Stealing a march on collocation

Deriving extended collocations from full text for student analysis and synthesis

James Thomas
Faculty of Arts, Masaryk University

Full text affords language learners many opportunities to observe a wide range of linguistic features whose typicality they can ascertain through corpus searches. The particular features investigated in this chapter revolve around the collocations of key words in texts. Given that knowing a collocation in no way guarantees its correct use, a procedure referred to as Collocation Plus has been developed in which learners explore the lexical and grammatical environments of collocations in the contexts in which they meet them. This is an important process in making receptive vocabulary productive. Learners may then formalise their findings into 'word templates' which are then available for production. This work combines some recent findings in linguistics, language acquisition and pedagogy to help learners produce language that is more accurate, fluent, idiomatic and sophisticated, whilst developing their autonomy in using the resources available and raising their consciousness of the processes involved.

Keywords: two-lexeme collocation; word template; Collocation Plus; topic trail; guided discovery; Sketch Engine; Hoey procedure

1. Introduction and overview

The author of the following sentence is an advanced researcher in computer science who is exposed to a great deal of text in his field in English.

1. *This process has to be carefully managed in order to prevent the violation user's privacy and to protect the community to be overburden of such questions.

While the sentence is not without its merits, the deviant uses of *prevent, protect* and *overburden*, in particular, suggest that the writer is constructing sentences by dropping words into syntactic slots and that his language study has paid little

DOI 10.1075/scl.69.05tho

attention to words' individual properties and how they interact. A remedy for this is the subject of this paper.

In a subcorpus of our Informatics Reading Corpus (IRC) (see Appendix for information about all corpora mentioned), which contains texts from this scientist's specific field only, *prevent* occurs 50 times and *protect* 31 times and all occur in their patterns of normal usage. Exposure to correct usage is clearly not enough to motivate a change in this scientist's English language behaviour: something more is needed to destabilise his interlanguage and move it closer to the desired target. There are dozens of such deviations in his 8,000-word paper. The word *deviation* is being used here partly to avoid the controversial distinctions in language pedagogy between *mistake*, *error*, *slip* and other terms (see Bartram & Walton 1991: 20), and partly to support the notion of *pattern of normal usage*, which is germane to this paper (cf. Hanks 2013).

Collocation is the construct around which this chapter revolves, and given the range of uses this term is put to, considerable space is devoted to explaining precisely what it means here and how a narrow definition is put into the service of language education. The paper then proceeds to 'steal a march' on collocation, an allusion to gaining an advantage from a situation.[1] After defining collocation, the concept is then extended linearly (syntagmatically) to include other typical co-occurring words, which is referred to as Collocation Plus. When these words are then clustered into semantic groups and formalized, 'word templates' start to evolve. The chapter then describes how learners can derive them from full text. While one aim is to obviate such deviant sentences as example (1) above, students with appropriate learning styles are inducted into a procedure offering a wide range of learning opportunities – such are its affordances.

Before launching into any of the above, it would be appropriate to consider how linguistic evidence from corpus studies overlaps with language acquisition studies to evolve new priorities and new attitudes in language education. This is a reference to the title of a chapter by Sinclair (2004): New Evidence, New Priorities, New Attitudes. Let us begin with *affordance*.

The notion of affordance takes an ecological view of language. When Gibson coined the term in this sense in 1979 (see Van Lier 2000), he described an affordance as the reciprocal relationship that exists between an organism and its environment. Van Lier's (2000: 252) application of this as a theory of learning extends 'relationship' to a property of the environment that affords further action. What becomes an affordance depends on what the organism does or wants, and what is useful for it. Thus the affordances of a text are the range

[1] The phrase *steal a march* occurs 18 times in the BNC.

of opportunities for the meaningful action that it affords. Engaged language learners will perceive linguistic affordances and use them for linguistic action. In many language learning situations, engagement amounts to little more than being herded through the tasks that textbooks and teachers provide. The concept of affordance is one of many borrowings from related fields employed in contemporary language education. But in the CorpusCorpus (CC), which contained the first 68 published studies that Boulton (2010) analysed in his overview of empirical DDL studies, *affordance* does not occur once. This, and the absence of many standard terms from language acquisition studies, leads one to believe that the stakeholders in pedagogical corpus work are not engaged in an important closely related field.

Conversely, there are numerous linguistic concepts, many of which have emerged or evolved through corpus studies, that rarely, if ever, appear in even the most recent teacher resource books or course books despite having much to contribute to language pedagogy. They include studies in collocation, colligation, chunks, linear unit grammar, schema, frame semantics, discourse studies, stylistics, pragmatics and ultimately the reconciliation of grammar and vocabulary. If the reasons can be traced to the lack of compatible teaching procedures, it is hoped that some of those presented in this chapter may inspire some interdisciplinary forays that can be tested in a variety of contexts.

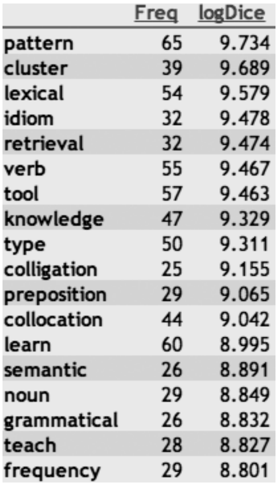
The pedagogical linguistic work done in the 1980–90s in the COBUILD project by Sinclair, Hanks, Hunston, Hoey, Krishnamurthy and many others produced language teaching resources that made possible a new lexical orientation towards language (McEnery & Hardie 2011: 79–81). Some of the work that these linguists did then can now be pursued by the current generation of internet-savvy teachers and students in guided discovery tasks, given the right tools and a new orientation towards language. This is happening regularly in my own classes and starting to appear in those of my teacher trainees.

The students who participate in various aspects of my research are well beyond the 'threshold' level at which their interlanguage is sufficiently developed to be able to operate in a classroom where the instruction is in the second language (L2), the course books are in the L2 and they are able to use monolingual learner dictionaries. These students have already learnt the most frequent uses of the most frequent vocabulary items in the language, have studied a wide variety of topics and have engaged in a great many learning activities. Furthermore, they are university students who can use their higher-order thinking skills to draw conclusions from data (see especially the revision of Bloom's taxonomy; Anderson & Krathwohl 2001). Most of the students are Czech and are studying for a Master's degree to prepare them to teach English in secondary school. There are also EFL teachers on a range of in-service courses in various parts of the world, as well as post-graduate academic writing students studying Informatics. This heterogeneous sample has been chosen to demonstrate that the approach proposed in this chapter has potential for a range of learners and teachers.

2. Sketch Engine

All of the corpus work discussed in this chapter has been undertaken by the author and his students using Sketch Engine. The pragmatic reason for this is that the software is under continuous development at Masaryk University and is entirely web-based, with over 80 preloaded corpora as well as tools to create corpora such as the above-mentioned IRC and CorpusCorpus. But there are linguistic and pedagogical reasons too why this software is preferred over the alternatives. The development of Sketch Engine has been much guided by lexicography, and this orientation is entirely compatible with the focus on lexis in contemporary language teaching (Thomas 2008). Its tools prove excellent allies in meeting Firth's (1957: 11) proclamation that "you shall know a word by the company it keeps".

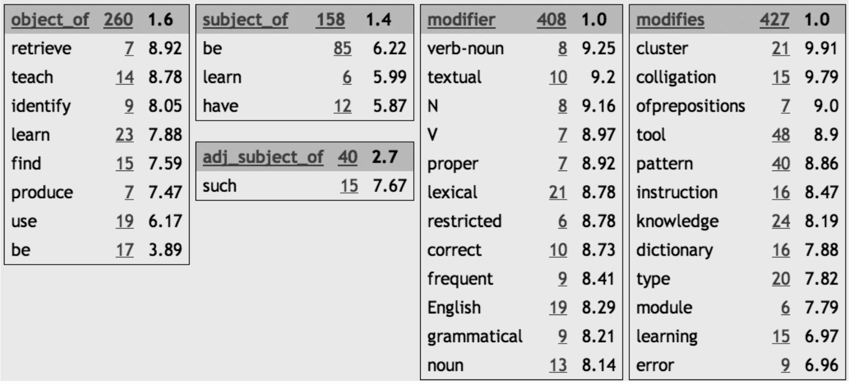
From any of its corpora, Sketch Engine generates various types of word lists such as collocations for which the user can adjust key variables, and frequency lists which can appear as strings of word forms (bundles), strings of parts of speech (syntagms), and strings that combine lemmas, word forms and parts of speech (hybrid n-grams). Central to this chapter's work on collocation is one particular collocation statistic, namely logDice, developed by Rychlý (2008), the originator of Sketch Engine. The logDice statistic generates lists of collocates where high-ranking items tend to accord with intuition (see Figure 1 for an example of the word *collocation* in the CorpusCorpus).



	Freq	logDice
pattern	65	9.734
cluster	39	9.689
lexical	54	9.579
idiom	32	9.478
retrieval	32	9.474
verb	55	9.467
tool	57	9.463
knowledge	47	9.329
type	50	9.311
colligation	25	9.155
preposition	29	9.065
collocation	44	9.042
learn	60	8.995
semantic	26	8.891
noun	29	8.849
grammatical	26	8.832
teach	28	8.827
frequency	29	8.801

Figure 1. Collocates of *collocation* in the CorpusCorpus (range –4 +4, sorted by logDice)

For this chapter, the most important tool is Sketch Engine's trademark 'word sketches', each of which is "an automatic, corpus-derived summary of a word's grammatical and collocational behaviour" (Kilgarriff et al. 2010: 372). The screenshot in Figure 2 shows the word sketch of *collocation* as it occurs in the CorpusCorpus. Each column represents a grammatical relationship ('gramrel') with the search word, and each column contains collocates of the search word in that grammatical relationship.



object_of	260	1.6
retrieve	7	8.92
teach	14	8.78
identify	9	8.05
learn	23	7.88
find	15	7.59
produce	7	7.47
use	19	6.17
be	17	3.89

subject_of	158	1.4
be	85	6.22
learn	6	5.99
have	12	5.87

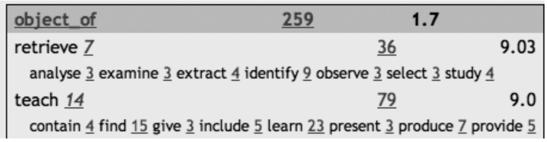
adj_subject_of	40	2.7
such	15	7.67

modifier	408	1.0
verb-noun	8	9.25
textual	10	9.2
N	8	9.16
V	7	8.97
proper	7	8.92
lexical	21	8.78
restricted	6	8.78
correct	10	8.73
frequent	9	8.41
English	19	8.29
grammatical	9	8.21
noun	13	8.14

modifies	427	1.0
cluster	21	9.91
colligation	15	9.79
ofprepositions	7	9.0
tool	48	8.9
pattern	40	8.86
instruction	16	8.47
knowledge	24	8.19
dictionary	16	7.88
type	20	7.82
module	6	7.79
learning	15	6.97
error	9	6.96

Figure 2. An extract of the word sketch of *collocation* in the CorpusCorpus sorted by significance

Furthermore, Sketch Engine includes an algorithm to generate a distributional thesaurus which lists words that occur in the same context, i.e. with the same collocations in the same grammatical relationships as the search word. Using this algorithm, the lists in a word sketch can be clustered semantically, as seen in Figure 3.



object_of	259	1.7
retrieve 7	36	9.03
analyse 3 examine 3 extract 4 identify 9 observe 3 select 3 study 4		
teach 14	79	9.0
contain 4 find 15 give 3 include 5 learn 23 present 3 produce 7 provide 5		

Figure 3. Extract from the word sketch of *collocation* with clustering turned on

Another feature of Sketch Engine is GDEX, its 'good examples' algorithm (Kilgarriff et al. 2008). This allows users to limit the number of sentences that appear at the top of a concordance page and to set parameters such as sentence length to ensure the sentences are suitable candidates for illustrative purposes. This function was initially developed for lexicographers, but it is useful in language learning as it addresses the frequent criticism that raw corpus data can be too rich and irregular to confront students with (e.g. Breyer 2009: 161).

Given these features, it seems incongruous that in the CorpusCorpus, Sketch Engine is mentioned in only eight articles – and more than twice in three only. Thomas (2015) devotes an entire book to demonstrating how using Sketch Engine to explore language questions brings forth the patterns of normal usage that are invaluable to many facets of language learning.

3. A constrained definition of collocation and its affordances

Of all the above-mentioned linguistic phenomena that have the potential to contribute to language education, the one that has the highest profile is collocation. Following the COBUILD project, it was championed in the language teaching community by Michael Lewis in his *Lexical Approach* (1993) and *Teaching Collocation: Further developments in the Lexical Approach* (2000). Giving a new priority to lexis, this approach attempted to displace sentence grammar as the organising unit for language teaching. Interestingly, its avid promotion of the teaching of collocation has outlived the approach itself (Timmis 2008).

Collocation is defined variously, with some linguists lumping general co-occurrence phenomena together, thus making the term available for a wide variety of uses from morphology to discourse studies. In one approach, Halliday and Hasan (1976: 284) referred to the use of topic-related words running through a text as collocation. *The BBI Combinatory Dictionary of English* (Benson et al. 1986), on the other hand, distinguished grammatical collocations from lexical collocations. They define the former thus:

> A grammatical collocation is a phrase consisting of a dominant word (noun, adjective, verb) and a preposition or grammatical structure such as an infinitive or clause. (Benson et al. 1986: ix)

The preferred term for this is *colligation* which Stefanowitsch and Gries (2003: 210) define as:

> The linear co-occurrence preferences and restrictions holding between specific lexical items and the word-class of the items that precede or follow them.

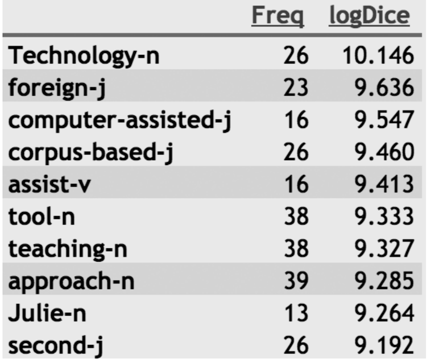
Thus pairings that consist of a lexical word plus grammar-function word (e.g. *damaging for*, *necessary to*, *agreement that*) will not be counted here as collocations. Nor will other pairings such as multi-word lexemes, where *lexeme* refers to the "smallest contrastive unit in a semantic system" (Crystal 1995: 454), be they single or multi-word units. Multi-word lexemes include phrasal verbs, delexical verbs (e.g. *give a lecture*, *do damage*), compound nouns (e.g. *case study*, *word list*, *mother tongue*) and combinations such as *nothing but* and *let alone* (cf. Jackson 1988: 11–15).

The 'two-lexeme' definition of collocation (TLC) espoused in this chapter is far too specific to permit any of these. Thus collocation here will refer to the pairing of independent lexical items only, e.g. REACH/CONSENSUS, CONSENSUS/SUPPORT, PUBLIC/CONSENSUS. These can be observed in context, whether or not directly adjacent as in:

2. This failure to reach a **public consensus** can do nothing but damage for the profession. (BNC)

Collocation does however include a combination of phrasal verbs with their typical subjects, objects and adverbs, e.g. BLOW UP/BALLOON OR STORM/BLOW UP. It certainly includes compound noun combinations with their adjectives and verbs, e.g. LACK/COMMON SENSE, CREDIT CRUNCH/BITES and EXERCISE/CASTING VOTE. Thus collocation is not a matter of two words, but of two lexemes.

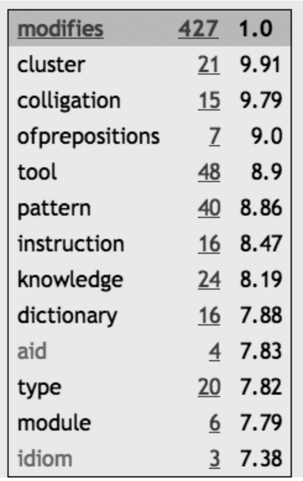
While learners find the two-lexeme criterion satisfying, they are not always able to determine how the items combine in actual usage. For example, the screenshot in Figure 4 shows the collocates of the compound noun *language learning* in the CorpusCorpus with their parts of speech (POS) indicated. Sketch Engine offers 'lempos' which, as this portmanteau term indicates, generates a list of collocates as lemmas plus their parts of speech. Learners find this particularly useful when the search word (node) is a compound itself. The most common compound noun in the CorpusCorpus is *language learning*, and one of its most frequent logDice collocates is also a compound, *corpus-based* (26 occurrences). Other hyphenated compounds include *internet-based* (7), *computer-assisted* (16) and *data-driven* (12). Single-word collocates of *language learning* include adjectives such as *effective*, *foreign* and *independent*; the nouns *application*, *concordancing* and *potential*; and the verbs *enhance*, *assist* and *integrate.*



	Freq	logDice
Technology-n	26	10.146
foreign-j	23	9.636
computer-assisted-j	16	9.547
corpus-based-j	26	9.460
assist-v	16	9.413
tool-n	38	9.333
teaching-n	38	9.327
approach-n	39	9.285
Julie-n	13	9.264
second-j	26	9.192

Figure 4. Lempos collocates of *language learning* in the CorpusCorpus

For single-word lexemes, which make up the vast majority of words we study and teach, the word sketch is preferred over collocation lists since these are unstructured and often contain words that form multi-word lexemes with the node. The structure of the word sketch, however, tellingly demonstrates the relationships between a node and its collocates: the syntactic role of the 'modifies' column is precisely that which exemplifies the words that form multi-word lexemes with the node as can be seen in Figure 5.



modifies	427	1.0
cluster	21	9.91
colligation	15	9.79
ofprepositions	7	9.0
tool	48	8.9
pattern	40	8.86
instruction	16	8.47
knowledge	24	8.19
dictionary	16	7.88
aid	4	7.83
type	20	7.82
module	6	7.79
idiom	3	7.38

Figure 5. The 'modifies' column from the word sketch of *collocation* in the CorpusCorpus sorted by significance

Another feature of collocation that appears in linguistics literature but not in pedagogical resources (as far as I am aware) is non-directionality (see Stubbs 2001: 64). The TLC definition of collocation, however, eschews non-directionality because the order in which the node and its collocates occur is determined by their syntactic roles. Thus, in a noun/verb collocation, the noun follows the verb when it is its object in an active clause, and collocating adjectives follow the noun when used predicatively. For example, a student wondering about uses of the word *strategy* can perform a word sketch in the BNC and find 6,144 modifiers (different from 'modifies'), e.g. *marketing*, *overall*, *teaching*, as well as 222 adjective subjects, e.g. *strategy appropriate*, *strategies available*. Such pre- and post-modification preferences fall within the realm of colligation and constitute patterns of normal usage.

The gramrels (syntagmatic) columns of a word sketch satisfy TLC's preference for syntax over non-directionality. Whereas a list of collocates contains no indication of the syntactic relationship between a node and its collocates, a word sketch most certainly does. For example, when students observed that the collocation

PERFORM/EXPERIMENT is mostly used in the passive, they learnt something about the words and their syntactic relationships, a feature of language they might never have considered a pattern of normal usage. At the same time, they have experienced converting data into information, the process we go through when making sense of corpus findings. Such are the affordances of guided discovery.

An obvious first step in helping learners improve is making them aware of their linguistic deviations. In a guided discovery procedure, students search corpora for the errors indicated by their peers and teachers, or more knowledgeable others (MKO) in Vygotskian terms. They find that the collocation *distantly similar* (as in (3) below) is not attested in the BNC and that the only collocating adverbs that express this notion are *remotely similar* (with 4 occurrences), and perhaps *vaguely similar* (2). The majority of adverbs preceding *similar* are boosters such as *very* (910), *remarkably* (88) and *strikingly* (20), along with hedges such as *somewhat* (76), *rather* (76) and *quite* (51). The raw frequencies, ranging from 2 to 910, are themselves telling.

3. *The Silent Way believed that there is nothing even **distantly similar** between the learning of the first and of the second language…
4. *…they are going to **pass** an **exam** tomorrow…

In Example 4, the situation needs considerably more teasing out: PASS/EXAM is a strong collocation, which is why learners are exposed to it. In fact, *pass* is both the most frequent and most significant logDice verb collocate of *exam* in the BNC with 129 occurrences in its word sketch. When this is marked as suspicious, the teacher must guide the student to examine the data more closely. In the context of PASS/EXAM, we find that *going to* only occurs twice, both in hypothetical situations – *if* and *be sure that*. And there are too few occurrences of *will* to consider this a pattern of normal usage. *Tomorrow* does not occur at all and neither do any other such adverbs of time. Where the problem is simply understanding meaning, it may be simpler, quicker and more effective just to consult a dictionary, as noted by Frankenberg-Garcia (2014); the point here, however, is that the processes of corpus consultation have their own benefits in the long term.

At issue here is the fact that a pair of lexemes does not tell the whole story of their joint behaviour. This is the motivation for helping students examine the patterns of normal usage of collocations, which is referred to here as Collocation Plus.

4. Collocation Plus (C+)

The patterns of normal usage of a collocation are determined by the words' parts of speech and by certain semantic constraints. As Hanks writes, "corpus pattern analysis shows that each word habitually participates in only a comparatively small number of patterns, and that most patterns are unambiguous in their interpretation" (2012: 54). The ambiguity that arises from the polysemy of many English words is problematic for learners (Schmitt 2010), but context disambiguates it and multiple contexts manifest the patterns which can be distilled and formalized. So when learners focus on vocabulary in extended collocations (C+), they develop their word knowledge into longer, holistic units, which primes them to co-select units of language.

We have seen that PERFORM/EXPERIMENT typically occurs in the passive, but only one of the 116 occurrences in the BNC uses *by*, contrary to some learners' expectations of the passive. Rather, *experiments BE performed* is followed either by punctuation indicating the end of a unit of information, or by one of a number of free prepositions (see below) launching prepositional phrases functioning as various types of adverbials. Further, following David Lee's (2001) classification of the BNC, 74 of the 116 occurrences are in academic texts (63%). The more or less synonymous CARRY OUT/EXPERIMENT occurs 37 times and is even more committed to the passive, but occurring less frequently in academic prose (54%) as might be expected of a phrasal verb. These extended collocations are not syntagms or bundles or multi-word units: they most closely resemble the collostructions of Stefanovitsch and Gries (2003), or the patterns proposed by Hanks (ongoing) in the *Pattern Dictionary of English Verbs* (PDEV).

Students can be guided to discover the patterns in which words occur, recording them systematically as they do so. To demonstrate this, we develop an extended collocation for the noun *scholarship* in its countable sense. Here is the word as some students came across it (Example 5):

5. A lecturer in nursing has been awarded a prestigious scholarship to undertake research into understanding what can be done to help older people who neglect to look after themselves. (Staffordshire University)

The verb/noun collocate under the microscope is AWARD/SCHOLARSHIP. The skeleton of *scholarship* can be represented as:

– someone has been awarded a scholarship to do something

Such skeletons heighten the learners' awareness of syntax as well, so we have already stolen a march on collocation. But until the semantic types are labelled, any 'someone' could be awarded a scholarship to do any 'thing'. This clearly will not do – our flights of fancy and our egalitarianism may embrace such inclusiveness, but for teaching purposes, we must insist on patterns of normal usage as a starting point. Let us now flesh out this skeleton, bearing in mind that one of the aims is to reduce the cognitive workload of remembering a great many collocations.

Our knowledge of the world tells us that an institution awards a scholarship to a student. Our knowledge of the world does not usually require us to identify the awarding institution: it is rarely found in corpus data. Sometimes the name of the scholarship implies the awarding institution, as in these two examples from the BNC:

6. In 1894 he was awarded a London county council scholarship with distinctions, which took him to the Kenmont Gardens Science School…
7. One previous winner, 14 year-old Alistair Cherry, who was awarded the Fender/Buddy Holly scholarship last year…

Similarly, scholarships are awarded to study *something*, but this is also rarely mentioned because the recipient may be known to the reader/listener and/or because the host institution or the course (i.e. the *somewhere*) is known for what it teaches, thereby invoking Grice's (1975) maxim of quantity: make your contribution as informative as is required for the current purposes of the exchange. It is included only if necessary, as in:

8. Sophie Green was awarded a scholarship to attend a summer course at Bryn Mawr College. (BNC)

When it is included, language learners are further able to focus on a *to*-infinitive clause expressing purpose. Elements of these templates that represent semantic types are enclosed in square brackets, following Hanks' PDEV formalism although somewhat simplified for pedagogical purposes:

– a scholarship is awarded [by an institution] [to a student] to study [a subject/skill] [somewhere]

Adding yet another layer of grammar focus to our vocabulary study, 21 of the 23 hits in the BNC of AWARD/SCHOLARSHIP use the passive. For this empirical reason, the template is given in the passive. In terms of the syntax exemplified by *has been awarded a scholarship*, the typicality and productivity of the syntagm can be determined by this CLAWS-based corpus query (Jakubíček et al. 2010):

– "VH." "VBN" "VVN" "AT0" "N.." (Default attribute 'tag')

This query returns 1,265 hits in the BNC (11.3 per million words), the most frequent lexical verbs in this syntagm being *give*, *offer* and *award* despite the lack of any lexical triggers. This lends support to Construction Grammar's notion that constructions have their own semantics.

With 80 hits, WIN/SCHOLARSHIP is almost four times more frequent in the BNC than BE AWARDED/SCHOLARSHIP (22 hits). So even though the learners came across the word *scholarship* in a text with the verb *award*, they came across *winning a scholarship* in their corpus searches. On the one hand, this adds to their cognitive workload, but it also enriches their semantic grasp of *win* and *scholarship*. Fifty of the 80 are followed by *to*, some as the infinitive marker – e.g. (9) and (10) – but the vast majority are free prepositions launching locative adjuncts expressing where the scholarship will be spent, as in (11), which is not the case when one is awarded a scholarship.

9. When the war was over, she **won a scholarship to** study ballet in London. (BNC)
10. No longer did a sixth former of limited means need to **win a scholarship to** go on to higher education. (BNC)
11. Frank, at the age of 16, had already **won a scholarship to** Trinity College in Cambridge. (BNC)

Thus:

- [someone] wins a *scholarship* to study [something]
- [someone] wins a *scholarship* to [an institution]

Collocation Plus gives priority to nouns because the key words in a text are typically nouns (Scott & Tribble 2006: 70). Key words are understood here as primary carriers of meaning in a text rather than statistically extracted items in the usual corpus linguistic sense (cf. Curado Fuentes, this volume). Excluding the citation, of the 16 tokens in the first sentence of this paragraph, eight are nouns including the compounds *collocation plus* and *key words*. Apart from one occurrence of *be*, there is also one delexical verb structure (*give priority*); one lexical adverb (*typically*); and the remainder are function words (*to, because, the, in, a*). For learners, nouns are more concrete than any other part of speech: course books are rich in them, as are tourist phrase books. But fluent, productive use of nouns involves co-selecting the right one with its appropriate structure.

Collocations are less needed for receptive purposes, as their meanings are by and large transparent, than for productive purposes for which students need to learn the patterns of normal usage. This is the beginning of the march we are stealing on collocation, and the reason why we have students observe them in the contexts where they meet them, as the impetus for studying *scholarship* briefly demonstrated.

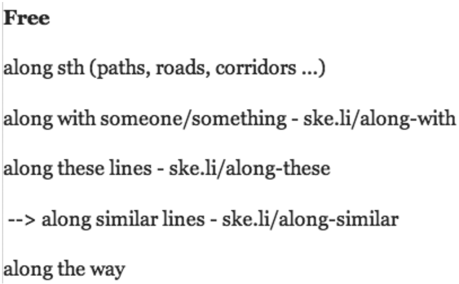
Up to this point, we have mostly focused on the lexical words and syntactic properties that constitute extended collocations. It is now time to consider the roles of prepositions in word templates. Cosme and Gilquin (2008: 259) observe that prepositions fall somewhere between grammar and the lexicon, and are often ignored by grammars and are regarded as lexically empty by lexicographers. This accounts for the linguistic difficulties teachers have teaching them and learners have learning them. As a key element in colligation, C+ procedures offer a glimmer of hope in teaching and learning prepositions.

As a starting point, the *Longman Grammar of Spoken and Written English* (Biber et al. 1999: 74ff) draws an important distinction between two types of prepositions: bound prepositions are closely related to the preceding word, invariable, semantically empty and not optional, while free prepositions are semantically full and head a prepositional phrase that functions as an adverbial. The adverbial may function as a circumstance in the Message of a clause or sentence, or as a linking adverbial in Organisational language (see 'M' and 'O' language below). For example, in the process of learning *allegation* for productive purposes, a guided discovery process led students to derive this template:

– [an injured party] makes an allegation about [a negative abstraction (potentially criminal, e.g. corruption) or a public figure]

This was corroborated by the 24 instances of MAKE/ALLEGATION/ABOUT in the BNC. In the process, it was also found that many [TALK] nouns are followed by *about*. In fact, the semantically tagged New Model Corpus (NMC – the same size as the BNC) identifies communication nouns as the most frequent nouns preceding *about*: 21,865 tokens, with approximately 75 types occurring more than 100 times. Similarly, Francis et al. (1998: 121) list 62 such types deriving from their COBUILD corpus work, including *gossip*, *lecture*, *instructions* and *prediction*. When studying words in this way, we do not try to memorise decontextualised lists of words and the prepositions that follow them; rather, we learn the whole structure including not only its prepositions, but also the semantic types of subjects and objects. These serve as mnemonics as well as exemplars of normal usage.

Similar principles and procedures apply to the adverbial uses of prepositions. Once introduced, students can observe their patterns of normal usage in corpora. Figure 6 shows how one student recorded his findings of *along* as a free preposition in our course wiki. The permalinks (starting ske.li) are direct links for Sketch Engine account holders to the data from which he drew his conclusions. Thus students who are required to fill in a gapped text with 'the correct preposition' would do well to consider if the gap is left or right facing, bound or free, respectively.



Free

along sth (paths, roads, corridors ...)

along with someone/something - ske.li/along-with

along these lines - ske.li/along-these

--> along similar lines - ske.li/along-similar

along the way

Figure 6. Extract from a student's wiki entry for *along* as a free preposition

This brings to a close the introduction to corpora, collocation and C+. The following sections propose applying C+ to key words derived from full texts.

5. Observing and using Topic Trails in full text

Contemporary language teaching resources consistently provide students with full texts to read and listen to. Full texts, whether authentic or not, provide a starting point for language learning and acquisition as they inevitably contain chunks of language that convey the propositions (messages) of the texts interweaving among chunks that structure or organise them. This division of text into message ('M' language) and organisation/orientation ('O' language) is the core of Sinclair and Mauranen's (2006) Linear Unit Grammar (LUG).

In sentence (12) from Ellis (2008: 396), the sections in bold express the propositions ('M' language), while the rest ('O' language) organises the relationships both between Messages across different levels of context, and with the reader:

12. Indeed, **every sentence is idiosyncratic**, as indeed **it is systematic**, too.

Although LUG was not developed with language learning in mind, such a straightforward division has the potential to be a great boon to authors of language teaching resources. It is not until the fourth last page of their book (p. 162) that the authors begin a brief foray into its potential for language teaching. In any case, LUG is another example of Sinclair's (2004) New Evidence, New Priorities, New Attitudes paper cited earlier. The language that expresses Organisation very often consists of fixed phrases, which in language teaching overlap with functional language.

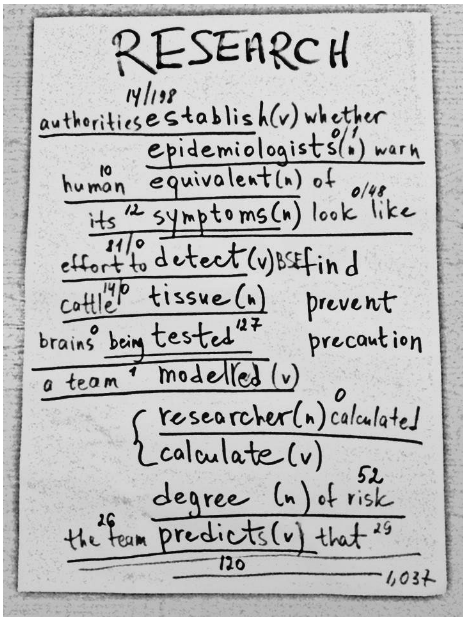
In the genre of academic prose, researchers such as Simpson-Vlach and Ellis (2010) have identified many hundreds of bundles, their "academic formulae", that are mostly 'O' language (e.g. *a function of*, *in response to*, *in this way*, *to some extent*, *on the basis of*). Given that they sought high-frequency items in a general academic corpus, it is statistically inevitable that 'O' language should be found. Students of academic prose can observe how such items are used in conjunction with the 'M' language of the texts that they work with, and, if they have access to a corpus of the texts in their field, they can observe multiple examples, arriving at evidence-based generalisations. For example, in the almost 7-million-word IRC, created by students for themselves, *in response to* occurs 78 times (11.7 per million words) and there is nothing but 'M' language either side of it. They observe that the chunk can be used in sentence-initial position (six times) and that *to* is mostly followed by noun phrases – it is an infinitive marker five times only. Charles (this volume) describes similar activities where students use self-compiled corpora to study frequently-occurring language chunks and their functions in academic discourse.

Since the primary focus of C+ is on the topic-based vocabulary derived from text, no more time need be devoted here to 'O' language. Of interest to C+ are the word templates of the key words in the text. Given that texts are rarely about one thing, we can tease out the topics in a text by listing the key words, as in the following samples of varied text types (see Appendix 1 for references):

- A Guardian review of Michael Haneke's film *Amour* (Bradshaw 2012) has such topics as classical music, family relationships, ageing (health and mortality) and cinema.
- In just one paragraph in Michael Cunningham's novel, *A Home at the End of the World* (p. 87), three topics emerge: (1) feelings about oneself – a negative past contrasting with a positive future; (2) secrets; and (3) shopping.
- In the introduction to a chapter entitled *The Meaning of Things in Time and Space* (Kral 2012: 209), there are words that represent the topics of meaning, time, space, attitude and people.
- In a science magazine article that paints a worst-case BSE scenario (Mackenzie 2002), there are words about animals, food, diseases and research.

Like any good, rich forest, there is more than one trail. The words of each topic form trails interweaving through the text. Students highlight each set in different colours and observe the weave. In addition to observing this feature of discourse, they arrive at rational lists of text-derived, topic-based vocabulary whose semi-preconstructed phrases can now be studied. It comes as no surprise that the key words are those around which Message chunks revolve. Nouns predominate: they manifest the main and related topics.

To put this into practice, a class was divided into four groups, each taking one of the topic trails in the above-mentioned BSE article. They started by locating the words and phrases that manifest their topic trail, writing them down the middle of a piece of chart paper; Figure 7 depicts how one group represented the RESEARCH topic trail. They then wrote the collocates beside them, which were later expanded into the word templates underpinning their usage in the text, following the earlier *scholarship* example.



RESEARCH

14/198
authorities establish(v) whether
0/1
epidemiologists(n) warn
10
human equivalent(n) of
0/48
its 12 symptoms(n) look like
81/0
effort to detect(v) BSE find
14/0
cattle tissue(n) prevent
0
brains being tested 127 precaution
a team 1 modelled (v)
0
researcher(n) calculated
calculate(v)
52
degree (n) of risk
26
the team predicts(v) that 29
120
1,037

Figure 7. The RESEARCH topic trail in an article about BSE, as drafted by a group of students

It is important to recognise what is omitted from a word template: grammatical features are only included if they are salient, so for example there are no auxiliary verbs forming the continuous, perfect and passive, nor modal verbs, nor articles, nor determiners. Word templates distil all of this out.

The important issue for learners now concerns the representativeness of these structures. At this stage of their observation, they have anecdotal evidence of the use of these words, which is no guarantee of their typical or canonical usage – the use of *award* vs. *win a scholarship* exemplified this. Gries's (2008: 425) comment that "authenticity does not automatically entail typicality" triggers the obligation to determine whether the extended collocations hitherto observed can be found in sufficient numbers to grant them the status of patterns of normal usage. After all, one swallow does not a summer make. Many a language learner will share Hoey's quandary (2000: 233):

> I was never sure whether the context was natural or typical. Unless one knows that the collocation one is learning is absolutely characteristic of the way the word is used, more than half the value one gets from learning the word in its context disappears.

In order to be certain that any structure is typical and therefore worth adding to our learning dossiers, we need to consult a corpus. With permission from Michael Hoey (personal correspondence), I refer to the procedure of checking the frequency of chunks in text against a corpus as the Hoey Procedure. He demonstrated this on his widely quoted Hammerfest sentence in *Lexical Priming* (2005: 5–7): rewriting a naturally-occurring sentence by using close synonyms can convey exactly the same propositional meaning and be grammatically accurate, but the result is "clumsy" at best – precisely because it avoids habitual collocations.

Once students assemble the topic trails into topic-based sets of key nouns and note their verb collocates – templates in the making – it is time to invoke the Hoey Procedure and check their frequency in a corpus. If found to be frequent, other collocates are also noted. We start by assembling collocations that appear in context, then check their frequency in the BNC. Figure 8 shows the frequencies that the students found by looking for the lemmas of both words within a span of 5 to the left and right. In most cases, they are significant enough for learners to consider them collocations worth adding to their English repertoire.

Extended collocations	**Collocations**	
sheep [animal] carry BSE [disease]	CARRY / DISEASE:	53
disease affects humans	DISEASE / AFFECTS:	91
infection passes to people	INFECTION / PASS: INFECTION / PASS TO:	20 10
efforts to detect BSE [disease] were abandoned	DETECT / DISEASE: ABANDON / EFFORT:	20 20
sheep [animal] show symptoms of BSE [disease]	SHOW / SYMPTOM:	94

Figure 8. Hoey Procedure applied to some extended collocations in the article on BSE

Being a general corpus, the BNC does not contain 100 million tokens of medical language, let alone the specifics of one topic trail in one article. As Hanks (2010: 1300) reminds us, "terminology in its purest form is rare in general language and typically found only in highly specialized texts." This renders the Hoey Procedure impractical when using a general corpus to deal with specialized topics. There are several alternatives. First, the clustering tool in the word sketch function often provides enough data to observe a pattern of normal usage. Second, students can use hypernyms and semantic sets in the slots. Those working at this level have enough knowledge of the world and of English to look through concordances and observe what sorts of things fill the slots, although rare words and cultural references can obfuscate the process. Finally, they can create a specialised corpus using one of Sketch Engine's tools. Space does not allow any elaboration here, but the process is described in Thomas (2015).

Over time, as students observe words in their extended collocations, the human tendency to categorise kicks in (Hanks 2012: 58; see also Tomasello 2005: 3–4). This can be given a nudge by asking them to store their vocabulary in structured categories. The top-level category is the part of speech of the target word. If the collocation AUTHORITIES/ESTABLISH is demonstrated to be representative usage, we can take this further: observing further patterns of normal usage is also of value. For example, *the* AUTHORITIES/ESTABLISH/WHETHER. The students performing the Hoey Procedure find that ESTABLISH/WHETHER is a frequent colligation and is a subset of ESTABLISH/WH-. Of the 186 instances of ESTABLISH/WHETHER, 130 are preceded by the to-infinitive marker. Of the 18 instances of AUTHORITY/ESTABLISH, half are passive. Different meanings of *establish* emerge from their extended collocations. Those who *establish* include *government*, *Act*, *company*, *case*, *Council*, *authority*, *law*, *time*, *evidence*, *agreement*, *study* and *treaty* (in order of frequency) which may be categorised as such semantic types as [Body] [Act].

Brown (2007: 258) claims that every set of complex skills is acquired through observing, focusing, practising, monitoring, correcting and redirecting. While the stages in the C+ procedure involve all of these processes, they vary in their cognitive demands. Many learners require only a little training in recognising instances of 'O' language at the beginning of sentences, since they are often separated from 'M' language by commas, and they come to recognise many of these discourse markers as formulaic. But it is their function in the text that holds the greatest interest. Neither is it particularly challenging to identify topic trails, but it is a good exercise in observing and focusing. In addition to locating exemplars of these phenomena, both of these activities involve observing aspects of how full text works. Observing and recording the use of key nouns with their collocates is a little more demanding, and noting or accounting for their syntactic relationships certainly requires higher-order thinking skills. Obtaining data from corpora to check the extent to which the collocations and extended collocations are canonical not only requires corpus training, but assumes the requisite metalanguage. This in itself is valuable language training and given the amount of metalanguage that non-linguist language students acquire concerning verb forms, aspects of nouns, clauses etc., it is not beyond them. Interpreting the data and extending the collocates to semantic types to create meaningful and viable word templates involves deciding which collocates in a word sketch are relevant. While clustering helps, this is often a cognitively demanding task.

The guided discovery learning advocated in C+ differs from that which is discussed in the literature, usually comparing deductive and inductive approaches (see Flowerdew, this volume). Discovery learning, according to Richards and Schmidt (2002: 162), is where "learners develop processes associated with discovery and inquiry by observing, inferring, formulating hypotheses, predicting and communicating". As Thornbury (no date) writes under the heading of Guided Discovery:

> Guidance is typically mediated by questions, each question challenging learners to advance their understanding one further step. Clearly, the notion of asking questions as a means of co-constructing learning maps neatly onto a sociocultural model of learning, where the teacher is working within the learners' *zone of proximal development* in order to scaffold their emergent learning.

The target in these situations is usually discrete grammar points, e.g. comparing *must* and *have to*, *will* and *going to*, present perfect and simple past. These are simply 'display questions' as the teacher has the 'right answer' at hand – there is no genuine discovery let alone room for exploration, interpretation, or classroom debate. There is no 'fuzzy'. Furthermore, in C+ the students are discovering features of a particular text and features of particular words. They are required to apply what they already know about English and dig deeper to discover for themselves facts about word grammar, the kernel of clauses.

In terms of developing a specific artefact, students build glossaries, undertaken as a task-based activity for groups of students to publish on their website, for example. An aspect of corpus work not yet touched on is the selection of illustrative sentences, without which no glossary would be complete. The GDEX algorithm described above is of considerable value here.

This paper has been noisily trumpeting word templates as a valuable procedure and resource for learners, but concedes that deriving word templates is not for everyone. However, even without excavating them themselves, pre-processed word templates can be used by students in productive activities. For example, in a task where students write an email to a radio station in response to a news story, they can use the word templates of the key words in the story for their 'M' language, and embed them in 'O' language chunks revolving around *disagree*, *insult my intelligence*, *demand an explanation*, *believe my ears*, for example. Students still have to grammatize the word templates and integrate them into their narrative.

6. Conclusion

The overall aim of this chapter has been to demonstrate some practical teaching applications of some findings from corpus linguistics. Searching corpora to confirm the language facts provided in grammar and course books focuses on the lower rungs of the hierarchy of language where the answers, already known to the teacher, are mostly either right or wrong. Such activities are also on the lower rungs of Bloom's taxonomy. But given that current corpus studies are revealing the highly patterned nature of language, the corpus tasks students undertake can lead them to such findings themselves. Not only do they thereby acquire holistic units of language, but the process furnishes them with a new framework for understanding language. For one thing, fuzzy is welcome.

Collocation Plus aims to inculcate a sense for the 'grammar of vocabulary' which is essential for students to turn receptive vocabulary into productive. Restricting itself to the two-lexeme definition of collocation, C+ also fosters sensitivity to syntax. This extends to full word templates, which being the skeletons of clauses, learners uncover by chipping away such grammatical elements as tense, aspect, articles, aspects of modality, that normally situate it in the real world. Learners observe the semantic sets that occupy the paradigmatic choices available, which often provide them with opportunities to recycle and extend their knowledge of vocabulary. In the nomenclature of Bloom's taxonomy, these activities can be seen as *analysis*. Constructing sentences from word templates, then fleshing out the skeleton, involves grammatizing them, which is a highly context-sensitive process and is an act of *synthesis*.

Focusing on prepositions within the bound and free framework leads learners to observe that the relatively small number of highly frequent prepositions on the right of a word are in some way bound to it, while the relatively large number of less frequent prepositions launch prepositional phrases expressing the circumstances of the clause. Graphic frequency lists depict this Zipfian tendency in a high proportion of cases. Students are often grateful to have a framework within which they can address the issue of prepositions.

This leads to another particularly important aim of the procedures advocated in this paper, namely, to explore new types of activities in which students operate at the discourse level, as instantiated by topic trails. While it may appear to students that this is simply a way of selecting vocabulary to study, it also depicts the interweaving of topics through text. This work awaits further study. Yet another aspect to C+ and word templates, though not within the scope of this chapter and to be pursued elsewhere, concerns suprasegmental phonology (cf. Aston, this volume), in particular tonic stress and vowel reduction. The vast majority of scientists that I have worked with over the last 15 years have never had any systematic training in this aspect of pronunciation.

Such a multi-tiered, multi-step teaching procedure requires a considerable amount of valuable class time which teachers need to be able to justify, to themselves first. For those who argue that it is overly time consuming, one word offers a robust retort: affordances. Students learn many things at the same time in investigating language in these ways. Through discovering the specific linguistic information they find in texts that revolves around word usage, they are initiated into the linguistic thinking that has been evolving in the last thirty years. To my mind the failure to acknowledge this, let alone inculcate it, is the biggest omission in the literature concerning the use of corpora in language teaching. Equipped with a view of language that revolves around patterns of normal usage, and some procedures for observing them in the texts they read, learners are well on the way to learner autonomy, which has been confirmed to me by many ex-students over the years. The following attested statements indicate how deeply they grasp and appreciate the value of such work:

– Why didn't anyone ever tell us this before?
– This really is how language works, isn' it? I had no idea!
– I haven't written an article without consulting corpora for years now.
– No dictionary could ever tell me that.
– Thank you for making yourself redundant!

References

Anderson, L.W. & Krathwohl, D.R. (eds). 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. London: Longman.

Bartram, M. & Walton, R. 1991. *Correction: Mistake Management – A Positive Approach to Language Mistakes*. Hove: Language Teaching Publications.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Benson M., Benson, E. & Ilson, R. 1986. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. Amsterdam: John Benjamins. DOI: 10.1075/sl.11.2.20boo

Boulton, A. 2010. Learning outcomes from corpus consultation. In *Exploring New Paths in Language Pedagogy: Lexis and Corpus-Based Language Teaching*, M. Moreno Jaén, F. Serrano Valverde & M. Calzada Pérez (eds), 129–144. London: Equinox.

Breyer, Y. 2009. Learning and teaching with corpora: Reflections by student teachers. *Computer Assisted Language Learning* 22(2): 153–172. DOI: 10.1080/09588220902778328

Brown, H.D. 2007. *Teaching by Principles: An Interactive Approach to Language Pedagogy.* Harlow: Pearson Education.

Cosme, C. & Gilquin, G. 2008. Free and bound prepositions in a contrastive perspective: The case of 'with' and 'avec'. In *Phraseology: An Interdisciplinary Perspective*, S. Granger & F. Meunier (eds), 259–274. Amsterdam: John Benjamins. DOI: 10.1075/z.139.23cos

Crystal, D. 1995. *The Cambridge Encyclopedia of the English Language*. Cambridge: CUP.

Ellis, N.C. 2008. Usage-based and form-focused language acquisition: The associative learning of constructions, learned attention, and the limited L2 endstate. In *Handbook of Cognitive Linguistics and Second Language Acquisition*, P. Robinson & N. Ellis (eds), 372–406. London: Routledge.

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

Francis, G., Hunston S. & Manning E. 1998. *Collins Cobuild Grammar Patterns,* 2: *Nouns and Adjectives*. London: Collins.

Frankenberg-Garcia, A. 2014. The use of corpus examples for language comprehension and production. *ReCALL*, 26(2): 128–146. DOI: 10.1017/S0958344014000093

Grice, P. 1975. Logic and conversation. In *Syntax and Semantics, 3: Speech Acts*, P. Cole & J. Morgan (eds), 41–58. New York NY: Academic Press.

Gries, S.T. 2008. Corpus-based methods in analysis of second language acquisition data. In *Handbook of Cognitive Linguistics and Second Language Acquisition*, P. Robinson & N. Ellis (eds), 406–431. London: Routledge.

Halliday, M.A.K. & Hasan R. 1976. *Cohesion in English*. Longman: London.

Hanks, P. 2010. Lexicography, terminology and phraseology. In *Proceedings of the XIV Euralex International Congress*, A. Dykstra & T. Schoonheim (eds), 1299–1306. Afûk, Ljouwert: Fryske Akademy. ⟨http://www.euralex.org/elx_proceedings/Euralex2010/122_Euralex_2010_9_HANKS_Terminology,%20Phraseology,%20and%20Lexicography.pdf⟩ (5 July 2014).

Hanks, P. 2012. How people use words to make meanings: Semantic types meet valencies. In *Input, Process and Product: Developments in Teaching and Language Corpora*, J. Thomas & A. Boulton (eds), 54–69. Brno: Masaryk University Press.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge MA: The MIT Press. DOI: 10.7551/mitpress/9780262018579.001.0001

Hanks, P. Ongoing. *Pattern Dictionary of English Verbs*. ⟨http://deb.fi.muni.cz/cpa⟩ (3 May, 2014).

Hoey, M. 2000. A world beyond collocation: New perspectives on vocabulary teaching. In *Teaching Collocation: Further Developments in the Lexical Approach*, M. Lewis (ed.), 224–243. Hove: Language Teaching Publications.

Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Jackson, H. 1988. *Words and their Meanings.* Harlow: Longman.

Jakubíček, M., Kilgarriff, A., McCarthy, D. & Rychlý. P. 2010. Fast syntactic searching in very large corpora for many languages. In *PACLIC*, 741–747. Waseda University: Institute for Digital Enhancement of Cognitive Development.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the 13th EURALEX International Congress*, E. Bernal & J. DeCesaris (eds), 425–432. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. ⟨http://www.euralex.org/elx_proceedings/Euralex2008/026_Euralex_2008_Adam%20Kilgarriff_Milos%20Husak_Katy%20McAdam_Michael%20Rundell_Pavel%20Rychly_GDEX_Automatically%20Finding%20Good%20Di.pdf⟩ (5 July 2014).

Kilgarriff, A., Kovář, V., Krek, S., Srdanović, I. & Tiberius, C. 2010. A quantitative evaluation of word sketches. In *Proceedings of the 14th EURALEX International Congress*, A. Dykstra & T. Schoonheim (eds), 372–379. Leeuwarden: Fryske Academy.

Lee, D. 2001. Genres, registers, text types and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* 5(3): 37–72.

Lewis, M. 1993. *The Lexical Approach.* Hove: Language Teaching Publications.

Lewis, M. (ed.). 2000. *Teaching Collocation: Further Developments in the Lexical Approach*. Hove: Language Teaching Publications.

McEnery, T. & Hardie, A. 2011. *Corpus Linguistics: Method, Theory and Practice.* Cambridge: CUP. DOI: 10.1017/CBO9780511981395

Richards, J. & Schmidt, R. 2002. *Longman Dictionary of Language Teaching and Applied Linguistics*, 3rd edn. Harlow: Longman.

Rychlý, P. 2008. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing* (*RASLAN*), P. Sojka & A. Horák (eds), 6–9. Brno: Masaryk University Press.

Schmitt, N. 2010. Key issues in teaching and learning vocabulary. In *Insights into Non-Native Vocabulary Teaching and Learning*, R. Chacón-Beltrán, C. Abello-Contesse & M. M. Torreblanca-López (eds), 28–40. Bristol: Multilingual Matters.

Scott, M. & Tribble, C. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins. DOI: 10.1075/scl.22

Simpson-Vlach, R. & Ellis, N. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31(4): 487–512. DOI: 10.1093/applin/amp058

Sinclair, J.M. 2004. New evidence, new priorities, new attitudes. In *How to Use Corpora in Language Teaching* [Studies in Corpus Linguistics 21], J.M. Sinclair (ed.), 271–299. Amsterdam: John Benjamins. DOI: 10.1075/scl.12.20sin

Sinclair, J.M. & Mauranen, A. 2006. *Linear Unit Grammar: Integrating Speech and Writing* [Studies in Corpus Linguistics 25]. Amsterdam: John Benjamins. DOI: 10.1075/scl.25

Stefanowitsch, A. & Gries, S.T. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–243. DOI: 10.1075/ijcl.8.2.03ste

Stubbs, M. 2001. *Words and Phrases.* Oxford: OUP.

Thomas, J. 2008. Impatience is a virtue: Students and teachers interact with corpus data – now. In *Proceedings of the 8th Teaching and Language Corpora Conference*, A. Frankenberg-Garcia (ed.), 463–469. Lisbon: ISLA-Lisboa.

Thomas, J. 2015. *Discovering English with Sketch Engine*. Brno: Versatile.

Thornbury, S. No date. *An A–Z of ELT* [Blog]. 〈http://scottthornbury.wordpress.com〉(10 March, 2012).

Timmis, I. 2008. The lexical approach is dead: Long live the lexical dimension. *Modern English Teacher* 17(3): 5–10.

Tomasello, M. 2005. *Constructing a Language: A Usage-based Theory of Language Acquisition*. Harvard: Harvard University Press.

Van Lier, L. 2000. From input to affordance: Social-interactive learning from an ecological perspective. In *Sociocultural Theory and Second Language Learning*, J. Lantolf (ed.), 245–259. Oxford: OUP.

Appendix 1: Text examples cited

Bradshaw, P. 2012. Cannes 2012 Amour – review. *The Guardian*, 20 May. ⟨http://www.theguardian.com/film/2012/may/20/amour-haneke-film-review⟩

Cunningham, M. 2004. *A Home at the End of the World*. New York NY: Picador.

Kral, I. 2012. *Talk, Text and Technology: Literacy and Social Practice in a Remote Indigenous Community*. Bristol: Multilingual Matters.

MacKenzie, D. 2002. vCJD deaths will rise if UK sheep have BSE. *New Scientist*, 9 January. ⟨http://www.newscientist.com/article/dn1772-vcjd-deaths-will-rise-if-uk-sheep-have-bse.html#.U1y58K2SyKw⟩

Staffordshire University website. No date. National scholarship awarded to Staffordshire University lecturer. ⟨https://www.staffs.ac.uk/news/national-scholarship-awarded-to-staffordshire-university-lecturer-tcm4242978.jsp⟩

Appendix 2: Corpora cited

All corpora are tagged with the TreeTagger (⟨https://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html⟩) for English. The last column indicates the number of tokens. All the corpora are accessible via Sketch Engine.

British National Corpus (BNC)	Retagged with the TreeTagger for English	112,985,133
New Model Corpus (NMC) super sensed	Corpus of texts created by web crawling; in addition to the Treetagger, it also has semantic tagging and named entity labels	115,074,168
Informatics Reading Corpus (IRC)	Corpus of academic articles that doctoral students of informatics upload	6,690,531
CorpusCorpus (CC)	Corpus of empirical research articles on using corpora in language teaching	1,119,024