2020.

Classification of abstract and concrete nouns using the adjective as a predictor variable

Valentina Ravest Córdova[195]
[vale.ravest.c@gmail.com]
Pontificia Universidad Católica de Valparaíso, Chile

Different Spanish grammars, such as the *Nueva gramática de la lengua española (*Real Academia Española, 2009) and the *Gramática descriptiva de la lengua española (*Bosque and Demonte, 1999), have mentioned the problems that exist in classifying nouns in the categories of abstract and concrete. Among these are the difficulty in defining the very notions of concrete and abstract, the lack of morphological criteria, and the lack of agreement among the different grammarians on their characteristics and definitions. These works use their own criteria to define what they consider to be an abstract and a concrete noun, although they recognize that there are obstacles to classifying them correctly because the definitions often encounter cases of nouns that defy them. This has generated difficulties within the field of lexicology, since there are still insufficient criteria to classify nouns in these categories. Until today, they were only classified based on semantic criteria, but these cannot be considered sufficient. For example, the idea that abstract nouns designate entities separate from the things themselves, such as characteristics or properties referring to form (Bello, 1847), which fails to encompass the singularities of the different nouns, making classification difficult.

Based on this lack of objective criteria for classification, this paper aims to establish a method of automatic, empirical and objective classification of nouns in the categories of abstract and concrete using as a predictive variable the adjectives with which they co-occur in their context of occurrence. In order to achieve this objective, it is hypothesized that the adjective is the predictive element that allows this classification. The proposed classification method was divided into five parts. The first consisted of the constitution of a set of experimental data. On the one hand, the selection of a random sample of 262 nouns in English and 248 in Spanish, divided into abstract and concrete, subdivided into the subcategories of activities, doctrines and feelings, within the abstract ones, and of weapons, fruits and minerals within the concrete ones, from the dictionaries *Diccionario de uso del español de América y España, DUEAE,* in its CD-ROM version (Battaner, 2003),

[195] Degree in Language and Literature, with a major in Applied Linguistics, from the Pontificia Universidad Católica de Valparaíso. Technical Assistant in the FONDECYT project "Automatic translation of taxonomies of discourse markers from multilingual corpora", in the following areas: Research in the field of natural language processing. Grammatical and linguistic analysis. Development of lexical classification systems. As a continuation of my undergraduate thesis, development of an automatic classification system for abstract and concrete nouns.

and the *Diccionario de la lengua española* (RAE, 2014). On the other hand, a working textual corpus was created from a random sample of 5000 contexts of occurrence of each of these nouns in the EsTenTen corpus (Renau and Kilgarriff, 2013). In the second part of the process, we proceeded to extract the adjectives that co-occur with the nouns in the corpus, using a Perl script that takes advantage of the morphosyntactic tagging already included in the corpus, in order to identify the adjectives. In the third part, the data of nouns and adjectives were arranged in a matrix, in which nouns were placed in the rows and adjectives in the columns, assigning binary values to each cell to indicate the co-occurrence of adjectives and nouns. Fourth, the agglomerative clustering statistical technique was applied using the R program, version 3.5.1 (Ihaka and Gentleman, 2018), which plots the results in the form of a dendrogram where the similarity of nouns as a function of shared adjectives is reflected. Finally, using the 50 adjectives with the highest frequency of occurrence for concrete nouns and for abstract nouns, the classifier was constructed, in which it is possible to enter any noun without categorization and classify it immediately.

The classifier was able to meet its goal correctly with both abstract and concrete nouns, in English and Spanish. Abstract nouns, in both languages, were classified with 0.81 accuracy, while concrete nouns were classified with 0.89 accuracy.

These results suggest the possibility of classifying abstract and concrete nouns through mechanical procedures and objective criteria, which raises new possibilities for lexicography and promotes new avenues of research in lexicology.

A web page offers a demo of the classifier and other project data. The demo accepts any noun as input and as output offers a classification as concrete or abstract. The address is: http://www.tecling.com/chungungo