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.[8]

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

[8]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.

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)[9] consisting of a massive archive of hybrid *n*-grams of co-occurring combinations of POS tags, lexemes and specific word forms.

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., 2013c]. 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

[9] www.natcorp.ox.ac.uk

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.

3.3 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 3.1 depicts the five main steps involved in obtaining the lexico-syntactic patterns, the processess involved, and the input and output of each process. Briefly, the first step is the linguistic processing of the Diana-Araknion corpus (See Section 3.3.2). In the next step, a DSM matrix is constructed with the frequencies of the lemmas in each one of the contexts (see Section 3.3.3). Step 3 focuses on clustering semantically related lemmas, that is, those lemmas that share a set of contexts (see Section 3.3.4). 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 3.3.5). 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 3.3.6).

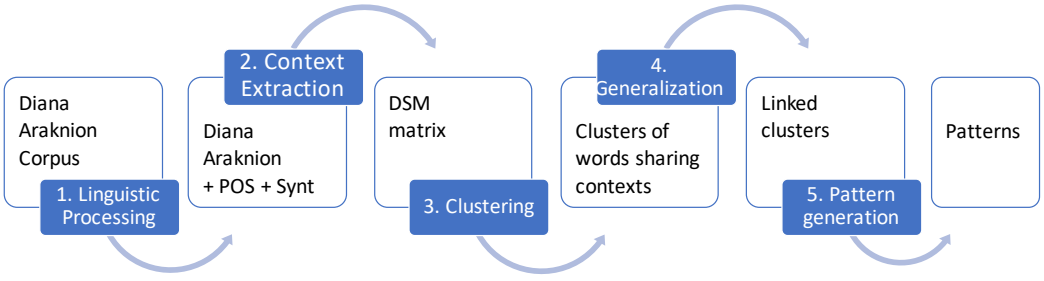

Figure 3.1: Main steps in DISCOver methodology