To appear in Machine Translation, DOI 10.1007/s10590-007-9028-8, with different layout and editing

Collocations in a Learner Corpus

Nadja Nesselhauf
University of Heidelberg

Studies in Corpus Linguistics 14
2005. xii, 332 pp., John Benjamins

Reviewed by
Joachim Wagner
National Centre for Language Technology
School of Computing
Dublin City University
Dublin 9, Ireland

Collocations are a feature of natural languages that are not well addressed by current models used for NLP. Language is full of word combinations that occur more frequently than expected, are semantically opaque, or show surprising syntactic restrictions. In her book, Nesselhauf studies collocations in a corpus of second language learner (L2) English. She focuses on German advanced learners of English and describes how these learners use collocation and what mistakes they make. She analyses what factors influence the mastery of collocations and what material interferes with learning.

The structure of the book is very clear. After setting the scene and aims of the study, Nesselhauf develops a definition of collocations suitable for her analysis, introduces the corpus and explains the methodology. Chapter 3 then details the observed verb-noun combinations to which Nesselhauf limits her analysis. This is the longest chapter as deviations are categorised in a detailed error taxonomy with numerous examples and quantitative information. In the next chapter, she looks at the influence of target language and native language material on the deviant collocations produced. Chapter 5 analyses which factors correlate with correct and deviant use of collocations. Besides various features of the collocations itself (intralinguistic factors) details about the learner and the context in which the essay is written (extralinguistic factors) are also considered. Finally, conclusions are drawn for both psycholinguistics (how learners store and produce language) and pedagogy (how collocations should be taught).

There are many different definitions of collocations in the literature. Nesselhauf quickly gives an overview on 23 pages and then develops her own working definition that she employs to categorise collocation candidates that were manually extracted using syntactic patterns. Her definition relies on the combinatorial potential of verb and noun in the sense in which they are used in the verb-noun collocation. The definition is supposed to make the task of distinguishing collocations from free combinations on the one hand and idioms on the other hand easier and more reliable. In practice, this means that, for each word sense, evidence must be collected for how it can combine with other words. Here, Nesselhauf employs two dictionaries as the primary source of information, but also uses the British National Corpus and native speaker judgements in cases not clear from the dictionary entries alone. The classification procedure is presented in detail and gets quite complicated. (The native speakers do not classify the collocation candidates directly. The classification is based on acceptability judgements of artificial combinations that are derived from the combinations in question. In addition, two types of collocations are distinguished.) Unfortunately, there are no diagrams nor pseudo-code to help the reader to grasp the procedure or to verify that they have understood it correctly. Nevertheless, Nesselhauf seems to have achieved the aim of providing both a clear working definition of verb-noun collocations and a reproducible and applicable procedure that keeps subjective judgements to a minimum.

While MT researchers might be interested in Nesselhauf's methodology in order to apply it to a corpus of MT output for evaluation and error analysis, the remainder of the book is probably less relevant to MT. The error taxonomy follows from the syntactic patterns of the collocations and can mostly be read off the section headings of chapter 3. The examples and statistics of collocation use draw attention to the vast amount of possible deviations. However, MT output might be less prone to such deviations as MT systems use language models or chunks of target language material while language learners follow different strategies and often are highly creative.

The book also cannot act as a substitute for a good introduction to collocations or even pre-fabricated units and phraseology in general. Nesselhauf focuses on establishing the required background to develop her own definition. For example, Ludewig (2005) grants more than three times the space to the review of the notion of collocations. Readers with a strong background in statistical NLP (see for example the chapter on collocations in Manning and Schütze 1999) will be surprised to see that the classification is purely qualitative. Possibly, an integrated approach combining qualitative and quantitative descriptions of collocations as in Bartsch (2004) can bridge the gap. Recent advances in quantitative criteria can be found in the compendium compiled by Evert (2005) and in the work of Pecina and Schlesinger (2006).

Readers with a general interest in language, psycholinguistics or language teaching will find interesting results and surprising previous findings confirmed. For example, Nesselhauf's analysis supports previous results that there is no correlation between proficiency and use of collocation. The data actually shows a negative correlation of years spent learning English and relative collocation use. As Nesselhauf studies various factors and interfering material, there are many such nuggets in the text. The section on implications for teaching is also long (20 pages), but in parts too abstract for the layperson. The findings are related to previous suggestions for teaching. Nesselhauf's recommendations include placing more emphasis on common collocations even at advanced levels and teaching how collocations are used in context and the ways in which they are restricted.

The following three reviews came to my attention at a very final stage of writing this review during an occasional web search. Rosi (2005) contains a long description of the contents and a critical evaluation. Cobb's review (2006) is a bit shorter. He identifies that statistical significance testing was not used in some places where it should have, but in general is also very positive. Both target an audience of general linguistics. Reppen (2007) is the shortest review of the three and focuses on the pedagogical implications. He concludes that the book fills an important gap.

References

Bartsch, Sabine (2004): Structural and Functional Properties of Collocations in English – A corpus study of lexical and pragmatic constraints of lexical co-occurrence. Gunter Narr Verlag Tübingen

Cobb, Tom (2006): Collocations in a Learner Corpus (review). In *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, Volume 63, Number 2, pp. 293-295, draft http://www.lextutor.ca/cv/nesselhauf.htm (accessed Nov 05 2007)

Evert, Stefan (2005): The Statistics of Word Cooccurrences - Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/ (accessed Nov 05 2007)

Ludewig, Petra (2005): Korpusbasiertes Kollokationslernen - Computer-Assisted Language Learning als prototypisches Anwendungsszenario der Computerlinguistik, Peter Lang Frankfurt

Manning, Chris and Schütze, Hinrich (1999): Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA

Pecina, Pavel and Schlesinger, Pavel (2006): Combining Association Measures for Collocation Extraction. In *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006),* Sydney, Australia

Rosi, Fabiana (2005): Collocations in a Learner Corpus, Linguist List 16-2896, Oct 07 2005. ISSN: 1068 – 4875. http://linguistlist.org/issues/16/16-2896.html (accessed Nov 05 2007)

Reppen, Randi (2007): Collocations in a Learner Corpus. In *Studies in Second Language Acquisition*, Volume 29, Issue 01, pp 136-137, Cambridge University Press, http://journals.cambridge.org/download.php?file=%2FSLA%2FSLA29_01%2FS0272263107270068a.pdf&code=b968f949cf974016e4a6c55bf429b91c#page=10 (accessed Nov 05 2007)