properties of the Spox construction, supporting its recognition as a discourse marker. In a next step, it proposes three possible roots of motivation for change since the 19th century: (1) the co-occurring "*remind*" context – e.g. *Speaking of catching flies reminds me of political economy.* (1831, MAG, NewEngMag), (2) the sentence-initial adverbial – e.g. *Speaking of the general question of doing government work by contract, I expressed the view*[…] (1910, MAG, Scribners), and (3) the "*now that*" clause – e.g. *And now that we're speaking of profits, Mr. Crashly, I have this thought to put before you.* (1950, FIC, SomethingValue). It will be argued that each root contributes primarily, but not exclusively, to some aspect/s of the establishment and consolidation of the discourse marker status of the Spox construction. In conclusion, this paper shows that the development of the Spox construction as a multi-word discourse marker is rooted in the formal and conceptual blending of all three input domains.

References

Brinton, L. J. 2008. The Comment Clause in English: Syntactic Origins and Pragmatic Development. Cambridge: Cambridge University Press.

Fraser, B. 2009. Topic orientation markers. *Journal of Pragmatics*, 415: 892-98.

Ferrara, K. W. 1997. Form and function of the discourse marker anyway: implications for discourse analysis. *Linguistics*, 352. 343-78.

Gohl, C., & Günthner, S. 1999. Grammatikalisierung von weil als Diskursmarker in der gesprochenen Sprache. *Zeitschrift für Sprachwissenschaft*, 18(1): 39-75.

Lewis, D. M. 2011. A discourse-constructional approach to the emergence of discourse markers in English. *Linguistics*, 492: 415-43.

Prevost, S. 2011. A propos from verbal complement to discourse marker: a case of grammaticalization? *Linguistics*, 492, 391-413.

(Association) measure for measure: Comparing collocation dictionaries with co-occurrence data for a better understanding of the notion of collocation

Sabine Bartsch (TU Darmstadt)
Stefan Evert (FAU Erlangen-Nürnberg)
Thomas Proisl (FAU Erlangen-Nürnberg)
Peter Uhrig (FAU Erlangen-Nürnberg)

Lexical collocations are a complex phenomenon for which neither traditional nor cognitive linguistic theories have yet found satisfactory definitions that would allow for a lexicographically convincing operationalisation. Despite the pervasiveness of collocations in language and their importance for our understanding of the structure of human language as well as for many applications, their definition and characterisation leave many questions unanswered.

Corpus-based studies of collocation and the development of collocation extraction tools have been influenced by two principal views: (a) an empirical notion of collocation (Firth 1957), which builds upon the recurrent co-occurrence of lexical items in more or less clearly defined contexts; (b) phraseological notions of collocation which are prevalent in lexicography and characterise collocations on the basis of their semantic, syntactic and distributional irregularity (cf. Hausmann 1979; 1984; 1985; Manning & Schütze 1999: 184). Other, related definitions equate collocations with lexicalised multiword expressions (as is often done in computational linguistics, e.g. Choueka 1988) or focus on their cognitive reality, using evidence from priming studies (Durant & Doherty 2010) or plausibility judgements (Lapata *et al.* 1999).

The operationalisation of such collocation definitions, which is necessary to allow for their reliable identification in corpora, remains a notoriously difficult issue. The situation is similar for related questions such as the choice of an appropriate quantitative measure of the association between co-occurring words, the influence of the quality and size of the corpus, and the qualitative evaluation of automatically extracted collocation candidates.

The aim of this paper is to report on work towards a better understanding of different notions of collocation and their operationalization and towards gauging the reliability of automatic collocation identification in large corpora. To this end, the research reported in this paper compares a sample inventory of collocations listed in two specialized collocation dictionaries (the pre-corpus era BBI and the corpus-based OCD2; see also Lea 2007) with measures of statistical association in linguistic corpora of different sizes and with different levels of linguistic pre-processing. The resulting collocation candidates are manually evaluated against a well-defined subset of data from the two dictionaries.

It will be shown that at least for common general language collocations such as those listed in dictionaries, smaller and cleaner corpora such as the BNC deliver better lists of collocation candidates than larger, but noisy web corpora. Furthermore it will be shown that syntactically annotated data are not only superior for collocation extraction from BNC-like corpora (as shown in previous work), but also for web-based corpora despite the relatively low accuracy of automatic syntactic annotation tools on such data.

Comparing different statistical association measures – such as log-likelihood ratio, t-score, chi-squared, several variants of Mutual Information and directional measures such as ΔP – with the dictionary data, we discover some surprising facts: MI² (Daille 1994: 193) and t-score correspond better to lexicographers' intuitions than the classical MI measure that has long been popular in computational lexicography; the widely-held belief that the chi-squared test is unsuitable for collocation identification (Dunning 1993) is not always true; finally, the relative usefulness of an association measure depends much less on the quality, amount and annotation of the corpus data than on the particular notion of collocation to be identified (i.e. on which dictionary is used as a "gold standard").

Thus – returning to the original question as to the concept of collocation – we can show that the collocation dictionaries used in the present study differ substantially with respect to their view of what should be listed as a collocation, which may (at least in part) be due to the fact that one of the two was created in the pre-corpus era. The automatic evaluation allows us to compare both dictionaries against the corpus findings but also to compare to what extent the explicit definition of collocation (as stated by the editors) and the implicit definition (i.e. the selection of collocations) correspond.

References

BBI: *The BBI Combinatory Dictionary of English. A Guide to Word Combinations*. Amsterdam: Benjamins, 1986. [3rd ed 2010.]

Choueka, Y. 1988. Looking for needles in a haystack. In *Proceedings of Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications): RIAO "88,* Cambridge, MA, 609-23.

Daille, B. 1994. *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7. http://www.bdaille.com/index.php?option=com_docman&task=doc_download&gid=8. Accessed 27.03.2015.

Dunning, T. E. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61-74.

Durrant, P. & A. Doherty. 2010. Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, 6(2), 125-55.

Firth, J. R. 1957. Papers in Linguistics 1934–1951. Oxford: OUP.

Hausmann, F. J. 1979. Un dictionnaire de collocations est-il possible? *Travaux de linguistique et de littérature*, 17. 187-95.

Hausmann, F. J. 1984. Wortschatzlernen ist Kollokationslernen. *Praxis des neusprachlichen Unterrichts*, 31. 395-406.

Hausmann, F. J. 1985. Kollokationen im deutschen Wörterbuch: Ein Beitrag zur Theorie des lexikographischen Beispiels. In H. Bergenholtz & J. Mugdan eds. *Lexikographie und Grammatik*. Tübingen: Niemeyer, 118-29.

Lapata, M., S. McDonald & F. Keller. 1999. Determinants of adjective-noun plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL 1999)*, Bergen, Norway, 30-36.

Lea, D. 2007. Making a collocations dictionary. *Zeitschrift für Anglistik und Amerikanistik*, 55(3), 261-72.

Manning, C. D. & H. Schütze 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA/London: MIT Press.

OCD2: *Oxford Collocations Dictionary for Students of English*. 2nd ed. Oxford, 2009.

From light verb constructions to negative polarity items: The cases of *take notice of* and *make mention of*

Eva Berlage (University of Hamburg)

It is well known that English has a series of so-called negative polarity items (NPIs) whose occurrence is restricted to or strongly preferred in non-assertive contexts. Typical examples include the *any* class of items (e.g. *any*, *anybody*, *any longer*, *any more*, *anything*), various grammatical items (e.g. *much*, *either*, *ever*, *yet*), the modal auxiliaries *dare* and *need*, a few lexical verbs (e.g. *bother* + infinitival, *budg*e, *faze*) and a vast range of idioms such as *can be bothered*, *give a damn*, *see a (living) soul* etc. (for a more comprehensive list, see e.g. Huddleston/Pullum et al. 2002: 823; von Bergen/von Bergen 1993). Far less well known is the evolution of these polarity sensitive items and the fact that the development of some of them can be linked to such historical processes as lexicalisation or its counter-image (for the