A Small System Storing Spanish Collocations*

Igor A. Bolshakov, Sabino Miranda-Jiménez

Center for Computing Research
National Polytechnic Institute, Mexico City, Mexico
igor@cic.ipn.mx, sabino@correo.cic.ipn.mx

Abstract. Collocations are defined as syntactically connected and semantically compatible pairs of content components, like Spanish *prestar atención* 'give attention', *presidente del país* 'president of the country', *país grande* 'large country' or *muy bien* 'very well'. The collocation databases are important for numerous applications of computational linguistics. A small system storing Spanish collocations is reported. Each collocation is accessible from both its components while querying. As compared with its Russian prototype, its size is now spare and the collocation types are limited to a few most productive ones. However, the available relation of the hyponym/hyperonym type between the systemic dictionary entries permits to infer some additional collocations at runtime. The actual statistics of the system is given.

1 Introduction

We call collocation a syntactically connected and semantically compatible pair of content words,[1] like Spanish *abandonarse a la desesperación* 'to fall into despair', *vocero del gobierno* 'press secretary of the government', *trabajar fuertemente* 'to work hard' or *muy grande* 'very large', where the two principal components are underlined. Just as in [1, 2], we treat as collocations word combinations of any stability, including full idioms and free word combinations.

As it is extensively explained in [1], collocation databases (CDBs) in any language can be used in language learning and in numerous tasks of text processing, among them lexicalized parsing, word sense disambiguation, malapropism detection & correction, automatic translation, revealing text cohesion, and segmentation of texts to paragraphs. Thus the creation of CDBs is topical for any language. However, the situation with CDBs poorly corresponds to the needs mentioned. For example, there is an excellent English collocations dictionary [10] but no good English CDB, so that all applications of CDBs different from word learning are left unprovided. Among other European languages only Russian has now a large CDB [1, 2]. For Spanish, there are neither dictionaries nor CDBs.

This paper reports on a small system called **CrossLexica-Esp** whose principal part is just a Spanish collocation database. In distinction from common dictionaries, each

* Work done under partial support of Mexican Government (CONACyT, SNI) and CGPI-IPN.

[1] Unlike some other authors, who use this term to refer to mere co-occurrences, e.g., [3]