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

3.1 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 3.2).

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 3.3). 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]

[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

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. 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][5]
   b. aterrizar_v [>:dobj:avioneta_n][6]

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[7].

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 3.4).

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.

---

patterns tend to have similar relations" survived in Turney and Pantel [2010] have also influenced this work.

[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]accident_n[>:mod:mortal_a]

[6]to_land[>:dobj:small_plane_a]

[7]All patterns obtained are available at `http://clic.ub.edu/corpus/`

The article is structured as follows. After presenting the related work in Section 3.2, the methodology applied for obtaining the constructions is described in Section 3.3. The evaluation of our methodology is presented in Section 3.4 and, finally, the conclusions and future work are drawn in Section 3.5.

3.2 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 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