3.3.1 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][10]). *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.

3.3.2 The Corpus

As shown in Figure 3.1, corpus creation is the first step in the process of obtaining lexico-syntactic patterns. Specifically, we built the Diana-Araknion[11] corpus, a Spanish corpus which consists of approximately 100 million tokens[12] (corresponding to 3 million sentences) gathered mainly from the Spanish Wikipedia

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

[11] All corpora are available at `http://clic.ub.edu/corpus/` or per-request

[12]Concretely, the Diana-Araknion has 93,987,098 tokens and 1,321,174 types.

(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[13] 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 Spanish 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.

3.3.3 Matrix

To generate the frequency matrix (see Step 2 in Figure 3.1), 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 3.3.3.1) and then we describe the matrix in more detail (see Section 3.3.3.2).

3.3.3.1 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 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.

[13] `http://nlp.lsi.upc.edu/freeling`.

3.3.3.2 *Lemma-Dep* Matrix

The matrix proposed in this work is a lemma-context matrix, hereafter *lemma-dep* matrix, based on syntactic dependencies[14]. 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 3.1:

$$context = [dep\_dir : dep\_lab : context\_lemma] \tag{3.1}$$

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* [15] are generated: [$<$:dobj:afeitar_v], [$>$:mod:largo_a] and [$>$:de_sp:pn_n][16]. These contexts are represented in the *lemma-dep* matrix.

In [$<$:dobj:afeitar_v], '$<$' indicates that the verb *afeitar_v* [17] 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* [18] is a modifier of *barba_n*, and in [$>$:de_sp:pn_n] the proper noun (*Jaime* in Figure 2) is

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

[15] 'beard'

[16] This context is the result of substituting the proper name "Jaime" by "pn_n".

[17] 'to shave off'

[18] 'long'