Web-Based Measurements of Intra-Collocational Cohesion in Oxford Collocations Dictionary*

Igor A. Bolshakov[1] and Sofia N. Galicia-Haro[2]

[1] Center for Computing Research (CIC),
National Polytechnic Institute (IPN), Mexico City, Mexico
igor@cic.ipn.mx

[2] Faculty of Sciences
National Autonomous University of Mexico (UNAM)
Mexico City, Mexico
sngh@fciencias.unam.mx

Abstract. Cohesion between components of collocations is already acknowledged measurable by means of the Web, and cohesion measurements are used for some applications and extraction of new collocations. Taking a specific cohesion criterion *SCI*, we performed massive evaluations of collocate cohesion in Oxford Collocations Dictionary. For three groups of modificative collocations (adjective—noun, adverb—adjective, and adverb— verb) *SCI* distributions proved to be one-peaked and compact, with rather close mean values and standard deviations. Thus we suggest a reliable numeric criterion for extraction of collocations from the Web.

Introduction

Let us transitorily define collocations as syntactically linked and semantically compatible pairs of content words. They are rather specific for each language, so electronic collocation databases compiled beforehand are needed for many applications (text editing, foreign language learning, syntactic analysis, word sense disambiguation, detection & correction of errors etc.). Though the tools for automatic collocation extraction are being developed more than 15 years [7], large electronic collocation databases do not exist to the date for well-known languages.

The Web is acknowledged now as a huge corpus for automatic collocation extraction and this it is supposed possible with a numeric criterion of coherence between collocates [2, 6]. An application of corpus-oriented criteria to Web statistics theoretically is not grounded, since the Web counts occurrences and co-occurrences in pages, not words. Since a theory allowing to recalculate the numbers of relevant pages to the numbers of words occurred in them does not exist nowadays, we are free to use both formulas recommended for corpuses, re-conceptualizing page numbers as word numbers, and any analogous formulas operating by numbers of relevant pages.

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