Dictionaries and spoken language: the beginnings of a second corpus revolution?

Dirk Siepmann (University of Osnabrück, Germany)

The present study starts from the observation that traditional lexicography has tended to rely on corpora of written text. It is hypothesized that this might be to the detriment of covering the commonest colloquial lexical units which carry the main burden of everyday conversation, are usually acquired early on in life and are therefore more deeply anchored in speakers' brains than units first encountered in the course of education.

This hypothesis receives support from a detailed examination of the treatment accorded four English and four French high-frequency words and three English and three French high-frequency phrasemes in ten different dictionaries as well as from a spot check on twenty medium-frequency phrasemes.

The methodology used for this purpose proceeded in seven steps. The first step involved generating several frequency lists, viz. a) a frequency list of the 3000 most common lemmas in the spoken portion of the *Corpus de référence du français contemporain* and a large corpus of spoken British and American English and b) three lists of the 3000 most common 3-, 4- and 5-grams; word frequencies and multi-word strings were identified using the relevant functions of the *Sketchengine*. In the second step, lemmas either labelled 'colloquial' in several dictionaries or identifiable as significantly more frequent in speech than in writing were extracted from frequency list A, and lexical bundles that did not constitute phrasemes were eliminated from frequency list B. In the third step, a random selection was made of four items from the first list and three items from the second list. In the fourth step, the frequency data obtained for individual lemmas were compared with native-speaker usage ratings. The fifth step involved a detailed analysis of the lexico-grammatical and pragmatic features of the selected items, following the corpus-driven approach to habitual co-occurrences of words ('usuelle Wortverbindungen') developed at the *Institut für Deutsche Sprache* (Steyer, 2009, 2013). This approach is based on three methodological premises which draw inspiration from the British tradition of text analysis established by Firth and Sinclair, viz. a) it derives structure from the data during the analysis rather than in advance; b) it foregrounds language as use; and c) it lets the data speak for itself, allowing the observer to form an unbiased picture of authentic language in use. The data are listed in terms of node words and their primary and secondary collocates (e.g. *never occurred to me* in the case of the English node *idea*) and are then subjected to thorough scrutiny with a view to determining the internal structure and typical variation found with node-collocate pairs and establishing the presence or otherwise of node-collocate pairs with similar characteristics. This fifth step thus involves two sub-stages (cf. also Hanks, 2013, p. 92): the first involves grouping the evidence into recurrent semantic-pragmatic patterns; the second is the assignment of meaning to each pattern.

The general finding is that current lexicographic descriptions of spoken French are often patchy and inadequate with respect to various lexico-grammatical features, while the description of spoken English is somewhat more advanced though far from satisfactory.

Most notably, there was found to be a dearth of information on the collocational range of colloquial items in most of the French and the English dictionaries under investigation, although, at least theoretically, native-speaker lexicographers could have retrieved some collocations from memory. There is on the whole a considerable uniformity in the content of both monolingual and bilingual dictionaries, with French monolingual learners' dictionaries and bilingual dictionaries tending to adopt collocations and sense divisions from standard reference works such as the *Petit Robert*.

Since the extraction of collocations is an essential prerequisite for the determination of meanings, it is hardly surprising to find that the marking of sub-senses may not be sufficiently clear for the encoding needs of non-native speakers of either English or French. Here the clearest examples are the highly polysemous verb *lâcher* and the complex preposition *autour de* in French. The only way to enable learners to gain an overview of, and ultimately to make productive use of, a verb like *lâcher* is to opt for splitting up the various senses derived from its literal meaning 'ne plus tenir' rather than lumping them all together, as PR does. It would be unrealistic to expect learners to derive such specific uses as the following from the general sense, especially since this use features a non-human subject:

(1) C'est un film qui vous prend à la gorge dès le départ et ça ne vous lâche plus.

Another main finding concerns the rudimentary treatment of common multi-word items with a clear discoursal function. Thus, nine of out of ten dictionaries fail to record the sense in which the common French preposition *autour de* is used to indicate that someone or something is at the centre of a particular endeavour, and all the dictionaries give low priority to the use of French *n'importe quoi* as a discourse marker. Most seriously perhaps, there is no indication of the typical contextual embedding or common lexical collocations of discourse markers (e.g. - *N'importe quoi*. – *Ah si*.). The aforementioned spot check on twenty medium-frequency items shows that coverage differs between dictionaries and that differences between American and British English often go unrecorded.

Most of the examples found in the dictionaries under investigation illustrate written usage, and of those that illustrate spoken usage many lack some naturalness feature or other. There is some evidence that different types of exemplification may be needed for different words (cf. Hausmann, 2005). *lâcher* is one example of a low-collocability word which cannot be illustrated by means of typical co-occurrences. With such words, users will need a large number of (at least) sentence-length examples to grasp the various meanings of the word. The case is different with the French noun *look*, a clear example of a high-collocability word which requires little exemplification beyond information on collocation.

Three dictionaries deserve special comment: DAFLES, *Harraps* and *Longman Dictionary of Contemporary English*. Although it has several compensating strengths which fall outside the scope of this study (cf. Verlinde, Binon & Selva, 2006), DAFLES has almost no entries for colloquial words, a fact which may be due to its corpus base.

By contrast, Harraps achieves a remarkable harmonization of descriptive and pedagogic needs. A measure of the overall quality of this monument to bilingual lexicography is the inclusion of a large number of colloquial senses of both *lâcher*

and *péter* and the provision of illustrative sentences from which students may confidently extrapolate personal choices. Like all current dictionaries, however, even Harraps is still weak on discoursal items. Longman stands out from the other monolingual dictionaries in recording a fairly large number of colloquial collocations and in achieving broad coverage of multi-word markers.

There are two important theoretical lessons to be drawn thence. The first, which concerns corpus linguistics, is that medium-sized spoken corpora like the CRFC or the BEspoken corpus will shed light on lexical patterns and collocations about which even very large mega-corpora of written language are completely uninformative. This means that there may well be a second corpus revolution ahead which will apply Sinclair's famous dictum that 'the language looks rather different when you look at a lot of it at once' to the investigation of intimate and colloquial language use.

The second theoretical lesson is that colloquial words, far from being stylistically 'inferior' substitutes of more formal words, are imbued with their own specific shades of meaning, phraseology, and pragmatics. It is as if there is a primary lexis which is even more deeply submerged in the routines of everyday life and thus even less accessible to native-speaker intuition than the secondary written lexis, but there can be no doubt that such lexis is communicatively prior and that its detailed description is of crucial importance to second or foreign language learners. The present study suggests that much of the primary lexis of French and English, and very probably other languages as well, remains almost undescribed in respect of many of its features, with dire consequences for foreign learners aspiring to acquire native-like proficiency. Reliance on the rarer and clumsier words or lexical units may make their language use sound stilted and unnative-like (e.g. *il n'abandonne jamais son portable* rather than *il ne lâche jamais son portable*).

References

Atkins, B.T. et al. (2010). *Collins/Robert French-English/English-French Dictionary*, Ninth edition. Glasgow / London: Collins.

Cambridge Advanced Learner's Dictionary. Retrieved from http://dictionary.cambridge.org/. Cambridge University Press.

Collins Cobuild English Learner's Dictionary. Retrieved from https://www.collinsdictionary.com/. Collins

Rey-Debove, J. (Ed.). (1999). *Dictionnaire du français*. Le Robert & Cle International.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge: MIT Press.

Hausmann, F. J. (2005). Isotopie, scénario, collocation et exemple lexicographique. In *L'exemple lexicographique dans les dictionnaires français contemporains. Actes des "Premières Journées allemandes des dictionnaires" (Premières Journées allemandes des dictionnaires, Klingenberg am Main, 25-27 juin 2004) (Lexicographica Series Maior 128).* Heinz, M. (Eds).. Tübingen: Niemeyer, 283-292.

Verlinde, S., Selva, T., Bertels, A. and J. Binon (Eds.). *Dictionnaire d'apprentissage du français langue étrangère ou seconde*. Leuven : Institut Interfacultaire des Langues vivantes. Retrieved January 24, 2015 from http://ilt.kuleuven.be/blf/search.php

Nicholson, K. & G. Pilard. (2012). *Harrap's Slang - Dictionnaire d'argot et d'anglais familier.* Paris: Larousse.

Langenscheidt-Redaktion. (2010). *Langenscheidts Handwörterbuch Deutsch Französisch/Französisch-Deutsch*. Berlin: Langenscheidt.
Longman Dictionary of Contemporary English. Retrieved from http://www.ldoceonline.com/. Longman
Macmillan English Dictionary. Retrieved from http://www.macmillandictionary.com/. Macmillan Publishers Limited
Meißner, F. J. (1992). *Langenscheidts Wörterbuch der französischen Umgangssprache*. Berlin: Langenscheidt.
Oxford Advanced Learner's Dictionary. Retrieved from http://www.oxfordlearnersdictionaries.com/. Oxford University Press
PONS. *Großwörterbuch Französisch. Deutsch-Französisch/Französisch-Deutsch*. Stuttgart: PONS.
Robert, P., Rey-Debove, J. & A. Rey. (Eds). (2008). *Nouveau Petit Robert: dictionnaire alphabétique et analogique de la langue française*. Paris: Le Robert.
Sketchengine Retrieved from http://www.sketchengine.co.uk/. Lexical Computing CZ
Stevenson, A. (2007). *Harrap's Unabridged PRO French-English/English-French*. Ottawa: Laurier Books Ltd.
Steyer, K. (2013). *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht (Studien zur Deutschen Sprache 65).* Tübingen: Narr.
Steyer, K. & A. Brunner. (2009). *Das UWV-Analysemodell. Eine korpusgesteuerte Methode zur linguistischen Systematisierung von Wortverbindungen*. Mannheim: Institut für Deutsche Sprache.
TLF: Trésor de la langue française*. Dictionnaire de la langue du XIXe et du XXe siècle (1789–1960), publié sous la direction de Paul Imbs*. Paris: Éditions du Centre National de la Recherche Scientifique 1971–1994. Retrieved January 24, 2015 from http://atilf.atilf.fr.
Verlinde, S., Binon, J. & T. Selva. (2006). *Corpus, collocations et dictionnaires d'apprentissage Langue française.* 150, 84-98.