

# Lexicography and phraseology

Magali Paquot

[DRAFT – To appear in: Douglas Biber & Randi Reppen (eds.), *The Cambridge Handbook of Corpus Linguistics*. Cambridge: Cambridge University Press.]

## 1. Introduction

Corpus linguistics has contributed to lexicography in a number of ways. It has provided the methods and tools for lexicographers to better assess the relative importance of different words and their different uses. It has led to the development of innovative approaches to the lexicographical treatment of meaning, grammar, and pragmatics and extended the entire scope of lexicographic research (Teubert, 2001; Hanks, 2009; Hanks 2012a). However, it is probably in the lexicographical description of phraseology that corpus linguistics has had the most revolutionary effect. Evidence of word use in corpora has shown to an unprecedented extent that words are not isolates but rather combine with each other in preferred syntagmatic patterns to acquire meaning (e.g. Sinclair, 1991; Hanks, 2012b).

This distinctive vision of language (and language study) is rooted in Firth's contextual theory of meaning and developed in the work of M. A. K. Halliday, J. Sinclair and M. Hoey. Stubbs (1993: 2-3) lists 9 principles underlying British linguistics in the Firthian tradition, four of which general ideas are of particular significance for the lexicographical treatment of phraseology: (1) Language should be studied in actual, attested, authentic instances of use, not as intuitive, invented, isolated sentences; (2) Linguistics is concerned with the study of meaning: form and meaning are inseparable; (3) There is no boundary between lexis and syntax: lexis and syntax are interdependent; and (4) Much language use is routine. The latter principle places collocation, i.e. "a relation of affinity which holds between words in a language, and which is revealed by the typical co-occurrence of words" (Seretan 2011: 10) [see also Xiao, this volume], at the forefront of any linguistic or lexicographic description of language and has been articulated by Sinclair in the form of the 'idiom principle':

The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments. (Sinclair 1991:110)

In Sinclair's view of language, phraseology is central: phraseological items, whatever their nature, take precedence over single words (Granger and Paquot, 2008: 29). This is perhaps best illustrated in the tendency of very frequent words such as *make*, *know*, or *fact* to derive their meanings from the context in which they occur. As put by Moon (2007: 168), "Evidence shows that their most frequent uses are often delexicalized: thus semantically independent meanings of *take* such as 'remove, move, steal, escort' are less common than its use in structures such as *take a step*, *take part*, *take a long time*".

Sinclair's pioneering corpus work was first put into practice lexicographically in the *Collins Cobuild English Language Dictionary* (CELLE), a monolingual dictionary for learners of English published in 1987. The concern for word combinations was not new in pedagogical lexicography: phraseology had long been recognized as an essential component of native-like fluency and idiomatic language use (Pawley and Syder, 1983) and an early and groundbreaking attempt to help learners to encode meaning in a native-like way by focusing on collocations, idioms and speech formulae can be traced back to Hornby et al.'s (1942) *Idiomatic and Syntactic English Dictionary* (cf. Cowie, 1999: 42), though the word

combinations “were identified from experience of language and language teaching” (Moon 2007: 168).

Since the CCELD’s publication in 1987, the use of corpus data has spread rapidly to English pedagogical lexicography. Dictionaries for other languages, specialised dictionaries and bilingual dictionaries are now also following suit. The aim of this chapter is to assess the impact of corpus data on the description of phraseology in various types of English dictionaries.<sup>1</sup> The focus is on electronic dictionaries as ‘Today lexicography is largely synonymous with electronic lexicography and many specialists predict the disappearance of paper dictionaries in the near future’ (Granger, 2012: 2). Section 2 starts with a brief description of the Sketch Engine, probably one of the corpus query systems most frequently used by lexicographers today to analyse the preferred environment of words. It then shows how the use of corpora has been instrumental in developing new access routes to phraseological units in learners’ dictionaries, before providing an overview of studies that have dealt with word combinations in English lexicography, and more particularly their coverage, i.e. how many and which word combinations are listed in dictionaries. Section 3 reports on a study that investigates the usefulness of phraseological information in learners’ dictionaries to answer the specific needs of non-native students and researchers who have to write in academic settings and makes out a case for the use of a wider range of specialised corpora in dictionary making. The chapter ends with suggestions for further and better integration of corpus-data (and corpus-query tools) to improve the coverage of and access to phraseology in dictionaries.

## 2) In search of phraseology: from dictionary making to the study of dictionaries

A wide range of software tools and resources are available today to assist in the identification and analysis of word combinations. The Sketch Engine is a corpus tool originally developed to answer the specific needs of lexicographers. It features all the characteristics of a good concordancer, i.e. display of concordance lines in KWIC format, sorting options, frequency distribution, collocation statistics (see Rayson this volume; Xiao this volume) as well as unique components such as Word sketches. These are corpus-based summaries of a word’s collocational behaviour and provide separate lists of collocates ordered by decreasing frequency or statistical significance for each grammatical relation a word participates in. Figure 1, for example, provides a sample of the Word sketch for the verb *suggest*. The first two columns display the most frequent subjects and objects of the verb *suggest*. The third column shows that *suggest* is commonly used in the past participle form followed by the preposition *by* and lists the nouns that are typically introduced by the preposition. The fourth column provides a list of adverbs that typically follow the verb *suggest*.

Other features of the Sketch Engine that are particularly useful for the lexicographic description of phraseology include:

1. A thesaurus which provides a list of ‘nearest neighbours’ for a word, i.e. words that share collocates with the search word and may therefore be potential synonyms (or possibly antonyms). For example, the 10 nearest neighbours for the noun *argument* are *claim, idea, view, theory, interpretation, explanation, account, concept, principle* and *reason*.
2. The Sketch-Diff option compares word sketches for two words and identifies the collocates that they have in common and those that are unique to each word.

---

<sup>1</sup> Moon (2008:314) has examples of corpus-based lexicographic projects in other languages.

See Kilgarriff and Kosem (2012) for a comprehensive overview of the Sketch Engine.

**suggest** (verb) Corpus of Academic Journal Articles (CAJA) freq = 64485 (683.4 per million)

| subj_NP     | 20787 | 4.5   | NP           | 15017 | 2.3   | PP_by-i     | 1748 | 4.5   | AVP        | 948 | 0.7   |
|-------------|-------|-------|--------------|-------|-------|-------------|------|-------|------------|-----|-------|
| evidence    | 1350  | 45.28 | role         | 436   | 29.96 | study       | 57   | 14.96 | here       | 135 | 40.9  |
| study       | 916   | 28.09 | possibility  | 254   | 31.75 | theory      | 36   | 16.41 | otherwise  | 69  | 42.5  |
| research    | 644   | 32.2  | way          | 218   | 19.89 | analysis    | 32   | 11.53 | earlier    | 62  | 38.15 |
| result      | 601   | 24.86 | presence     | 207   | 26.69 | fact        | 27   | 15.0  | above      | 58  | 38.61 |
| analysis    | 417   | 20.69 | relationship | 205   | 20.18 | model       | 26   | 8.21  | only       | 52  | 20.8  |
| finding     | 375   | 32.25 | need         | 169   | 24.46 | observation | 21   | 15.85 | either     | 38  | 29.98 |
| literature  | 306   | 29.38 | mechanism    | 161   | 21.51 | literature  | 19   | 14.06 | strongly   | 36  | 27.92 |
| theory      | 298   | 22.93 | effect       | 159   | 11.29 | finding     | 18   | 13.63 | instead    | 34  | 27.1  |
| model       | 285   | 13.43 | existence    | 157   | 29.01 | result      | 18   | 7.05  | in         | 23  | 26.26 |
| datum       | 215   | 13.55 | approach     | 134   | 15.04 | author      | 16   | 15.12 | previously | 22  | 20.6  |
| author      | 197   | 28.49 | model        | 132   | 8.47  | datum       | 16   | 6.64  | further    | 21  | 19.67 |
| work        | 190   | 15.79 | importance   | 119   | 21.93 | research    | 15   | 8.67  | not        | 21  | 6.18  |
| observation | 162   | 22.93 | difference   | 112   | 11.61 | presence    | 13   | 10.85 | again      | 20  | 20.41 |
| researcher  | 116   | 22.72 | use          | 109   | 13.53 | experiment  | 12   | 8.87  | that       | 18  | 28.03 |
| article     | 99    | 17.43 | link         | 93    | 22.48 | evidence    | 12   | 7.8   | much       | 18  | 16.1  |
| paper       | 99    | 15.96 | change       | 93    | 9.47  | researcher  | 10   | 11.52 | elsewhere  | 17  | 24.12 |
| approach    | 96    | 9.68  | reason       | 83    | 14.85 | Figure      | 9    | 11.63 | below      | 16  | 22.27 |
| hypothesis  | 95    | 16.53 | involvement  | 82    | 22.62 | work        | 9    | 5.02  | also       | 16  | 7.83  |
| Research    | 93    | 24.72 | factor       | 73    | 8.61  | referee     | 8    | 19.81 | therefore  | 12  | 13.15 |
| report      | 90    | 16.85 | degree       | 72    | 14.89 | example     | 8    | 5.34  | perhaps    | 10  | 14.92 |

**Figure 1: a sample of the WordSketch for the verb ‘suggest’**

Lexicographers can also make use of admirable online corpus-based resources such as the Database of Analysed Texts of English (DANTE) and the Pattern Dictionary of English Verbs (PDEV). DANTE is a lexical database that provides a fine-grained description of the meanings, grammatical and collocational behaviour, and text type characteristics of over 42,000 English words. It was created for lexicographers and computational linguists, using a custom-built corpus of 1.7 billion words uploaded in the Sketch Engine. The PDEV is an ongoing corpus-driven project in which a procedure called Corpus Pattern Analysis (CPA) is applied to identify the various patterns in which a verb is used and then discover how exactly meanings arise from each of the patterns. The completed project will contain around 5,800 verb entries (Hanks, 2012b: 424-429). See Hanks (2013) for a detailed discussion of Corpus Pattern Analysis.

The systematic exploitation of corpus data in the process of dictionary making, together with the advent of electronic dictionaries, has resulted in the reshaping of lexical entries. This is especially true of electronic learners’ dictionaries where phraseological information may surface in various elements of the microstructure:

1. The full-sentence definition format in which words are defined in context “to represent the match between phraseology and meaning” (Moon, 2008a: 251) as illustrated in the following example from the *Collins-Cobuild Advanced Learner’s English Dictionary* (6<sup>th</sup> edition):

If you **make an issue of** something, you try to make other people think about it or discuss it, because you are concerned or annoyed about it. (CCAD6)

2. The use of corpus-derived examples which indicate typical lexicogrammatical patterns and collocations.

In the *Oxford Advanced Learner's Dictionary* (8<sup>th</sup> edition), for example, a selected list of adjective and verb collocates are highlighted in the examples that are used to illustrate sense 1 of the noun *issue*:

**1 [ countable ]** an important topic that people are discussing or arguing about  
*a **key/sensitive/controversial issue***

*This is a **big issue** ; we need more time to think about it.*

*She usually writes about environmental **issues**.*

*The union plans to **raise the issue** of overtime.*

*The party was divided on this **issue**.*

*You're just **avoiding the issue**.*

*Don't **confuse the issue**.* (OALD8)

3. The display of a restricted set of collocations and fixed phrases in the body of a lexical entry.

In the *Macmillan English Dictionary for Advanced Learners* (2<sup>nd</sup> edition), for example, each example sentence is preceded by the collocation it illustrates highlighted in bold, as shown with sense 22 of the verb *take*:

**22 [transitive]** to have or to show a feeling or opinion

**take offence** I'm afraid she took offence at my remarks.

**take (an) interest** He's never taken much interest in his kids.

**take a view/attitude** I take the view that children should be told the truth.

(MEDAL2)

Fixed phrases (most particularly idioms and phrasal verbs) are often treated as sub-entries. In the *Longman Dictionary of Contemporary English* (5<sup>th</sup> edition), for example, fixed phrases with the noun *point* are listed as separate senses of the word:

**14 to the point** dealing only with the important subject or idea, and not including any unnecessary discussions:

*Her comment were brief and to the point.*

**15 make a point of doing something** to do something deliberately, even when it involves making a special effort:

*He made a point of spending Saturdays with his children.*

*I always make a point of being early.*

(LDOCE5)

4. Collocation boxes where salient collocates as identified by a statistical analysis of corpus data are organised by part of speech (Figure 2).



**Figure 2: Collocation box for sense 1 of the noun ‘talent’ in MEDAL2**

LDOCE5 is the only dictionary that provides collocation boxes and phrase banks for almost each word while MEDAL2 deserves special mention for offering collocation boxes at the level of a word sense rather than for the word in general (see Figure 2).

Some of these features are typical of just one dictionary: the full sentence definition format, for example, is a hallmark of the Collins Cobuild series (Hanks, 1987). Features such as corpus-based examples are mainstream in today’s electronic pedagogical lexicography while a selected list of specific collocations and phrases is often offered in bilingual dictionaries.

Lexicographic research has addressed the question of accessibility of phraseological information in dictionaries (e.g. De Cock and Granger, 2004; Herbst and Mittmann, 2008) and a few studies of dictionary use have investigated whether dictionary users were able to locate in different dictionaries the collocations they needed to perform a given task (e.g. Lew, 2012). However, most research has concentrated on aspects of coverage, i.e. how many and which phraseological units are listed in dictionaries (e.g. De Cock, 2002; Coffey, 2006; Siepmann, 2008; Götz-Votteler and Herbst 2009).

Two studies by Walker (2009) and Moon (2008b) stand out as they compare different types of dictionaries. Walker (2009) examined the way in which collocations are treated in three learners’ dictionaries (COBUILD5, LDOCE4 and OALD7), three collocation dictionaries (i.e. the *Oxford Collocations Dictionary for Students of English*, 2002; the *BBi Dictionary (revised edition) of English Word Combinations*, 1997, BBI; the *LTP Dictionary of Selected Collocations*, 1999, LTP) and two dictionaries of business English (the *Oxford Business English Dictionary for Learners of English*, 2005, OBED; the *Longman Business English Dictionary*, 2000, LBED). The selected dictionaries are corpus-based except for two collocation dictionaries, i.e. BBI and LTP. The study focused on eighteen semantically-related nouns (e.g. *issue*, *aspect* and *factor*) and verbs (e.g. *run*, *head* and *manage*) from the domain of business English. Findings from a corpus-based analysis of the collocational behaviour of these lexical items were used as a benchmark for assessing the treatment of collocations in the different dictionaries. Two corpora were queried, i.e. the Bank of English and a more specialised corpus of business English made up of the commercial and financial data files from the British National Corpus (6.3 million words). The collocations included in the entries for each of the eighteen selected items in the different dictionaries were listed and compared with those revealed by the corpus-based analysis. Major findings include:

1. There is a lack of consistency in the collocations recorded in the three learners’ dictionaries. Only 5% of the collocates listed in the learners’ dictionaries appear in the

three dictionaries and 24% appear in two of the three dictionaries. Put differently, 71% of the collocates appear in just one of the three dictionaries (see also Coffey, 2006).

2. Most of the collocations included in the learners' dictionaries correspond to those identified in the corpora but the dictionaries tend to record the most frequent collocates (e.g. *key/main issue* rather than *contentious issue*). Walker (2009: 290) argued that, as a result, the learners' dictionaries often include the same collocates in entries for near synonyms, thus failing to highlight the differences in their semantics.
3. A similar picture emerges from the analysis of the two dictionaries of business English. OBED and LBED include many more collocations from the field of business and commerce (e.g. *growth target*, *earnings target*) but, like the learners' dictionaries, they prioritize the most frequent collocates.
4. There is very little agreement in the collocates recorded in the three collocation dictionaries: only 3% of the total number of collocates listed are found in all three dictionaries, and 82% appear in only one of the three dictionaries.
5. The *Oxford Collocations Dictionary for Students of English* is the only one of the three collocation dictionaries which is corpus-based. Not surprisingly (and perhaps luckily enough), it was found to contain the largest number of collocates which were the same or similar to those revealed by the corpus analysis (ibid, 295).

One year earlier than Walker, Moon (2008b) adopted a similar methodological framework to evaluate the coverage of collocations in monolingual dictionaries for native speakers of English, monolingual learners' dictionaries and bilingual French-English dictionaries. After a general description of the collocational behaviour of the three English words *river*, *rivet* and *riven*, as observed in the 450-million word Bank of English, the study examined how it is represented in the different types of dictionaries. Moon's analysis is particularly enlightening in that it offers a diachronic perspective to current lexicographic practice and places emphasis on "the function of phraseological information in relation to the needs and interests of the target users" (ibid, 333). The analysis of the *Collins English Dictionary* (2003) and the *New Oxford Dictionary of English* (1998) showed that dictionaries for native speakers scarcely represent the phraseological patterns of *river*, *rivet* and *riven* as identified in the Bank of English. This lack of phraseological information most probably stems from the fact that, in a monolingual dictionary for native speakers, information presented has essentially been for decoding: "a primary role of a native-speaker dictionary is to list and explain the lexical items of a language" (Moon, 2008b: 318). Learners' dictionaries, on the other hand, have from the outset taken proactive steps to help learners encode in idiomatic English (cf. Rundell, 1999). Moon (2008b) found that preferred collocates, prepositional selections and pattern structures are incorporated into definitions and examples in the OALD7, the LDOCE4 and the COBUILD2 (see also De Cock and Granger (2005) for a discussion of how learner corpora can be used to identify learners' difficulties and to improve aspects of prevention of error in learners' dictionaries). Although bilingual dictionaries have always shown more awareness of phraseology than monolingual native-speaker dictionaries, Moon (2008b) found that the phraseology of French nouns such as *fleuve* and *rivière* is poorly covered in the Collins Robert French-English English-French Dictionary (2002, 6<sup>th</sup> edition) and the Oxford Hachette French Dictionary (2001, 3<sup>rd</sup> edition). The two dictionaries focus primarily on the translations of compounds (e.g. *river basin* and *bassin fluvial*) and not on other types of phraseological units such as collocations.

The evidence of current research clearly shows that learners' dictionaries have played a pioneering role in the description of phraseology and that other kinds of dictionaries have lagged behind. Today, however, the most exciting developments are to be found in bilingual

lexicographic research. The lack of large corpora for other languages than English is one of the greatest impediments to the successful treatment of bilingual phraseology and Ferraresi et al.'s (2010) study is particularly important in that it presents Web corpora as viable alternatives and valid reference resources for lexicographic purposes. In the first part of the study, they made use of a very large Web-derived corpus of British English (ukWaC, 1.9 billion words) and the BNC to extract collocational pairs with three English lexical headwords, i.e. the adjective *hard*, the noun *point* and the verb *charge*. The extracted pairs were ranked according to the Log-likelihood measure and the top 30 pairs for each headword extracted from the ukWaC and the BNC were merged into a single alphabetically ordered list. A lexicographer was then asked to flag the 129 different collocational pairs that he believed could be included in the English half of an English-French bilingual dictionary. The results of the expert validation indicate that over 70% of these pairs automatically taken from the two corpora could well be relevant. While the BNC and ukWaC share 45% of the validated collocational pairs (e.g. *melting* + *point*, *hard* + *cash*, *charge* + *offence*), each corpus also contains between 25% and 30% of validated collocational pairs not found in the other. A closer look at these pairs revealed that the ukWaC performs slightly better than the BNC: it offers a better coverage of different word senses and provides “a more up-to-date snapshot of language in use” (Ferraresi et al., 2010: 353). For example, the ukWaC includes more instances of collocations illustrating the ‘take as payment’ sense of the verb ‘charge’ as in *charge* + *fee*, *VAT*, *penalty* or *rent* as well as instances of the pattern ‘charge + PERSON’ (e.g. *customer*), a pattern that is not found in the BNC.

In the second part of the study, they show that data of high linguistic quality can also be obtained from a Web-derived corpus of French (frWaC, 1 billion words). For example, they used the frWaC to extract a list of the 60 most frequent noun collocates in a span of 1 to 3 words to the right of two translation equivalents of the verb *charge*, namely *inculper de* and *accuser de*. Potential translation equivalents for 12 out of 16 collocational pairs found in the ukWaC were found in the resulting list (e.g. *charge* + *burglary* ~ *inculper/accuser* + *vol*, *charge* + *connection* ~ *inculper/accuser* + *complicité*, *charge* + *conspiracy* ~ *inculper* + *conspiration*). They were all validated by an English to French professional translator with French as a native language. The corpus also proved particularly useful to identify preferred collocational pairs and lexicogrammatical patterns in the target language as well as larger but perhaps less lexicalized phrases such as ‘*faire payer* + NOUN’.

Apart from collocations, a wide range of less salient word combinations remain largely disregarded in current lexicographic practice. Corpus-based approaches to phraseology, however, have uncovered the essential functions played in language by n-grams or lexical bundles, i.e. “recurrent expressions, regardless of their idiomaticity, and regardless of their structural status” (Biber et al., 1999: 990) such as *he is*, *is that the*, *I don’t know what*, *can I have a*, *as suggested by*, and *if it were accepted that* (see also Biber and Gray, this volume). Granger and Lefer (2012) investigate whether the n-gram method can be used to enhance the quality of English-French bilingual dictionaries. They used the French part of the Label France translation corpus, i.e. a 1 million word parallel corpus made up of French magazine articles translated into English, to extract automatically a list of c. 6,000 2-to 5-grams with a minimum frequency of 20 which they analysed manually: only complete sequences were retained and lemmatized (e.g. the sequences *dans le cadre de/du/d’/des* put under the phrasal lemma *dans le cadre de*). Granger and Lefer (2012) came up with a final list of 425 n-grams which they then checked against two online subscription-based dictionaries, i.e. Le Grand Robert & Collins version 2.0 (RC) and Hachette Oxford (HO). They found that 15% of the selected lexical bundles are absent from the French-to-English parts of both dictionaries. The majority of ‘absent’ lexical bundles are longer units of 3 to 5 words which function as discourse organizers or complex adverbs of time/space (e.g. *avant même que*, *ce*

*n'est sans doute pas, de ce point de vue*). The authors also showed that more than a quarter of the lexical bundles appear in an example with no particular highlighting, e.g. *au quotidien, loin d'être, on assiste à, sous la direction de*.

In the second part of the study, Granger and Lefer (2012) used the English part of the Label France translation corpus to identify frequent translation equivalents of two lexical bundles, i.e. *de plus en plus (de)* and *sur le plan (de)*, and compare corpus-derived translation equivalents with those found in three French-English bilingual dictionaries: RC, HO and the Larousse French-English dictionary (LA). They found that bilingual dictionaries offer two main translations for *de plus en plus*: *more and more* and 'comparative + comparative' (e.g. *hotter and hotter*). However, the most frequent equivalent in the translation corpus is *increasingly*, which is only mentioned in the LA. The translation corpus also revealed a translation equivalent which is conspicuous by its absence from the 3 dictionary entries, 'ever + comparative'. Granger and Lefer's study thus also provides compelling evidence that parallel corpora can be used to improve the number and accuracy of translation equivalents.



## Summary and critical standpoint

The treatment of phraseological units differs significantly across dictionaries both in terms of coverage and access. Co-occurrence analysis lay at the core of the pioneering COBUILD project and collocations now feature prominently in (at least) British pedagogical lexicography. By contrast, a whole range of recurrent phrases with essential discourse functions have yet to find the place they deserve in learners' dictionaries, dictionaries for native speakers and bilingual dictionaries alike. As regards access, techniques range from highlighting a restricted number of word sequences in examples to providing lists of salient collocates in collocation boxes.

There are many different types of English corpora [see Davies, this volume] but the most widely used corpus in lexicography is the large monolingual reference corpus. Today, lexicographers at Oxford, for example, have at their disposal a corpus of over 2 billion words that represent a range of material from different subject areas (e.g. business, computing, law), regions of the world (e.g. Australian English, Canadian English as well as new varieties such as Hong Kong English), and types of writing (e.g. academic papers, newspapers, novels, blogs).<sup>2</sup> Lexicographers often compare the different parts of a reference corpus so as to label the words that are more typical of informal, formal, spoken, literary, or technical English, or to flag the words that are only used in specific geographical varieties of English. By contrast, so far, the lexicographical treatment of phraseology has largely been undifferentiated: no domain, genre or register labels are attached to phraseological units in dictionaries. Corpus-based studies have however shown that different genres and text types are characterised by different phraseological profiles (e.g. Biber and Conrad 1999; Luzón Marco 2000; Gledhill, 2000).

While the large monolingual reference corpus is an extraordinary source of lexicographic data, other types of corpora certainly deserve a more prominent place on the lexicographer's computer: specialised corpora, parallel corpora, and learner corpora. As regards the use of learner corpora, they certainly have a major role to play in the prevention of phraseological errors. Learner corpus research has revealed the huge impact of the first language on the learner phrasicon (see Paquot and Granger, 2012 for a review of studies of formulaic language in learner corpora) and, until very recently, L1-orientation was still considered as unrealistic from the perspective of publishers (cf. Gilquin et al., 2007). Recent research has however illustrated the clear necessity to adapt dictionaries to users' needs and with the advent of electronic dictionaries, L1-orientation appears to be one way forward to answer the call for more customisation of lexicographic data (cf. De Schryver, 2003: 182-185).

In the words of Granger (2012:2), "The innovations afforded by the electronic medium can radically transform every facet of dictionary design and use" (see Granger and Paquot, 2012 for a discussion of significant innovations in electronic lexicography). In the area of phraseology, however, publishers have at best used the opportunities opened up by the shift of dictionaries towards the electronic so as to offer more collocations and phrases and improve search options to find these word combinations.

---

<sup>2</sup> <http://oxforddictionaries.com/words/about-the-oxford-english-corpus>

### 3) Towards a genre-based approach to the lexicographical treatment of phraseology in electronic monolingual learners' dictionaries

Corpus-based studies have highlighted the crucial role of recurrent word combinations such as *prime example*, *final point*, *noted above*, *worth noting*, *as follows*, *as a result*, *the evidence suggests that*, *it is possible to*, or *in the case of* in academic texts (e.g. Curado Fuentes, 2001; Biber et al, 2004; Biber & Barbieri, 2007; Pecman, 2008). They have generally supported Gledhill's view that "there is a shared scientific voice or 'phraseological accent' which leads much technical writing to polarise around a number of stock phrases" (Gledhill 2000: 204). As put by Nation (2001: 178), 'vocabulary choice is a strong indicator of whether the writer has adopted the conventions of the relevant discourse community'. Mastery of these word combinations is therefore particularly important, especially for the large proportion of students and researchers for whom English is a non-native language. A growing number of university students have to write term papers, reports or their MA/PhD dissertations in English. For researchers worldwide, the stakes are even higher as they need to write and publish in English to achieve international recognition in their field.

Electronic monolingual learners' dictionaries today often use as a selling point the fact that they provide a variety of resources to help users produce written texts, especially in academic or professional settings. These features typically include a focus on academic words, vocabulary expansion material, special notes for helping learners avoid common errors, mid-matter sections focusing on specific discourse functions, exercises to enable users to practise what they have learned as well as collocation boxes. The study reported on here assesses the usefulness of collocation boxes for academic writing in the latest editions of the 'Big Five', i.e. the five major electronic monolingual learners' dictionaries of English:

1. *Cambridge Advanced Learner's Dictionary* (3<sup>rd</sup> edition, 2008) [CALD3]
2. *Oxford Advanced Learner's Dictionary* (8<sup>th</sup> edition, 2010) [OALD8]
3. *Collins-Cobuild Advanced Learner's English Dictionary* (6<sup>th</sup> edition, 2009) [CCAD6]
4. *Macmillan English dictionary for advanced learners* (2<sup>nd</sup> edition, 2007) [MEDAL2]
5. *Longman Dictionary of Contemporary English* (5<sup>th</sup> edition, 2009) [LDOCE5]

It focuses on collocation boxes for ten high-frequency verbs – *argue*, *demonstrate*, *illustrate*, *imply*, *indicate*, *prove*, *reveal*, *show*, *suggest*, and *support* – that centre around the production of knowledge in the process of academic investigation (cf. Meyer, 1997). Their pervasiveness in academic discourse is further supported by the fact that they belong to Paquot's (2010) *Academic Keyword List* (AKL). This is a list of 930 potential academic words, i.e. "words that are reasonably frequent in a wide range of academic texts but relatively uncommon in other kinds of texts and which, as such, might be used to refer to those activities that characterize academic work, organize scientific discourse and build the rhetoric of academic texts, and so be granted the status of academic vocabulary" (Paquot, 2010: 29).<sup>3</sup>

I refer to the ten verbs as 'verbs of evidence' as they enable writers to show that a phenomenon or fact (e.g. the data or research mentioned) constitutes or has produced evidence for something.

In terms of quantity and access route, the treatment of collocations varies considerably across the five learners' dictionaries. In OALD8 and CALD3, there is no collocation box for verbs of evidence: a limited number of collocations and phraseological units are highlighted in bold in example sentences. As shown in Table 1, CCAD6 provides access to a selected list

---

<sup>3</sup> Unlike Coxhead's (2000) *Academic Word List*, the *Academic Keyword List* includes the 2,000 most frequent words in English, thus making it possible to appreciate the paramount importance of core English words in academic prose. The *Academic Keyword List* is available from <http://www.uclouvain.be/en-372126.html>

of collocates for the verbs *argue*, *imply*, *indicate*, *prove* and *suggest* while MEDAL2 offers collocation boxes for *argue*, *show*, *suggest* and *support*. LDOCE5 is the only dictionary that systematically provides collocation boxes for all lexical entries. For each word, users can generally check: (1) a list of all the collocations that appear in the entry (generally in bold in examples), (2) a list of all the collocations that include that particular word but in other entries and (3) a list of corpus-derived collocations for that particular word (Figure 3).

**Table 1:** Collocation boxes for verbs of evidence in the Big Five

|             | CALD3 | OALD8 | CCAD6 | MEDAL2 | LDOCE5 |
|-------------|-------|-------|-------|--------|--------|
| argue       |       |       | √     | √      | √      |
| demonstrate |       |       |       |        | √      |
| illustrate  |       |       |       |        | √      |
| imply       |       |       | √     |        | √      |
| indicate    |       |       | √     |        | √      |
| prove       |       |       | √     |        | √      |
| reveal      |       |       |       |        | √      |
| show        |       |       |       | √      | √      |
| suggest     |       |       | √     | √      | √      |
| support     |       |       |       | √      | √      |



**Figure 3:** The verb ‘suggest’ in LDOCE5

In the electronic age, dictionary lookup is often described as an information retrieval activity (Bothma, 2011; Heid, 2011; Lorentzen & Theilgaard, 2012). In information retrieval, retrieval effectiveness has commonly been assessed by means of the measures of *recall* and *precision* (Salton, 1989; Ponte & Croft, 1998; Shafi & Rather, 2005). Recall is commonly defined as the proportion of relevant materials retrieved and precision is the proportion of retrieved materials that are relevant. Applied to dictionary lookup, and more specifically to the purposes of finding collocations in monolingual learners’ dictionaries, *recall* may be defined as the proportion of relevant collocations retrieved in a specific dictionary and computed as follows:

$$\text{Recall rate of relevant collocations} = \frac{\text{Relevant collocations found in dictionary}}{\text{Relevant collocations as found in corpus}}$$

*Precision* is the proportion of collocations listed in a specific dictionary that are relevant:

$$\text{Precision rate of relevant collocations} = \frac{\text{Relevant collocations found in dictionary}}{\text{Collocations found in dictionary}}$$

To investigate the usefulness for academic writing of the collocation boxes available in CCAD6, MEDAL2 and LDOCE5, I assessed the recall and precision rates of the ten most typical academic collocations of each verb of evidence as found in the *Corpus of Academic Journal Articles* (CAJA), i.e. a 90 million word corpus of research articles published in peer-reviewed journals (Kosem, 2010). I used the CAJA as the standard against which to evaluate the collocations reported in the learners' dictionaries as (1) it is the largest corpus of academic writing available, (2) it was compiled in recent years, (3) it represents a wide variety of academic disciplines and (4) unlike the academic component of the BNC, for example, it only includes full texts - not samples. As such, it is deemed to be a unique model of academic writing.

Using the Sketch Engine, I extracted Word Sketches for the verbs *argue*, *demonstrate*, *illustrate*, *imply*, *indicate*, *prove*, *reveal*, *show*, *suggest*, and *support*. To operationalize the concept of 'relevant collocations', collocates were ordered by decreasing frequency to identify the ten most frequent 'general' academic collocations rather than by statistical score as the latter option retrieved too many discipline-specific collocations (e.g. *prove* + *theorem*, *plaintiff* + *prove*). The analysis was restricted to collocates in subject and object positions as these two syntactic relations are clearly identifiable in CCAD6, MEDAL2 and LDOCE5, and thus comparable between the Word Sketches and the three dictionaries.

Recall and precision rates were computed for each verb in CCAD6, MEDAL2 and LDOCE5. Recall rates range from 10% for the verb 'prove' in LDOCE5 to 65% for the verb *illustrate* in the same dictionary, with a mean of 29.76%. In the case of the verb *support*, for example, only 7 collocates out of the 20 academic collocates as found in the CAJA are listed in LDOCE5. As shown in Table 2, the recall rate is thus 35%. No figures are given for COBUILD as there is no collocation box under the entry for the verb *support*. MEDAL2 only lists 6 collocates in the verb-object relation and its recall rate is thus of 30%. When the two syntactic relations are considered separately, the minimum recall rate drops to 0% (for the verb *prove* in a subject-verb relation in LDOCE5) and the maximum recall rate reaches 90% (for the verb *show* in a subject-verb relation in LDOCE5).

**Table 2: Recall rates for the collocates of the verb ‘support’ in LDOCE5, CCAD6 and MEDAL2**

| CAJA collocates                                                            | LDOCE5 | CCAD6          | MEDAL2 |
|----------------------------------------------------------------------------|--------|----------------|--------|
| NP (= object)                                                              |        |                |        |
| <i>argument</i>                                                            |        |                | √      |
| <i>claim</i>                                                               | √      |                | √      |
| <i>conclusion</i>                                                          | √      |                | √      |
| <i>finding, model, role</i>                                                |        |                |        |
| <i>hypothesis</i>                                                          | √      |                | √      |
| <i>idea</i>                                                                |        |                | √      |
| <i>notion</i>                                                              | √      |                |        |
| <i>view</i>                                                                | √      |                | √      |
| Subj_NP                                                                    |        |                |        |
| <i>analysis, argument, data, fact, finding, observation, result, study</i> |        |                |        |
| <i>evidence</i>                                                            | √      |                |        |
| <i>research</i>                                                            | √      |                |        |
| Total                                                                      | 7/20   |                | 6/20   |
| Recall                                                                     | (35%)  | not applicable | (30%)  |

MEDAL2 does not provide many collocation boxes for verbs of evidence but when it does, recall is often relatively good, as the dictionary innovates by offering collocation boxes at the level of the word sense, rather than for the word in general. For example, the collocation box under Sense 4 of the verb *support* (‘to show that an idea, statement, theory etc is true or correct’) lists eight abstract nouns that are frequently used as objects of the verb in academic texts: *argument, claim, conclusion, contention, hypothesis, idea, theory* and *view*. Among those nouns, six are typical academic collocations as found in CAJA (cf. Table 2).

Precision rates range from 6.8% for the verb *prove* in LDOCE5 to 75% for the verb *support* in MEDAL2 (Table 3). Academic collocations are often listed together with a number of collocations that are more typical of other text types and genres. In LDOCE5, for example, the 31 collocates of the verb *support* only include 6 highly frequent general academic collocates (*claim, evidence, hypothesis, idea, research* and *view*) together with nouns such as *accusation, cause, charity, effort, event, family, government, team, weight*, etc. Precision is particularly low in LDOCE5 with a mean of 23.9%.<sup>4</sup> The list of ‘collocations from other entries’ proves particularly problematic: different uses of the headword are juxtaposed and users are sometimes forced to scan through lists that may consist of more than 100 collocations (in the case of the lemma *show* for example) to find the appropriate collocation. In addition, verb and noun uses of lemmas such as *show* and *support* are listed together. Precision rates were not computed for the verbs *imply* and *prove* in CCAD6 and the verbs *argue, show* and *suggest* in MEDAL2. The collocation boxes for these verbs in the two dictionaries do not list nouns but mainly adverbs. There is a collocation box with object nouns for the verb *show* in MEDAL2 but it appears under a sense that is clearly not academic-like, i.e. ‘to behave in a way that allows people to know your feelings, opinions, or personal qualities’, and was therefore not included in the analysis.

<sup>4</sup> The policy of LDOCE5 to divide collocations into different lists on the basis of the source of information makes the collocation lists become in part repetitive (see Götz-Votteler and Herbst 2009 for a critical overview). Collocates that are repeated in the different collocation lists were however only counted once.

**Table 3: Precision rates for subject and object collocations of verbs of evidence in LDOCE5, CCAD6 and MEDAL2**

|                    | LDOCE5 |       |           | CCAD6 |       |           | MEDAL2 |       |           |
|--------------------|--------|-------|-----------|-------|-------|-----------|--------|-------|-----------|
|                    | CAJA   | Total | Precision | CAJA  | Total | Precision | CAJA   | Total | Precision |
| <i>argue</i>       | 4      | 7     | 57.1      | 2     | 4     | 50        | 0      | 0     |           |
| <i>demonstrate</i> | 6      | 19    | 31.6      |       |       |           |        |       |           |
| <i>illustrate</i>  | 13     | 26    | 50.0      |       |       |           |        |       |           |
| <i>imply</i>       | 5      | 7     | 71.4      | -     | -     |           |        |       |           |
| <i>indicate</i>    | 9      | 17    | 52.9      | 3     | 8     | 37.5      |        |       |           |
| <i>prove</i>       | 3      | 44    | 6.8       | -     | -     |           |        |       |           |
| <i>reveal</i>      | 5      | 16    | 31.3      |       |       |           |        |       |           |
| <i>show</i>        | 11     | 99    | 11.1      |       |       |           | -      | -     |           |
| <i>suggest</i>     | 6      | 19    | 31.6      | 4     | 8     | 50        | -      | -     |           |
| <i>support</i>     | 6      | 31    | 19.4      |       |       |           | 6      | 8     | 75        |

The results of a precision and recall analysis of the collocations included in a dictionary are crucially dependent on the definition of ‘relevant collocations’ put forward and may even vary according to (the quality of) the corpus data used to identify the relevant collocations. Recall and precision nevertheless prove particularly instructive tools for measuring how well a dictionary answers specific users’ needs.

Precision is most probably also an appropriate measure to quantify what has been variously called ‘information stress’ or ‘information overload’ in the literature. As put by Tarp (2009: 26), “a major problem in the present information age is not the absence of data from which the needed information can be retrieved, but the abundance of unstructured data.” Quite clearly, collocations in learners’ dictionaries generally are unstructured data. They cover a wide range of meanings that are often characteristic of different genres and text types. The list of collocations for the verb *show* in LDOCE5, for example, includes word combinations and senses as varied as *show sth’s limitations*, *show compassion*, *show a disposition to do sth*, *show a correlation*, *show a desire*, *show contempt*, *show mastery*, *show symptoms*, and *show your amusement*. More importantly perhaps, academic collocations are listed along with other word combinations that are not appropriate for academic writing. In MEDAL2, for example, there is a collocation box for the verb *argue*, sense 2 (“to give reasons why you believe that something is right or true”), a sense that is also quite common in academic writing, which offers the following list of adverb collocates: *consistently*, *convincingly*, *forcefully*, *passionately*, *persuasively*, *plausibly*, and *strongly*. Non-native writers can be seriously misled by this presentation of collocations as they are not given any help in deciding which word combinations are most appropriate in academic writing. Put differently, the treatment of phraseology in electronic learners’ dictionaries may lead non-native writers to believe that all collocations and phrases are suitable for all purposes (e.g. writing a research article or a short story, writing a letter to a friend or to a human resources manager). This echoes a question that was well-formulated by Williams in 2006: “Learner’s dictionaries are made for learners, but who are the learners in question?” To really help EFL learners write in English, future dictionaries must show greater awareness of the many different types of writing, genres and styles. Not only does this hold for the phraseological description of words, but also for other features of the microstructure such as sense ordering and example sentences (cf. Paquot, 2012: 165-166).

#### 4) Conclusion

The use of corpus data has considerably improved the coverage of phraseology in electronic dictionaries, but its lexicographical treatment has “still not found an adequate balance between the parameters of quantity and quality” (Götz-Votteler and Herbst, 2009: 57). In fact, much remains to be done. In terms of quantity, for example, the use of corpus-derived collocation boxes needs to be systematized. Today, they are often restricted to a limited set of highly frequent nouns or verbs. When available, collocation boxes are devoted to a small class of morpho-syntactic relationships such as ‘adjective + noun’ or ‘verb + noun’. Word Sketches provide different lists of collocates for the different patterns of a word (e.g. *argue* + *case*, *point* vs. *argue* + *for* + *importance*, *existence*, *view* or *suggest* + Ving: *using*, *adding*, *considering* vs. *suggest* + *as* + *cause*, *explanation*, *factor*) and these combinations should also make their way into dictionaries. Ideally, phrase banks should also feature more prominently in electronic dictionaries to address the current paucity of lexical bundles, especially where they are cohesive markers that fulfil a range of functions in writing (cf. Granger and Lefer, 2013: 163).

In terms of quality, a wider range of multi-word expressions certainly deserves to be granted headword status (cf. Heid and Gouws, 2006) and we are still very far away from Sinclair’s ‘ultimate dictionary’ that would contain “all the lexical items of a language, each one in its canonical form with a list of possible variations” (Sinclair, 2004: xxiv). It is also essential that the lexicographical treatment of collocations and other phraseological units be context sensitive. The *Macmillan English Dictionary for Advanced Learners* (1<sup>st</sup> edition, 2002) broke new ground by providing sense-differentiated collocations. Context, however, also includes other aspects of ‘collocational normality’. Fifteen years ago, Partington (1998: 17) wrote that “collocational normality is dependent on genre, register and style i.e. what is normal in one kind of text may be quite unusual in another”. This significant statement has not yet found an echo in commercial lexicography. It has, however, recently been put into practice in a number of innovative corpus-based academic lexicographical projects. The Louvain English for Academic Purposes Dictionary (LEAD), for example, is an integrated dictionary and corpus tool intended to help non-native speakers write academic texts in English (cf. Granger and Paquot, 2010; Paquot, 2012). Context-sensitivity is addressed by selective use of corpus data in the LEAD: according to users’ profiles, collocations and lexical bundles are linked to relevant concordance lines in discipline-specific corpora. As the corpus-query tool is fully integrated into the LEAD, users also have direct access to these specialised corpora and can therefore “participate in the social activity of negotiating meanings in a committed and informed way” (Teubert 2001 151-2).

The field of lexicography has been very much part and parcel of the corpus revolution and very few would argue today against the statement that, unlike many contemporary revolutions, it has been a real success story. The part played by corpora has become increasingly important and “no serious compiler would undertake a large dictionary project nowadays without one (and preferably several) at hand” (De Schryver, 2003: 167). Like many other fields, lexicography is now witnessing a second and extraordinarily rapid turnabout with the ‘Internet revolution’. The consequences for dictionaries and dictionary making are unprecedented and a discussion clearly falls beyond the scope of this paper (see Granger and Paquot, 2012). However, I would like to pinpoint two fundamental characteristics of the Web 2.0 that will undoubtedly have an impact on the role of corpus data in future lexicography. First, given the potential levels of connectivity and interoperability now available, it is to be expected (and desired) that future online dictionaries will feature more integration of corpus-query tools (cf. Asmussen, forthcoming). Users should be able to navigate to and from and use the best resource to answer each of their specific linguistic needs. To search for the most appropriate collocation, for example, a simplified and more user-friendly version of a Word Sketch could well fit the bill. Second, in the quest for more user-oriented and context-

sensitive data in lexicography, there is scope for new forms of online collaboration. The time is ripe for a dictionary-cum-corpus platform where users can upload their own corpora, so as to visualise patterns of word use in a context which reflects their individual field of interest,



## References

### Dictionaries

- [CALD3] Cambridge Advanced Learner's Dictionary (3rd edition), 2008.  
[CCAD6] Collins-Cobuild Advanced Learner's English Dictionary (6th edition), 2009.  
[LDOCE5] Longman Dictionary of Contemporary English (5th edition), 2009.  
[MEDAL2] Macmillan English dictionary for advanced learners (2nd edition), 2007.  
[OALD8] Oxford Advanced Learner's Dictionary (8th edition) 2010.

### Web resources

|                                     |                                                                                     |
|-------------------------------------|-------------------------------------------------------------------------------------|
| DANTE                               | <a href="http://www.webdante.com/index.html">http://www.webdante.com/index.html</a> |
| Pattern Dictionary of English Verbs | <a href="http://deb.fi.muni.cz/pdev/">http://deb.fi.muni.cz/pdev/</a>               |
| Sketch Engine                       | <a href="http://www.sketchengine.co.uk/">http://www.sketchengine.co.uk/</a>         |

### Other literature

- Asmussen, J. Forthcoming. Combined products: dictionary and corpus. In R. H. Gouws, U. Heid, W. Schweickard and H. E. Wiegand (eds.), *Dictionaries. An international encyclopedia of lexicography*. Volume 5.4 (suppl.). Berlin and New York: Walter de Gruyter.
- Biber, D., and Barbieri, F. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26:263-86.
- Biber, D. and Conrad, S. 1999. Lexical bundles in conversation and academic prose. In H. Hasselgård and S. Oksefjell (eds.), *Out of corpora: Studies in honour of Stig Johansson*, 181-190. Amsterdam: Rodopi.
- Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.
- Biber, D., Conrad, S. and Cortes, V. 2004. *If you look at ....: Lexical bundles in university teaching and textbooks*. *Applied Linguistics* 25(3): 371-405.
- Bothma, T. J. D. 2011. Filtering and adapting data and information in an online environment in response to user needs. In P. A. Fuertes-Olivera and H. Bergenholtz (eds.), 71-102.
- Coffey, S. 2006. 'Delexical verb + noun' phrases in monolingual English learners' dictionaries. *Proceedings of the XII EURALEX Congress*. Downloaded from [http://www.euralex.org/elx\\_proceedings/Euralex2006/](http://www.euralex.org/elx_proceedings/Euralex2006/)
- Coxhead, A. 2000. A new academic word list. *TESOL Quarterly* 34(2): 213-38.
- Curado Fuentes, A. 2001. Lexical behaviour in academic and technical corpora: implications for ESP development. *Language Learning and Technology* 5(3):106-29.
- De Cock, S. 2002. Pragmatic prefabs in learners' dictionaries. *Proceedings of the X EURALEX Congress*. Downloaded from [http://www.euralex.org/elx\\_proceedings/Euralex2002/](http://www.euralex.org/elx_proceedings/Euralex2002/)
- De Cock, S. and Granger, S. 2004. High frequency words: the bête noire of lexicographers and learners alike. A close look at the verb 'make' in five monolingual learners' dictionaries of English. *Proceedings of the XI EURALEX Congress*. Downloaded from [http://www.euralex.org/elx\\_proceedings/Euralex2004/](http://www.euralex.org/elx_proceedings/Euralex2004/)
- De Cock, S. and Granger, S. 2005. Computer learner corpora and monolingual learners' dictionaries: the perfect match. *Lexicographica* 20: 72-86.

- de Schryver, G.-M. 2003. Lexicographer's dreams in the electronic-dictionary age. *International Journal of Lexicography* 16(2): 143-99.
- Ferraresi, A., Bernardini, S., Picci, G. and Baroni, M. 2010. Web corpora for bilingual lexicography: A pilot study of English/French collocation extraction and translation. In R. Xiao (ed.), *Using corpora in contrastive and translation studies*, 337-59. Newcastle upon Tyne: Cambridge Scholars.
- Fuertes-Olivera, P. A. and Bergenholtz, H. (eds.) 2011. *e-Lexicography: the Internet, digital initiatives and lexicography*. London and New York: Continuum.
- Gilquin, G., Granger, S. and Paquot, M. 2007. Learner Corpora: the missing link in EAP pedagogy. *Journal of English for Academic Purposes* 6(4): 319-35.
- Gledhill, C. 2000. *Collocations in science writing*. Tuebingen: Gunter Narr Verlag.
- Götz-Votteler, K. and Herbst, T. 2009. Innovation in advanced learner's dictionaries of English. *Lexicographica* 25, 47-66.
- Granger, S. 2012. Introduction: Electronic lexicography – from challenge to opportunity. In S. Granger and M. Paquot (eds), 1-11.
- Granger, S. and Lefer, M.-A. 2012. Towards more and better phrasal entries in bilingual dictionaries. *Proceedings of the XV EURALEX Congress*. Downloaded from [http://www.euralex.org/proceedings-toc/euralex\\_2012/](http://www.euralex.org/proceedings-toc/euralex_2012/)
- Granger, S. and Meunier, F. (eds.) 2008. *Phraseology: an interdisciplinary perspective*. Amsterdam and Philadelphia: Benjamins.
- Granger, S. and Paquot, M. 2008. Disentangling the phraseological web. In S. Granger and F. Meunier (eds.), 27-49.
- Granger, S. and Paquot, M. 2010. Customising a general EAP dictionary to meet learner needs. In S. Granger and Paquot, M. (eds.), *eLexicography in the 21st century: New challenges, new applications*, 87-96. Louvain-la-Neuve, Presses universitaires de Louvain.
- Granger, S. and Paquot, M. (eds) 2012. *Electronic lexicography*. Oxford: Oxford University Press.
- Hanks, P. 1987. Definitions and Explanations. In J. Sinclair (ed.), *Looking up: an account of the COBUILD project in lexical computing*, 116-36. London and Glasgow: Collins.
- Hanks, P. 2009. The impact of corpora on dictionaries. In Baker, P. (ed.), *Contemporary corpus linguistics*, 214-36. London and New York: Continuum.
- Hanks, P. 2012a. The corpus revolution in lexicography. *International Journal of Lexicography* 25(4): 398-436.
- Hanks, P. 2012b. Corpus evidence and electronic lexicography. In S. Granger and M. Paquot (eds.), 57-82.
- Hanks, P. 2013. *Lexical analysis: norms and exploitations*. Cambridge, MA: MIT Press.
- Heid, U. 2011. Electronic dictionaries as tools: Towards an assessment of usability. In P. A. Fuertes-Olivera and H. Bergenholtz (eds.), 287-304.
- Heid, U. and Gouws, R. 2006. A model for a multifunctional electronic dictionary of collocations. *Proceedings of the XII EURALEX Congress*. Downloaded from [http://www.euralex.org/elx\\_proceedings/Euralex2006/](http://www.euralex.org/elx_proceedings/Euralex2006/)
- Herbst, T. and Mittmann, B. 2008. Collocation in English dictionaries at the beginning of the twenty-first century. *Lexicographica. International Annual for Lexicography*, 103-19.
- Kilgarriff, A. and Kosem, I. 2012. Corpus tools for lexicographers. In Granger, S. and Paquot, M. (eds), 31-56.
- Kosem, I. 2010. Designing a model for a corpus-driven dictionary of academic English. Unpublished PhD thesis. Aston University, Birmingham, UK. Downloaded from <http://eprints.aston.ac.uk/14664/>

- Lew, R. 2012. The role of syntactic class, frequency, and word order in looking up English multi-word expressions. *Lexikos* 22: 243-60.
- Lorentzen, H. and Theilgaard, L. 2012. Online dictionaries – how do users find them and what do they do once they have? *Proceedings of the XV EURALEX Congress*. Downloaded from [http://www.euralex.org/proceedings-toc/euralex\\_2012/](http://www.euralex.org/proceedings-toc/euralex_2012/)
- Luzón Marco, M.J. 2000. Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes* 19(1): 63-86.
- Meunier, F. and Granger, S. (eds.) 2008. *Phraseology in foreign language learning and teaching*. Amsterdam and Philadelphia: Benjamins.
- Meyer, P.G. 1997. *Coming to know: studies in the lexical semantics and pragmatics of academic English*. Tübingen: Gunter Narr Verlag Tübingen.
- Moon, R. 2007. Sinclair, lexicography, and the Cobuild Project: the application of theory. *International Journal of Corpus Linguistics*, 12(2): 159-81.
- Moon, R. 2008a. Sinclair, phraseology, and lexicography. *International Journal of Lexicography*, 21(3): 243-254.
- Moon, R. 2008b. Dictionaries and collocation. In S. Granger and F. Meunier (eds.), 313-36.
- Nation, P. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Paquot, M. 2010. *Academic vocabulary in learner writing: From extraction to analysis*. London and New-York: Continuum.
- Paquot, M. 2012. Automated customization or how to bring electronic dictionaries closer to their users. In Granger, S. and Paquot, M. (eds.), 163-186.
- Paquot, M. and Granger, S. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32: 130-49.
- Partington, A. 1998. *Patterns and meanings. Using corpora for English language research and teaching*. Amsterdam and Philadelphia: Benjamins.
- Pawley, A. and Syder, F.H. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. C. Richards and W. Schmidt (eds.), *Language and communication*, 29-59. London and New-York: Longman.
- Pecman, M. 2008. Compilation, formalisation and presentation of bilingual phraseology: problems and possible solutions. In F. Meunier and S. Granger (eds.), 203-222.
- Ponte, J. M. and Croft, W. B. 1998. A language modeling approach to information retrieval. *Computer and Information Science*, 98(3): 275-281.
- Rundell, M. 1999. Dictionary use in production. *International Journal of Lexicography*, 12(1): 35-53.
- Salton, G. 1989. *Automatic text processing*. Reading, MA: Addison-Wesley.
- Seretan, V. 2011. *Syntax-based collocation extraction*. Berlin: Springer.
- Shafi, S. M. and Rather, R. A. 2005. Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology. *Webology* 2(2). Downloaded from <http://www.webology.org/2005/v2n2/a12.html>
- Siepmann, D. 2008. Phraseology in learners' dictionaries: what, where and how? In F. Meunier and S. Granger (eds.), 185-202.
- Sinclair, J. 2004. Interview with John Sinclair conducted by Wolfgang Teubert. In R. Krishnamurthy (ed.), *English collocation studies: The OSTI report*, xvii-xxix. London and New York: Continuum.
- Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Stubbs, M. 1993. British traditions in text analysis: from Firth to Sinclair. In M. Baker, G. Francis and E. Tognini-Bonelli (eds.), *Text and technology: in honour of John Sinclair*, 1-33. Philadelphia and Amsterdam: Benjamins.

- Tarp, S. 2009. Beyond lexicography: new visions and challenges in the information age. In H. Bergenholtz, S. Nielsen and S. Tarp (eds.), *Lexicography at a crossroads. Dictionaries and encyclopedias today, lexicographical tools tomorrow*, 17-32. Bern: Peter Lang.
- Teubert, W. 2001. Corpus linguistics and lexicography. *International Journal of Corpus Linguistics* 6: 125-53.
- Teubert, W. 2004. Corpus linguistics and lexicography: The beginning of a beautiful friendship. *Lexicographica. International Annual for Lexicography*, 1-19.
- Walker, C. 2009. The treatment of collocation by learners' dictionaries, collocational dictionaries and dictionaries of business English. *International Journal of Lexicography*, 22(3): 281-299.
- Williams, G. 2006. Advanced ESP and the learner's dictionary: tools for the non-language specialist. *Proceedings of the XII EURALEX Congress*. Available from [http://www.euralex.org/elx\\_proceedings/Euralex2006/](http://www.euralex.org/elx_proceedings/Euralex2006/)