Joel Olofsson, and Julia Prentice. From construction candidates to constructicon entries. an

experiment using semi-automatic methods for identifying constructions in corpora. *Constructions and Frames*, 6(1):114–35, 2014. ISSN 1876-1933.

Marc Franco-Salvador, Rangel Francisco, Rosso Paolo, Taulé Mariona, and Martí M. Antónia. Language variety identification using distributed representations of words and documents. In *Proceedings of the 6th International Conference of CLEF on Experimental IR meets Multilinguality, Multimodality and Interaction*, Lectures Notes in Computer Science. Springer Verlag, 2015.

Pablo Gamallo, Alexandre Agustini, and Gabriel P Lopes. Clustering syntactic positions with similar semantic requirements. *Computational Linguistics*, 31(1):107–146, 2005.

A. E. Goldberg. *Constructions: A Construction Grammar Approach to Argument Structure*. Cognitive Theory of Language and Culture. University of Chicago Press, 1995. ISBN 9780226300863.

A. E. Goldberg. *Constructions at work*. Oxford University Press, 2006.

Adele E Goldberg. Argument structure constructions versus lexical rules or derivational verb templates. *Mind & Language*, 28(4):435–65, 2013.

Stefan Th. Gries and Nich C. Ellis. Statistical measures for usage-based linguistics. *Language Learning*, (65):1–28, 2015.

Stefan Th. Gries, Beate Hampe, and Doris Schönefeld. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, (16):635–76, 2005.

Zellig Harris. Distributional structure. *Word*, 10(23):146–62, 1954.

Jena D Hwang, Rodney D Nielsen, and Martha Palmer. Towards a domain independent semantics: Enhancing semantic representation with construction grammar. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 1–8. Association for Computational Linguistics, 2010.

George Karypis. CLUTO - a clustering toolkit. Technical report, University of Minnesota, 2003.

K. Kesselmeier, T. Kiss, A. Müller, C. Roch, T. Stadteld, and J. Strunk. Mining for preposition-noun constructions in german. In *Workshop on Extracting and Using Constructions in Natural Language Processing*, NODALIDA 2009, 2009.

Venelin Kovatchev, Maria Salamó, and M. Antònia Martí. Comparing distributional semantics models for identifying groups of semantically related words. *Procesamiento del Lenguaje Natural*, 57:109–116, 2016.

T.K. Landauer, D.S. McNamara, S. Dennis, and W. Kintsch. *Handbook of Latent Semantic Analysis*. University of Colorado Institute of Cognitive Science Series. Lawrence Erlbaum Associates, 2007. ISBN 9780805854183.

Gabriella Lapesa and Stefan Evert. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *TACL*, 2:531–545, 2014. URL `https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/457`.

Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.

Dekang Lin and Patrick Pantel. Dirt@ sbt@ discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–8. ACM, 2001.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–51, 2013.

George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995. ISSN 0001-0782.

Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1439, 2010.

K. Muischnek and H. Sajkan. Using collocation-finding methods to extract constructions and estimate their productivity. In *Workshop on Extracting and Using Constructions in Natural Language Processing*, NODALIDA 2009, 2009.

Brian Murphy, Partha Pratim Talukdar, and Tom M Mitchell. Learning effective and interpretable semantic models using non-negative sparse embedding. In *COLING*, pages 1933–50, 2012.

Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, December 2012. ISSN 0004-3702.

Yoshiki Niwa and Yoshihiko Nitta. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th Conference on Computational Linguistics*, volume 1 of *COLING '94*, pages 304–309, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. Idioms. *Language*, pages 491–538, 1994.

Matthew Brook O'Donnell and Nick Ellis. Towards an inventory of english verb argument constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, EUCCL '10, pages 9–16, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Lluìs Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *LREC*, pages 2473–9. European Language Resources Association (ELRA), 2012. ISBN 978-2-9517408-7-7.

Pavel Pecina. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44:137–58, 2010. ISSN 1574-020X.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. Multiword expressions in the wild?: the mwetoolkit comes in handy. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 57–60. Association for Computational Linguistics, 2010.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer Berlin Heidelberg, 2002.

Federico Sangati and Andreas van Cranenburgh. Multiword expression identification with recurring tree fragments and association measures. In *Proceedings of NAACL-HLT*, pages 10–18, 2015.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002–1010. Association for Computational Linguistics, 2010.

Ekaterina Shutova, Lin Sun, Elkin Darío Gutiérrez, Patricia Lichtenstein, and Srini Narayanan. Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. *Computational Linguistics*, 43(1):71–123, 2017.

Anatol Stefanowitsch and Stefan Th. Gries. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8(2):209 – 43, 2003.

Anatol Stefanowitsch and Stefan Th. Gries. Corpora and grammar. *Corpus Linguistics*, 2008.

Michael Tomasello. First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, 11(1-2):61–82, 2000.

Peter D Turney. The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research (JAIR)*, 33:615–55, 2008.

Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research (JAIR)*, 37(1):141–88, January 2010. ISSN 1076-9757.

Elena Tutubalina. Clustering-based approach to multiword expression extraction and ranking. In *Proceedings of NAACL-HLT*, pages 39–43, 2015.

David Wible and Nai-Lung Tsao. StringNet As a Computational Resource for Discovering and Investigating Linguistic Constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, EUCCL '10, pages 25–31, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Alison Wray and Mick Perkins. The functions of formulaic language: an integrated model. *Language and Communication*, 20(1):1–28, 2000.

Willem Zuidema. What are the productive units of natural language grammar?: a dop approach to the automatic identification of constructions. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 29–36. Association for Computational Linguistics, 2006.