Getting One's First Million... Collocations[*]

Igor A. Bolshakov

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

Abstract. Many-long-years-of experience in creating a very large database of Russian collocations is summarized. The *collocations* here described are syntactically connected and semantically compatible pairs of content components—single or multi-words. We begin from a synopsis of various applications of collocation databases (CDBs). Then we describe the main features of collocation components, syntactic types of collocations, and links of other nature between their components that amplify the applicability of the enclosing systems. All of the above-mentioned characterizes the CrossLexica system created for Russian but with a universal structure suited for other languages. The statistics of CrossLexica is given and discussed. It now contains more that a million collocations and more than a million WordNet-like links.

1 Introduction

The term *collocation* is widely used in both theoretical and computational linguistics but without a generally agreed definition. We define a *collocation* as a syntactically connected and semantically compatible pair of content words, like *full-length dress, well expressed, to briefly expose, to pick up the knife* or *to listen to the radio*, where the collocation components (hereafter ***collocatives***) are underlined.

Syntactical connectedness is understood as in dependency grammars [21] and it is in no way merely a co-occurrence of the collocatives in a short span of a text [1, 27]. The head collocative syntactically governs the dependent collocative, being adjoined to it directly or through an auxiliary word (usually a preposition). Sequentially, the collocatives can be at any distance from each other in a sentence, but are nearby in the dependency tree.

The stability of a collocation can be determined as rather high mutual information of its collocatives quantitatively evaluated within a large text corpus [20; 17, 18].

The idiomaticity of collocations is an intuitively graduated measure as illustrated below with the English verb *see*. In the phrase *to see a jungle* both collocatives have their direct meaning; in *to see a doctor* 'to visit a doctor for consultations or treatment' *doctor* has its direct meaning, while the meaning of *see* is supplemented by other elements; in *to see the reason* 'to have or understand the reason' the direct meaning of *see* is not traced at all; the phrase *to see light at the end of the tunnel* 'to

[*] Work done under partial support of Micra Inc., USA (1993–1994), Soros Foundation, Russia (1995), and Mexican Government (CONACyT, SNI, CGPI–IPN) (since 1999). Thanks to Dr. Patrick Cassidy for his valuable suggestions.