Exploratory Collocation Extraction

Stefan Evert[1], Brigitte Krenn[2]

[1] University of Osnabrück – Germany

[2] OFAI, Vienna – Austria

Keywords: collocations, co-occurrence, association measures, collocation extraction, evaluation

1. Approaches to lexical collocations

Lexical collocations are a fuzzy phenomenon for which linguistic theory has not yet found a satisfactory explanation. At the same time, they are important both for our understanding of the structure of human language and for many applications such as lexicography and natural language processing. Corpus-based studies of collocations as well as collocation extraction tools have been influenced by two basic views:

(a) An empirical notion of lexical collocations as recurrent combinations of words, which has developed from the ideas of Firth (1957). Proponents of this view are typically interested in studying sets of collocations extracted from a corpus. Since Firth was mostly concerned with co-occurrences that express semantic and conceptual relations (such as *dark – night* and *milk – cow*), collocation extraction techniques from this approach are usually based on spans of a few tokens around the instances of a given keyword and ignore the syntactic structure of sentences.

(b) A phraseological notion of collocations as pre-constructed syntactic units (Grossmann & Tutin 2003) or lexically determined elements in syntactic constructions (e.g. Choueka 1988), which is prevalent in the lexicographic treatment of word combinations and in most of computational linguistics. In this view, collocations are characterized by their semantic, syntactic and distributional irregularity (cf. Manning & Schütze 1999:184), i.e. by intrinsic properties of the word combinations rather than their distribution in corpora. The goal of such approaches is to extract a specific type of collocation – defined according to intensional linguistic criteria – with high precision and recall. In order to improve accuracy, it is common to consider only words that co-occur in a specific syntactic relation (e.g. verb-object), based on a (partial) syntactic analysis of the corpus text.

Views (a) and (b) approach the phenomenon of lexical collocations from opposite directions. Approach (a) starts from recurrent word combinations, defined in terms of empirical distributional criteria, and aims to describe and understand their observed linguistic properties.

Approach (b), on the other hand, starts from a theoretical analysis of lexical collocations (often resulting in a taxonomy of subtypes). Its goal is to develop methods to extract the desired type of collocation with high accuracy. This situation has led to much controversy (if not open hostilities) between adherents of the two views, which culminated in the recent publication of Hausmann (2004). However, a closer look reveals that both approaches face essentially the same problem: the difficulty of giving their object of study a precise definition.

For (a), it is necessary to operationalize the notion of "recurrence". Most researchers rely a mathematical criterion, namely that of significant statistical association, which may seem to be an objective and indisputable definition at first sight. However, statistical association can be quantified in many different ways, neither of which is obviously right or wrong (cf. the long-standing debate in mathematical statistics reported by Yates (1984)). In addition, methods for establishing the significance of an observed association face various mathematical problems that can often be traced back to characteristic properties of language data such as Zipf's law and the untenability of independence assumptions (cf. Evert 2004). As a result, a wide range of equally plausible association measures will extract entirely different sets of recurrent word combinations from a given corpus.

Approach (b) seems to have an advantage in the form of a clearer goal to guide the choice of a suitable association measure. Here, the problem lies in the theoretical analysis, namely the lack of a precise definition of the phenomenon of lexical collocations and a clear delineation of the relevant subtypes. The classifications that have been developed up to now – figurative expressions, support verb constructions, idioms, proverbs, etc. – are problematic for various reasons. While they often function well for a core set of instances, they invariably leave open a grey area of word combinations that exhibit properties of several different classes of collocations. An example is the distinction between support verb constructions and figurative expressions in German, which can be operationalized fairly well (cf. Krenn 2000). Nevertheless, a considerable number of instances are difficult to assign unanimously to one class or the other, as evidenced by the lack of complete agreement between expert annotators (Krenn, Evert & Zinsmeister 2004).

2. Towards exploratory collocation extraction

We have thus identified three key problems for corpus-based studies of lexical collocations: (i) to develop suitable mathematical definitions for the empirical notion of recurrent word combinations; (ii) to achieve a better theoretical understanding of the linguistic phenomenon of collocations; and (iii) to investigate the relation between (different quantitative definitions of) recurrence and (different qualitative types of) collocativity. The "traditional" approaches concentrate on (i) and (iii), respectively, to the extent that they have all but forgotten their common ground (ii). It is now obvious, though, that both sides must address all three issues in order to achieve their goals. Combining approaches (a) and (b), we suggest an incremental exploratory process that works in the following way:

Step 1: Sketch a provisional classification for the subtypes of lexical collocations with clear linguistic definitions for core instances (but allowing "grey areas" at the boundaries).

Step 2: Perform a series of evaluation experiments on different corpora to study the relation between collocativity (according to the criteria set down in step 1) and the various

quantitative definitions of statistical association, then identify the most suitable measure for each subtype of collocations.

Step 3: Use the association measures identified in step 2 to extract comprehensive sets of recurrent word combinations from large text corpora, pre-classified into the subtypes from step 1.

Step 4: Make a detailed linguistic analysis of the extracted data, paying special attention to the grey areas between different subtypes of collocations, where candidates cannot be clearly assigned to one category by the association measures.

Step 5: Refine the theoretical definition and classification of collocations based on the experience acquired in step 4, then repeat the process from step 2.

An essential component of this exploratory approach is the large number of evaluation experiments carried out in step 2, which require manual and conscientious annotation of candidate data according to the provisional classification. Such time-consuming tasks are only practicable when the amount of manual work can be reduced. Fortunately, this is indeed possible by carrying out evaluation experiments on a random sample from the candidate set whose results can then be extrapolated to the full data (Evert and Krenn, to appear).

References

Choueka, Y. (1988). Looking for needles in a haystack. In: *Proceedings of RIAO '88*, 609 – 623.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, IMS, University of Stuttgart. [manuscript available from http://www.collocations.de/EK/]

Evert, S. & B. Krenn (to appear). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. In: *Studies in linguistic analysis*. Oxford: The Philological Society, 1 – 32.

Grossmann, F. & A. Tutin, eds., (2003). *Les Collocations: analyse et traitement*. Amsterdam: De Werelt.

Hausmann, F. J. (2004). Was sind eigentlich Kollokationen? In: *Wortverbindung – mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003*. Berlin: de Gruyter, 309 – 334.

Krenn, B. (2000). The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations. Saarbrücken: DFKI & Universität des Saarlandes.

Krenn, B., S. Evert & H. Zinsmeister (2004). Determining intercoder agreement for a collocation identification task. In: *Proceedings of Konvens '04*. Vienna, Austria.

Manning, C. D. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Yates, F. (1984). Tests of significance for 2x2 contingency tables. *Journal of the Royal Statistical Society, Series A*, **147**(3), 426 – 493.