Exploratory Collocation Extraction

Stefan Evert, University of Osnabrück[†]
Brigitte Krenn, OFAI, Vienna[‡]

[†]*Institute of Cognitive Science, University of Osnabrück, 49069 Osnabrück, Germany*
stefan.evert@uos.de

[‡]*Austrian Research Institute for Artificial Intelligence, Freyung 6/6, 1010 Vienna, Austria*
brigitte@ofai.at

Exploratory Collocation Extraction

Conference themes: lexicographical approach, computational approach

Type of proposal: full paper

Keywords: collocations, cooccurrence, association measures, collocation extraction, evaluation

Abstract

Lexical collocations are a fuzzy phenomenon that has not yet been satisfactorily explained by linguistic theory. At the same time, they are important both for understanding the structure of 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 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 semantically motivated cooccurrences (such as "dark"/"night" and "milk"/"cow") that provide information about the objects and concepts in the world and their properties, collocation extraction is based on spans of a few tokens (or a full sentence) around the instances of a given keyword.

(b) A phraseological notion of collocations as pre-constructed syntactic units (Grossmann & Tutin 2003) or lexically determined elements of grammatical structures (e.g. Choueka 1988), which is prevalent in the lexicographic treatment of word combinations and in computational linguistics. In this view, collocations are characterised by semantic, syntactic and distributional irregularity (cf. Manning & Schütze 1999: 184), i.e. by intrinsic properties of the word combinations rather than their actual occurrences in corpora. The goal of such approaches is to extract a specific type of collocation – according to an intensional definition – with high precision and recall. In order to improve accuracy, it is common to consider only words that cooccur 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, which are defined by their empirical distribution in cor-

pora, 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 operationalise the concept of "recurrence". Most researchers use a statistical criterion, viz. significant association, which may seem to be an objective and indisputably 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 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 lexical collocations and a clear delineation of 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 different classes of collocations. An example is the distinction between support verb constructions and figurative expressions in German, which can be operationalised fairly well (cf. Krenn 2000). Nevertheless, a considerable number of instances are difficult to assign unanimously to one class or the other.

We have thus identified three key problems for corpus-based studies of lexical collocations: (i) to develop suitable (statistical) definitions 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 definitions of) recurrence and (different 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 strategy that works in the following way:

1. sketch a provisional classification of lexical collocations with clear definitions for core instances
2. perform evaluation experiments to find a suitable association measure for each class of collocations
3. extract recurrent word combinations from large corpora, using the measures identified in step 2
4. make a detailed linguistic analysis of the extracted data, paying special attention to the grey areas where candidates cannot be clearly assigned to one class by the association measures
5. refine the theoretical definition and classification of collocations, then repeat 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 and extrapolating the results to the full data set (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. and 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. and 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.

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.