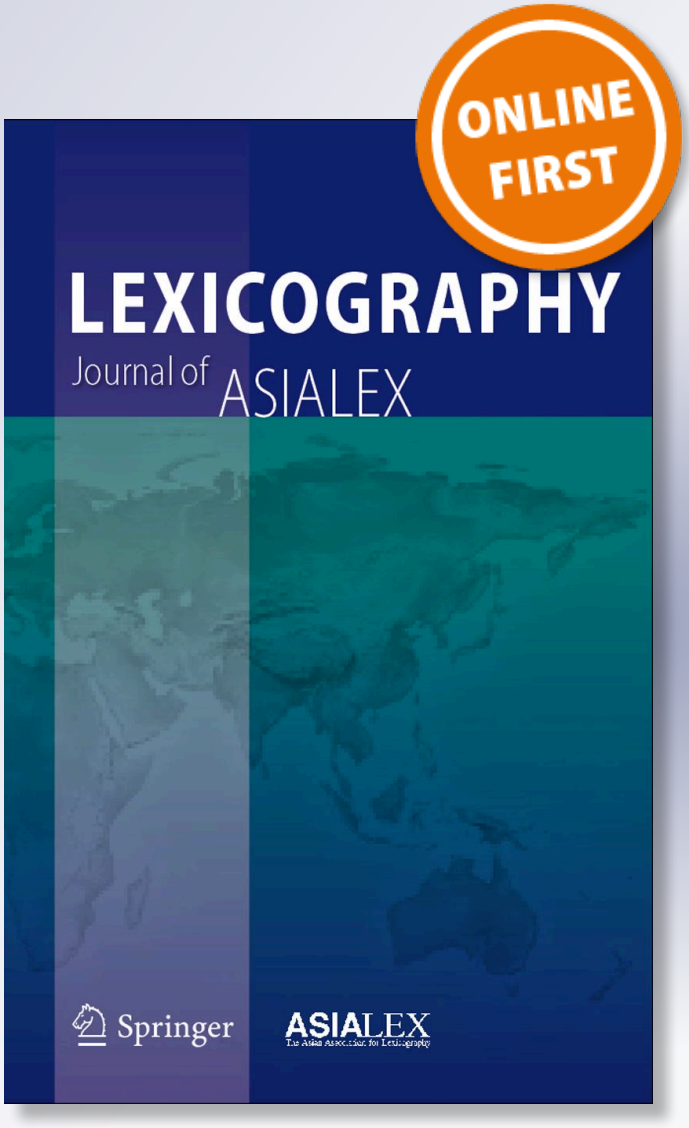
Searching for extended units of meaning—and what to do when you find them

Michael Rundell

Lexicography
Journal of ASIALEX

ISSN 2197-4292

Lexicography ASIALEX
DOI 10.1007/s40607-018-0042-1



Springer

Lexicography ASIALEX
https://doi.org/10.1007/s40607-018-0042-1

ORIGINAL PAPER

Searching for extended units of meaning—and what to do when you find them

Michael Rundell[1]

Received: 24 October 2017 / Accepted: 5 February 2018

Abstract Two of the key outcomes of corpus-linguistic research over the past 30 years have been the development of the idea that meanings are mostly constructed through context (undermining traditional notions of the individual word as an autonomous bearer of meaning); and the discovery that recurrence and regularity—our tendency to employ a limited number of conventionalized ways of expressing ideas—are essential features of the language system. Both findings have had a major impact on our understanding of how language works, and both have influenced the content of dictionary entries—contributing, for example, to improved word sense disambiguation, and to a greater emphasis on phraseology and collocation. However, there is still much to do. Ever-larger corpora and more powerful corpus-query tools reveal areas where we can further improve our description of languages, and thus provide better resources for users. In addition, the migration of dictionaries to digital media (removing space constraints) opens up new opportunities for doing this. In a characteristically far-sighted paper (Sinclair, Textus 9(1): 75–106, 1996), John Sinclair broadened the search for what he called "units of meaning" by investigating longer strings of words and identifying recurrent, and often quite extended, patterns of usage. Using this as a starting point, I will look at other examples in corpus data of the kinds of patterning Sinclair discussed, and we will see how current corpus-querying systems can help us identify these extended units of meaning. Finally, I will speculate about whether dictionaries should aim to describe these longer units, and if so, how this might work in practice.

Keywords Extended units of meaning · Collocation · Colligation · Semantic prosody · Multi-word sketch · Longest commonest match

✉ Michael Rundell
michael.rundell@lexmasterclass.com

[1] Lexical Computing Ltd, Brighton, UK

Published online: 20 March 2018

1 Theoretical background

1.1 Introduction

Corpus-linguistic research, applied to ever-growing volumes of language data, has undermined the notion of the individual word as the primary unit of meaning. Meanings are mostly constructed through context, not by slotting vocabulary items into spaces created by a formal syntax. In addition, many of the patterns which we use to encode meanings occur with remarkable frequency, as speakers draw upon a fairly limited repertoire of conventionalized "ways of saying". All of which suggests that recurrence is an essential feature of the language system. As Sinclair puts it, "By far the majority of text is made of the occurrence of common words in common patterns" (Sinclair 1991:108). It follows that a high percentage of what people say or write is predictable, because, as Hanks observes, "Although the number of possible combinations may in principle be limitless…the number of probable combinations…is rather limited" (Hanks 2013: 399).

However, although corpus analysis enables us to observe the inbuilt predictability of most language output, much of this is far from predictable to a learner or non-fluent user of a language. Even where a given word combination is semantically transparent, its status as a recurrent string, as a norm worth learning, is not necessarily obvious. This raises the question of what dictionaries can or should do to identify and describe recurrent patterns of usage, and this is one of the themes of this paper. I will start by looking at a paper written over 20 years ago by John Sinclair (Sinclair 1996), in which he explores the idea of "extended units of meaning". In many earlier works, Sinclair had already developed (and demonstrated the workings of) what he called "the idiom principle"—the idea that language users regularly resort to an inventory of "semi-preconstructed phrases that constitute single choices" (Sinclair 1991: 110). By the time, Sinclair was writing this, relatively small units of meaning, such as two-word collocations, and other so-called "lexical bundles" (Biber et al. 1999: 990ff), had already been extensively discussed, and were beginning to be accounted for in dictionaries. In his 1996 paper, however, Sinclair was interested in longer patterns, sometimes of considerable complexity, which the data show to be remarkably frequent.

From a lexicographic point of view, this interest in multi-word units of meaning is relatively recent. Traditional lexicographic practice has rested on the assumption that individual words are autonomous bearers of meaning—a view reflected in the names we give to dictionaries (which are called "word books" in many Germanic languages) and even in definitions of the word dictionary itself. For example:

> a book that gives a list of words in alphabetical order and explains what they mean (*Macmillan English Dictionary*, first edition, 2002)

This Macmillan definition (now superseded) exactly describes Cawdrey's *Table Alphabeticall* of 1604 (generally thought of as the first monolingual English dictionary), where the "definitions" typically consist of one or two

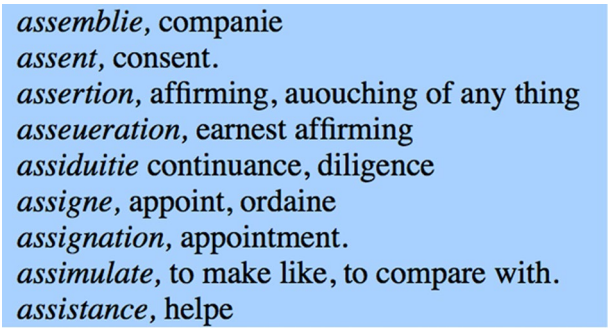


Fig. 1 Extract from Cawdrey's Table Alphabeticall, 1604

almost-synonymous words (see Fig. 1), while contextual information (showing the conditions in which word X is equivalent to word Y) is entirely absent.

This long-standing focus on the word—in linguistics generally and dictionaries in particular—is not so surprising, given that "in the majority of writing and printing conventions, words are separated by spaces…A text is therefore seen as a succession of discrete items, those items being words" (Sinclair 1996: 75). In this model, the meaning of an utterance is a concatenation of the meanings of the individual words which it comprises. In addition, when dictionaries have to deal with longer units of meaning such as idioms and phrasal verbs, these items have traditionally been relegated to the bottom of a main dictionary entry, so that *break out* and *break the bank* are "nested" at the end of the entry for *break*. Longer units such as these, as Sinclair notes, "are considered as marginal phenomena, almost aberrations" (Sinclair 1996: 76).

1.2 Understanding how meanings are created

All of this changed under the impact of corpus study, which provided the empirical basis for a radically different understanding of how meanings are created—a model of meaning summed up in Sinclair's well-known observation that "Many if not most meanings depend for their normal realization on the presence of more than one word" (Sinclair 1998). A number of pre-corpus scholars had already begun to develop the idea that meaning is at least partly dependent on context and co-text, rather than being an inherent property of individual words. As far back as 1755, Samuel Johnson had recognised that "It is not sufficient that a word is found, unless it be so combined as that its meaning is apparently determined by the tract and tenour of the sentence" (Johnson 1755). Much more recently, J. R. Firth demonstrated that features such as collocation, colligation, and phraseology had a central (rather than marginal) function in the language system and that the meaning of a word could not be fully understood without knowing "the company it keeps".

Similar ideas were developing among scholars involved in the teaching of English as a second language. Several years of experience as a language teacher brought Harold Palmer—working in Japan during the 1920s and 1930s—to the realization

that "it is not so much the words of English nor the grammar of English that makes English difficult … The vague and undefined obstacle to progress … consists for the most part in the existence of so many odd comings-together-of-words." (Palmer 1933, quoted in Cowie 1999: 52–53). The first major learner's dictionary of English, compiled in Japan by Palmer's protégé A.S. Hornby, adapted the standard lexicographic model by including information about syntactic behaviour, phraseology, and (to a lesser extent) collocation. Since Hornby, pedagogical dictionaries have paid increasing attention to longer units of meaning. Salient two-word collocations are extensively covered in today's dictionaries, while in the *Longman Language Activator* (Summers 1993), the lists of near-synonyms which lexicalize a given concept make no distinction between single words and multi-word expressions. Thus, the concept "Usually" is instantiated not only by words like *generally* and *routinely*, but also by longer units such as *nine times out of ten* and *as a rule*. From the speaker's point of view, these are all equally valid choices, whose selection depends on the meanings they convey, not on their status as "words" or "phrases". In a significant recent development (facilitated by the migration of dictionary text from print to digital platforms), many pedagogical dictionaries now treat phrasal verbs and idiomatic phrases as headwords in their own right, only loosely connected to the entries under which they were formerly "nested". Thus, gradually, dictionaries' exclusive focus on single words has given way to a more mixed picture, where the role of longer units of meaning is recognised at both microstructural and macrostructural levels.

These developments in lexicographic practice reflect theoretical insights, gained through the study of corpus data, into how meanings are created and understood. Hanks proposes (e.g., Hanks 2013: 73f.) that words on their own do not have meanings; rather, they have "meaning potentials". These meaning potentials are activated by specific contextual features, and many of the resulting patterns (of word plus syntactic and/or lexical context) recur frequently enough in corpus data to be regarded as normal units of meaning. In addition, since the primary job of a dictionary is to account for linguistic norms, Sinclair's interest in longer units of meaning (beyond those already described in dictionaries) is a logical next development.

1.3 Sinclair's 1996 paper: The search for units of meaning

In this paper, Sinclair looks in detail at corpus data for a number of words and expressions, including the verb *brook* and the multi-words *true feelings* and *naked eye*. In the case of the latter, a frequent though more or less opaque combination, he discovers a complex network of recurrent patterns. To summarize the main features of this patterning: the two positions to the left of the node, which we will refer to as N-1 and N-2, are typically filled by *to the* [*naked eye*] or *with the* [*naked eye*]. As we move further to the left, things become more interesting. At position N-3, the language data reveal a strong preference for words relating to visibility: this slot tends to be filled by verbs like *see*, *spot*, and *perceive*, or by adjectives such as *visible*, *evident*, and *detectable*. For the verbs in N-3, there is a colligational preference for use with a modal at N-4 (especially *can* or *could*): *…these could be seen with the naked eye from a helicopter*. Finally, Sinclair finds that a semantic prosody of "difficulty"

is evident in 85% of the instances in his sample: we see this in expressions like *too faint to be seen by the naked eye* or *barely visible to the naked eye*.

What Sinclair presents here is a model of a lexical item consisting of several words which—unlike items which we categorize as idioms or fixed phrases—can tolerate "a great deal of internal variation". Distinct linguistic features such as collocation, semantic preference, colligation and semantic prosody all combine to create one of the "semi-preconstructed" phrases which Sinclair referred to in his earlier work, and which he sees as effectively a single lexical choice. In addition, despite all the internal variation found in his set of naked eye phrases, "there is always a clearly preferred selection right down to the actual words". This calls to mind Halliday's hypothesis—made long before corpus data were sufficiently abundant to confirm it—about "the ability of a lexical item to 'predict' its own environment" (Halliday 1966:160).

There are of course many exceptions to the patterns which Sinclair focuses on. In position N-2, for example, prepositions other than *to* and *with* can sometimes occur, and there are some sentences which do not exhibit any of the semantic preferences Sinclair identifies as being typical, such as:

To the naked eye, he is easily one of their the fittest…

As anyone who has spent much time looking at corpus data knows, it is not difficult to find exceptions to whatever generalizations emerge from one's analysis. However, for dictionary makers, such exceptions are of far less interest than the norms.

From the point of view of practical lexicography, two key messages emerge from Sinclair's investigations. First, the need to broaden our notions of what constitutes a lexical unit to be accounted for in a dictionary: "So strong are the co-occurrence tendencies of words, word classes, meanings and attitudes that we must widen our horizons and expect the units of meaning to be much more extensive and varied than is seen in a single word." The second and related point is that, in the theory of meaning Sinclair proposes in this paper, "the idea of a word carrying meaning on its own would be relegated to the margins of linguistic interest, in the enumeration of flora and fauna, for example". From a lexicographic point of view, this is a significant reversal of traditional practice, where the word is central and longer units are seen as anomalies.

1.4 Some new examples

With regard to the specific items he investigates, Sinclair's analysis is compelling. However, before we consider the implications for practical lexicography, it seems advisable to get a clearer idea of just how pervasive this sort of patterning is. Therefore, in the section which follows, we will attempt a similar analysis on several other items. The corpus used here is the LexMCI corpus, which was the main evidence base for the creation of the DANTE lexical database (Convery et al. 2010). LexMCI

is a collection of about 1.7 billion words of contemporary English, and is described in more detail on the DANTE website: http://www.webdante.com/the_corpus.html.

1.4.1 wreak

Number of instances in the corpus: 2334

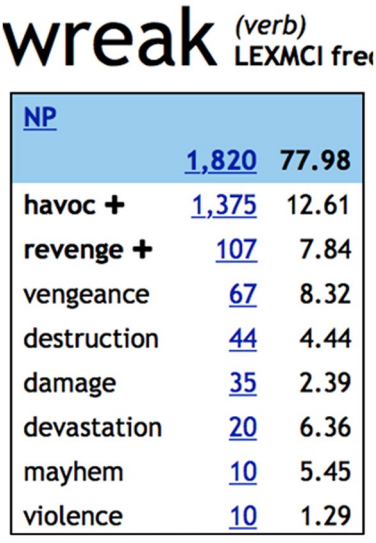
A word sketch for *wreak* (Fig. 2 shows an extract) vividly illustrates the stand-out feature of this verb.

This verb's typical objects fall into just two semantic classes: revenge or some form of chaos and destruction. There is some variation in the choice of collocates, but *havoc* is so dominant that the collocation *wreak havoc* is almost a fixed phrase. Regardless of the object type, a prepositional phrase follows the noun object in almost 60% of cases, with *on* being by far the most frequent preposition. The usual pattern is *wreak havoc/revenge on*, but in about 14% of cases the object is modified by an adjective such as *untold*, *terrible*, or *enormous*.

1.4.2 untoward

Number of instances in the corpus: 992

This is an unusual adjective. A noun complement follows in about 45% of cases, and two semantic types dominate: words meaning (roughly) "event" or "consequence". The most frequent of these is *incident*: the collocation *untoward incident* makes up over 13% of all instances of *untoward*. Typically occurring in the context of discussions about health and safety, it could be seen almost as a technical term in its own right:


wreak (verb) LEXMCI fre

NP		
	1,820	77.98
havoc +	1,375	12.61
revenge +	107	7.84
vengeance	67	8.32
destruction	44	4.44
damage	35	2.39
devastation	20	6.36
mayhem	10	5.45
violence	10	1.29

Fig. 2 Extract from a Word Sketch for *wreak*

Staff also need some avenue for whistleblowing, to voice concerns about untoward incidents or bad practice.

The striking colligational feature of *untoward* is that it functions, in a majority of cases, as a postpositive adjective, usually preceded by the words *nothing* (201 instances), *anything* (172), or *something* (50):

He was keen to reassure parents that nothing untoward was going on.

After initially denying that anything untoward had happened… he later confessed to raping the boy.

In these and many similar cases, there is a broadly negative semantic prosody. In addition, when that does not apply, a conditional appears in around 10% of the postpositive examples:

If you spot anything untoward, let me know.

Wreak and *untoward* are relatively rare words in English, so it would not be surprising if the range of patterns in which they appear is limited. However, even here, there is clear evidence of recurrent "units of meaning", some of which are quite extended.

1.4.3 remiss

Number of instances in the corpus: 502

A cursory glance at a concordance of *remiss* immediately reveals a limited set of overlapping patterns, which recur in various permutations. The main ones are:

- *remiss*+*of*+[personal pronoun or noun]: 151 instances

  It would be remiss of me to not thank the rest of the staff.

- *remiss*+[to-infinitive (within 5 words)]: 195 instances

  How remiss of you to forget to pay the monthly storage fee.

- *It* [*is/seems* etc.] *remiss*: 197 instances

  It seemed a little remiss not to book anyone truly spectacular to launch the new slot.

- *would be remiss*: 224 instances

It would be very remiss; however, if I did not thank the Baths Superintendent and his wife and staff.

- *remiss*+*not* (within 5 words): 177 instances

I do feel that we have been somewhat remiss as a council by not welcoming what has been done by Mr Banks and Pentland Ferries.

- *remiss+in*: 97 instances

We are remiss in not noticing this link was missing before now.

The number of corpus instances, where none of these patterns is present is extremely small—probably no more than 2%. Not only that, the data show that some combinations recur with extraordinary regularity, notably this one, which makes up almost 30% of all cases:

It would be [*very/extremely* etc.] *remiss* [*of*] [*me/her* etc.] [*not*] *to* ….

1.4.4 sink in

Our focus here is the phrasal verb.[1] A simple search for the lemma *sink*-verb followed immediately by *in* (*sink in* is a non-separable phrasal verb) generates a concordance of 3819 lines. However, well over half of these are instances of the verb followed by a prepositional phrase, and therefore not relevant for the present purposes. For example:

Their fishing boat sank in the Bristol Channel between Penarth Pier and Cardiff Bay.

He was so sunk in his despair, he scarce observed the change.

To filter out the noise, a search string was created using CQL (Corpus-Query Language), a syntax used in Sketch Engine for specifying complex searches. The CQL query[2] reduced the output from 3819 lines to 1592, and a random sample of 600 was taken from this "candidate" set. Finally, manual methods reduced the 600 candidates to 470 bona fide instances of the phrasal verb *sink in*, and this dataset forms the basis for what follows.

One of the most striking features of this verb's behaviour is its colligations. Colligation refers to a word's observable preference for occurring in—or for avoiding—a particular form, a particular position in the sentence, or a particular grammatical function (see Hoey 2005: 43ff. for a fuller description). Three facts stand out. First, the verb has a strong preference for the infinitive form, with 144 of 470 instances being infinitives. Second, in well over 70% of cases, *sink in* appears at—or very close to—the end of a sentence or clause. 265 instances

[1] In this case, it is not possible to give a reliable figure for the number of occurrences of our target word in the corpus, since (as the following paragraph shows) instances of the phrasal *sink in* are interspersed arbitrarily with cases where the verb *sink*, in its usual meanings, is followed by the preposition *in*.

[2] The CQL query used was: [lempos="(sink)-v"] [lemma="in"] [tag!="CD" & tag!="JJ" & tag!="N.*" & word!="the" & lemma!="a"]. This finds all cases of *sink*-verb (lemma) followed immediately by *in*, but it excludes cases where *in* is followed by a number (CD) (to eliminate cases like *sank in* 1815), an adjective (JJ), a noun (N.*) or the words *a* or *the*. It would no doubt be possible to find a more elegant solution, but this immediately removed well over 2000 non-relevant cases from the raw sample.

are immediately followed by punctuation; in a further 50 or so cases, we find the pattern *sink in*+adverb+punctuation (e.g., *it has not sunk in yet*.); and in 20 or so other cases, *sink in* is followed (without intervening punctuation) by a conjunction such as *and* or *but* introducing a new clause. In those cases, where *sink in* is not clause- or sentence-final, it is often followed by a *that*-clause (47 instances) or occasionally a *wh*-clause (6). And thirdly, in almost a quarter of cases *sink in* occurs in a broadly negative environment, such as:

It had not really sunk in until I spoke to mum.

I don't think the shock of it all has sunk in yet.

The realization that they are 'just like us' has yet to sink in.

Looking now at the characteristic co-text of *sink in*, there are three common types of subject (referring to what it is that "sinks in"): information (instantiated by words like *message*, *words*, and *news*), impact (*implications*, *scale*, *realisation*, *gravity*, *impact*, *extent*), and—most frequently—the pronoun *it*:

Read that again. Let it sink in.

I miss him already and it has not really sunk in.

"*It still hasn't sunk in*, " *says McGoldrick.*

A key fact about the meaning of *sink in* is reflected in another recurrent contextual feature: in a high proportion of cases, there is some indication that "sinking in"—the full absorption of new information—is a process, and it takes time. This feature is realised in a number of ways:

- co-occurring with *start* or *begin* (40 instances)

  Now that she has done that, the shock begins to sink in.

  It is just starting to sink in now, but when my name was announced, I was just dumbstruck.

- in the pattern *take*+time marker+*to sink in* (51 instances):

  …the reality took a little while to sink in

  This came as a complete shock to me and has taken a few days to sink in.

  …who knows how long it will take for all the implications to sink in?

  (As a variation on this, we also find patterns like *let/allow/give something (time) to sink in*.)

- co-occurring with adverbs such as *gradually*, *finally*, *slowly*, *eventually*, or in questions or negatives with *yet* (*has it sunk in yet*?)

- the use of the progressive form (typically with an adverb like *still* or *only just*):

The awful truth was slowly sinking in.

… the shock of what happened is still sinking in

Finally to the role (if any) of semantic prosody in the way *sink in* behaves. In his analysis of *naked eye*, Sinclair found a semantic prosody of "difficulty" in a high proportion of corpus instances. Prosodies like this differ from collocation or characteristic co-text in that they are not instantiated by specific lexical items. In the case of *naked eye*, the sense of something being difficult is generally present, but the ways in which this is conveyed lexically can be quite diverse. In the case of *sink in*, the picture is less clear. "Bad" situations outnumber "good" ones by about three to one, so we are more likely to encounter instances like this:

As the scale of the catastrophe sank in, he began to fear for his family…

…than like this:

As the initial euphoria sinks in you say to yourself, "*What do I do now*?"

However, both good and bad types are outnumbered by cases which are "neutral" (or cases, where it is impossible to tell one way or the other), and typical subjects like *implications*, *message*, *scale*, or *realisation* are not inherently positive or negative.

1.4.5 Discussion

Much of what Sinclair found in his analysis of *naked eye* holds true for words like *wreak*, *untoward*, *remiss*, and especially *sink in*. In every case, corpus analysis reveals—beneath the surface variation—patterns which appear repeatedly in the language data, including frequently co-occurring words and clear colligational preferences. As well as further undermining the idea of words as independent bearers of meaning, the examples discussed here support what Sinclair calls "the case for compound lexical items" which may be of considerable length.

The evidence of usage leaves little doubt that such extended units of meaning are a pervasive feature of everyday language and that "the independence of the choice of words is compromised, because other patterns cut across them and constrain them" (Sinclair 1996). To understand why our language output should be "compromised" in this way, and why it relies so much on recurrent patterns (of whatever length), it is helpful to invoke Michael Hoey's notion of "lexical priming". Hoey proposes that "every word is primed for use in discourse as a result of the cumulative effects of an individual's encounters with the word" (Hoey 2005: 13). As the data for *sink in* suggest, we are likely to encounter this verb in one of a limited set of contexts, showing one of a limited number of selectional and colligational preferences and (sometimes) semantic prosodies. These are its "primings", and they influence us when we use the word ourselves. The process is circular and self-reinforcing, and for Hoey, such

primings are "the driving force behind language use, language structure and language change" (Hoey 2005: 12).

2 Implications for practical lexicography

It is already well established, on the basis of empirical language study that words tend to occur frequently with certain other words in predictable patterns and contexts. This insight has informed many of the innovations in pedagogical dictionaries over the last 20 years, notably a sharper focus on collocation. In his 1996 paper, Sinclair showed that these networks of co-occurrence could be considerably more extensive than had been envisaged in earlier corpus studies. Therefore, if we accept that extended units of meaning are a significant feature of the language system, it follows that a dictionary which aims to describe normal usage should find ways of incorporating such information. This raises two questions: how can lexicographers identify recurrent extended units in a reliable and time-efficient way, and, once found, how should these extended units be accounted for in dictionaries?

2.1 Finding extended units in a corpus

Since the beginnings of corpus-based lexicography, efforts have been made to identify recurrent patterns in the language and to describe them in dictionaries. When the main (or only) analysis tool was the concordance, finding patterns could be a laborious process, and one whose outcomes were not necessarily complete or systematic. However, lexical profiling software, of which the best-known example is the Word Sketch (Kilgarriff et al. 2004), has transformed this operation. A Word Sketch presents the lexicographer with lists of a word's most significant collocates. Lists are grouped according to the grammatical relations they instantiate (such as verb+NP, ADJECTIVE+noun, and ADVERB+adjective), and collocates can be ranked according to their frequency or their salience. Word Sketches also list prepositions which typically follow a word (and in some cases those that precede it), and sometimes also list "constructions" such as *that*-clauses or infinitive clauses. In all cases, a further click takes the user to a concordance of the selected pattern. The automatic detection of patterns like these is a well-researched topic in natural language processing, and methods for extracting this information are well established and widely used. Finding shorter units of meaning (typically, two words which regularly co-occur) is now a relatively straightforward process.

If we broaden our search horizons to take in Sinclair's extended units of meaning, what kind of software tools will we need? In Sketch Engine, some progress has already been made in this direction, with two features which have been added fairly recently: multi-word sketches and "longest-commonest match" (Kilgarriff et al. 2015). For a multi-word sketch, the starting point is any two-word collocation, and the observation that in many cases a third collocate is found in the corpus data. For example, when we look at instances of the common collocation *seek+advice*, we

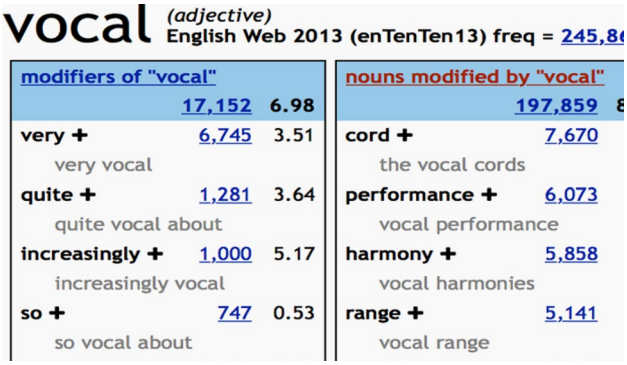


Fig. 3 Extract from a Word Sketch for *vocal*

find that many include an adjective referring to the *type* of advice being sought: *professional*, *legal*, *medical*, *financial*, and so on. These are effectively three-word collocations, and in the large corpora available now, it is easy to find numerous examples of this type. The second feature, longest-commonest match, is based on the concordance for a search word, and identifies any multi-word string which accounts for a high proportion of the corpus instances. (Technical aspects of this feature—how the algorithm works, and how it could be improved—are discussed in Kilgarriff et al. 2015.)

To give an idea of how these functions work, Fig. 3 shows an extract from a Word Sketch for the adjective *vocal*. The corpus used is the very large (> 20-billion-word) EnTenTen13 Web corpus of English (freely available in Sketch Engine), and in this case, collocates are ranked according to frequency, not salience. The screenshot shows the top four collocates for two grammatical relations:

Each of the eight collocates shown in this extract has a multi-word string, in grey, below the collocate, and this is the longest-commonest-match. Thus, at the collocate *quite*, the algorithm has detected that the sequence "quite vocal about" is especially frequent. All the collocates in this extract (*very*, *quite*, *increasingly*, etc.) are shown in bold and followed by a + sign, and this gives the user access to a multi-word sketch. Clicking the + symbol at increasingly brings up a new Word Sketch for the composite item *increasingly vocal*, which occurs 1000 times in this corpus (Fig. 4):

This multi-word sketch shows that combinations such as *increasingly vocal critic* are fairly common and that (see top right-hand column) prepositional phrases with *in* or *about* often follow *increasingly vocal*. However, the word which occurs most frequently with *increasingly vocal* is the verb *become* (in the bottom right-hand column), accounting for 269 instances out of the 1000 examples. Because of the frequency of this combination, the word *become* has its own + symbol, and this brings up an even more granular multi-word sketch for the combination "become increasingly vocal" (Fig. 5):

We have now drilled down almost as far as we can, but our final observation is that this three-word string is often followed by a PP with either *about* or *in*. If we click on one of the numbers (64 or 62), we bring up a concordance like this (Fig. 6):

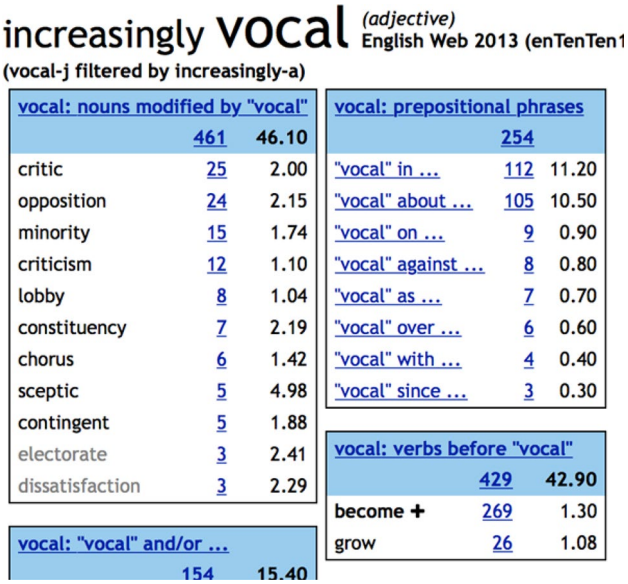


Fig. 4 Extract from a Word Sketch for *increasingly vocal*

Fig. 5 Extract from a Word Sketch for *become increasingly vocal*

There are 62 instances of the multi-word unit "become increasingly vocal in". Even here, the highlighted words to the right of the node draw our attention to nouns which are common collocates of *vocal*. At this level, it is feasible to scan all the concordances in the "old-fashioned" way, and this reveals that the seven word string "become increasingly vocal in [one's] criticism of" (with 13 of the 64 concordance lines) is the commonest pattern—and a clear example of what Sinclair referred to as an extended unit of meaning.

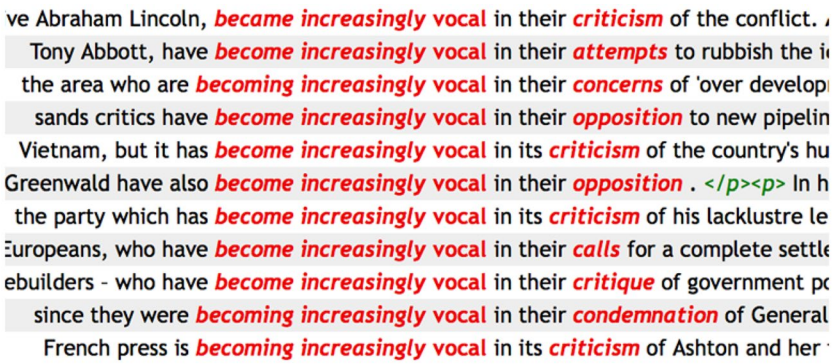


Fig. 6 Extract from a concordance for *become increasingly vocal in*

We have seen that even currently-available corpus-querying tools—provided they are working with sufficiently large corpora—can help us to detect recurrent multi-word units of considerable length. For purposes of linguistic research, the software already works well, but it does not yet meet the needs of time-poor lexicographers. Their task is to analyse corpus data and identify all of the lexicographically-relevant facts about a word or phrase—effectively, all frequently-occurring patterns of whatever type—and to complete this operation as rapidly as possible. This requires a high degree of automation. Automation works optimally when lexicographers are not required to make too many subjective choices, and can feel confident that the software has provided them with a complete set of each linguistic feature they are looking for (see, e.g., Rundell and Kilgarriff 2011). The "classic" Word Sketch meets both these needs. However, the recently-added functionality which facilitated our analysis of *vocal* is not yet ideal for lexicographic purposes. It raises doubts about the level of "recall" (have all relevant extended units been found?) and it adds a degree of complexity to the task which will result in a significant (and probably unacceptable) overhead in terms of time.

None of this is unsurmountable. The search mechanisms do their job and the information is all there if you have the time and skill to find it. Therefore, the issue is largely one of optimising the presentation, so that all relevant information is made available to the working lexicographer in an easy-to-digest form.

Possible improvements include a function which extracts what the software sees as the most significant multi-word strings in which a search word participates, and presents them in a single list ranked by frequency or salience. A further refinement would be to classify such n-grams and present separate lists for different types of expression. These could include lexical collocations of more than two words (such as *seek professional advice*), or prepositional phrases that frequently follow the search word. Many of these *n*-grams will be semi-fixed patterns with slots for words belonging to a particular semantic set, and the system could show which actual words, or which types of word, typically fill the slot.

For example: *vocal in* [*one's*] [*criticism*, *opposition*, *condemnation*, *concerns*] or *X takes* [time marker] *to sink in*. The goal is to maximize the usefulness of the information for lexicographers, and this will call for some further processing, some design tweaks, and perhaps for the use of data visualization techniques.

2.2 Implementation: what goes in the dictionary, and why

Let us assume that improvements in corpus-querying software can be made which will provide lexicographers with a user-friendly overview of the significant extended units typical of a given search word. We then need to consider whether this is the kind of information dictionaries should include—and if so, how this might be done.

Over the last 30 years or so, corpus-based research has led to major changes in our understanding of how meanings are created and interpreted. The idea of words as semantically autonomous is no longer sustainable, at least for mainstream non-specialist discourse, and this insight is increasingly applied to the content of dictionary entries. In older monolingual dictionaries, the description of meaning was often limited to short definitions, and one-word translation equivalents performed a similar function in bilingual dictionaries. However, meaning is now understood to be distributed across longer sequences of words, and the scope of dictionary entries has broadened significantly to reflect this view, taking in a range of features such as selectional restrictions, constructions, collocation, and phraseological conventions.

Things now need to take another step forward. With larger corpora and more powerful search tools at our disposal, we are learning even more about the pervasiveness of what Sinclair called the "idiom principle", and of the extent to which meanings are conveyed through multi-word strings which may be longer than previously suspected. Given that the goal of corpus lexicography is to describe linguistic norms—to produce what Hanks calls "an inventory of normal uses of each word in a language" (Hanks 2013: 92)—it follows that dictionary entries should expand further, to incorporate the kind of information discussed in this paper.

There are clear benefits here for computational applications, such as word sense disambiguation. Indeed, as Sinclair predicted in his paper on extended units, the approach he proposes should mean that "some of the problems of conventional description are much reduced—for example, there will be little word-based ambiguity left when this model has been applied thoroughly" (Sinclair 1996). But what about human users of dictionaries? Many of the units described both by Sinclair and in this paper are compositional in nature. Expressions such as *barely detectable to the naked eye* or *increasingly vocal in one's* (*criticism/opposition* etc.) are certainly frequent, but they do not pose particular problems of comprehension. However, the same is true of many two-word collocations, but these are nevertheless regarded as worth describing, especially in pedagogical dictionaries. Even if their meaning is transparent, their form is often unpredictable, and the same rationale applies to longer units. As examples of normal, frequent usage, they can help dictionary users to understand the conventions of mainstream discourse, or sometimes of specific types of discourse. (Some extended units are especially typical, for example, of journalistic, academic, or technical registers.) Non-fluent speakers are thus provided

with models for production which will help them to avoid unnatural, non-idiomatic language.

In his last paper, written in 2007 and published posthumously, John Sinclair returned to this theme. Observing once again "the tendency of words to occur together more often than their frequency would predict", he saw far-reaching implications for lexicography, because "the definiendum…is no longer a simple entity, a headword" (Sinclair 2007/2010: 37). What does this mean in practical terms? In online dictionaries, the trend now is to show units like phrasal verbs (*let down*, *let on*) and idiomatic phrases (*let the cat out of the bag*) as separate headwords in their own right (instead of being appended to a base lemma such as *let*). However, extended units of the type discussed in this paper are in a different category. Rather than requiring the status of full and separate headwords (which would in any case be problematic because of the high degree of internal variation they exhibit), they belong more naturally at the main entry—for example, as part of the information provided at words such as *vocal*, *remiss*, or *sink in*. In most pedagogical dictionaries, entries for complex words like these already provide information about the word's syntactic behaviour and collocational preferences. The function of this is to help the user to understand the word more fully and to use it naturally in productive mode. A description of the extended units which are frequently associated with a word can be seen as a further enhancement of this type of information. Implementing this in dictionaries will no doubt involve tough decisions regarding design and presentation. But this is part of a wider challenge which dictionary makers face. Adding new categories of data to dictionary entries which are already information-rich poses the challenge of how to give users access to the facts which they need at a given moment, while also making it easy for them to ignore data types which do not interest them (see Rundell 2015: 308–309 for a fuller discussion). However, without the space constraints of paper-based dictionaries, there is no reason why we cannot solve these problems, and the resulting dictionary entries would provide a new and deeper level of information on how words naturally and typically behave and combine in text.

Acknowledgements I am grateful to Vojtěch Kovář of the Sketch Engine team for his helpful comments on the functions discussed in Sect. 2.1.

References

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

Convery, C.Ó., P. Mianáin, M.Ó. Raghallaigh, S. Atkins, A. Kilgarriff, and M. Rundell. 2010. The DANTE Database (Database of ANalysed Texts of English). In *Proceedings of the XIV EURALEX Congress*, ed. Anne Dykstra, and Tanneke Schoonheim. Leeuwarden: Fryske Akademy.

Cowie, A.P. 1999. *English Dictionaries for Foreign Learners: A History*. Oxford: Oxford University Press.

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

Halliday, M.A.K. 1966. Lexis as a Linguistic Level. In *Memory of J. R. Firth*, eds. C.E Bazell, J.C Catford, M.A.K. Halliday, R.H. Robins. 148–162. London: Longman.

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

Johnson, S. 1755. *Preface to A Dictionary of the English Language*. Edited by Jack Lynch. http://andromeda.rutgers.edu/~jlynch/Texts/preface.html.

Kilgarriff, A., P. Rychly, P. Smrz, and D. Tugwell. 2004. The Sketch Engine. In *Proceedings of the Eleventh Euralex Congress*, ed. Geoffrey Williams and Sandra Vessier, 105–116. France: UBS Lorient.

Kilgarriff, A., Baisa, V., Rychlý, P., Jakubíček, M. 2015. Longest–commonest Match. In *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of the eLex 2015 conference*, ed. Kosem, I., Jakubíček, M., Kallas, J., Krek, S, 397–404. Ljubljana/Brighton

Rundell, M. 2015. From Print to Digital: implications for Dictionary Policy and Lexicographic Conventions. *Lexikos* 25: 301–322.

Rundell, M., and A. Kilgarriff. 2011. Automating the Creation of Dictionaries: Where Will It All End? In *A Taste for Corpora. A tribute to Professor Sylviane Granger*, ed. F. Meunier, S. De Cock, G. Gilquin, and M. Paquot, 257–281. Amsterdam: Benjamins.

Sinclair, J.M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J.M. 1996. The Search for Units of Meaning. *Textus* 9 (1): 75–106.

Sinclair, J.M. 1998. The Lexical Item. In *Contrastive Lexical Semantics*, ed. E. Weigand, 1–24. Amsterdam: Benjamins.

Sinclair, J.M. 2007/2010. Defining the Definiendum. In *A Way with Words: Recent Advances in Lexical Theory and Analysis - A Festschrift for Patrick Hanks,* ed. G-M. de Schryver, 37–47. Kampala: Menha Publishers.

Summers, D. (ed.). 1993. *Longman Language Activator*. London: Longman.