14

What can a corpus tell us about lexis?

David Oakey

1 Corpus linguistics and lexis

A corpus linguistic approach to language study naturally leads to a focus on linguistic form: Corpus analysis software first presents language as written forms to be seen by the researcher on a screen, rather than as sounds in the air to be recorded and transcribed, or as processes in the mind to be inferred through experiment. In the English language, the basic meaningful written form is the word, and the study of words, their forms and meanings, all falls within the scope of lexis. Corpus linguistics, therefore, has a lot to tell us about lexis.

Lexis is usually viewed as one side of a separation between the form of a language and its meanings, between its syntax and semantics, and between its grammatical structures and the words used in them. Many decades of analysis of ever-larger text corpora, however, have revealed patterns of language use that blur the traditional boundaries between lexis and grammar, and a common theme of corpus linguistics has been a rejection of a clear distinction between the two systems. Halliday (1991: 32), for example, developed the notion of *lexicogrammar* which views lexis and grammar as opposite ends of a single continuum, in a similar way to how waves and particles are seen as complementary aspects of light, and the use of corpus data was crucial to the development of Halliday's ideas (Halliday, 1966: 159 cited in Oakey 2020: 3). Lexis, as seen by corpus linguists, now involves not only the meaning relations between words themselves but also how meanings arise from the grammatical configurations in which words are used in the real world.

Corpus linguistics has thus had an effect on researchers' theoretical stances towards language: When electronic corpora were first collected in the 1960s, they were in direct opposition to the dominant linguistic research paradigm of the time, which focused on internal mental representations of language for which corpus evidence was irrelevant. In the opinion of a contemporary reviewer, for example, 'many linguists will be uninterested in pursuing their researches into LANGUAGE with the questionable aid of a million words of typographic USAGE' (Maverick 1969: 75) (emphasis in the original).

DOI: 10.4324/9780367076399-14

Since then, however, within theoretical linguistics more value has been placed on evidence from language use. Murphy (2010: 5), for example, has pointed out that theoretical models of the mental lexicon (language "in here" or "I-language") need to be consistent with the observed features of usage (language "out there" or "E-language") that are provided by reference to a corpus. Taylor (2012: 1) similarly argues that while the goal of linguistic theory should be a theory of language in the mind, it 'must begin with a study of language as encountered' and that I-language and E-language should be 'aligned as closely as possible'. Taylor even goes on to liken language as represented in the brain to a corpus, and states that 'knowledge of a language consists in knowledge of the kinds of facts that are recorded in a corpus and that can be extracted from it' (ibid: 3). From corpus linguists there have also been proposed lexically based theories of language which are based primarily on corpus evidence: Lexical priming (Hoey, 2005) and the theory of norms and exploitations (Hanks 2013), for example, both draw on evidence from language use and prioritise the role of lexis over grammar in making meaning.

Thus the lexis described in this chapter is both similar to and different from lexis from the pre-corpus era. The following sections describe how corpus linguistics can offer the researcher lexical insights in a wide range of areas and highlight where our existing knowledge of language can be extended and where new discoveries can be made.

2 Word frequency lists

The simplest use of a corpus in relation to lexis is to show word frequency (see Chapter 10, this volume), and counting words pre-dates the invention of computers by several centuries. The earliest wordlists are in the form of concordances, or indexes, to sacred texts such as the Bible or the Koran. These concordances aim to reveal more to followers of a religion by providing an index of where each word occurs in the text, together with some of the surrounding context to show how the word is used.

The forerunners of modern corpus linguists proceeded from counting and indexing words in single texts to collecting ever-larger samples from many texts which represented the variety of language being studied. Producing concordances by hand is a labour-intensive task, and manually counting words and their meanings in these text collections was an effort that 'still boggles the mind' (Gilner 2011: 69). In the first half of the twentieth century, psychologist George Zipf used a corpus of 44,000 words of American newspaper English to investigate the relationship between word length, variety, and frequency (Zipf 1935: 24). In the field of English language teaching Irving Lorge and Edward Thorndike, in their *Semantic Count of English Words*, created a list of words and their different senses (Lorge and Thorndike 1938), cross-referenced with the Oxford English Dictionary, based on various corpora eventually totalling 4.5 million words. Various scholarly committees then combined several existing wordlists, culminating in West's *General Service List of English Words* (West 1953) of 2,000 headwords and their most frequent meanings and derivations.

The introduction of computers to the study of language built on this previous work and essentially continued doing things the same way, only faster. In 1964, computer-generated English word counts were produced from the million-token Brown Corpus of American written English, based on samples of texts from different varieties of fiction and non-fiction (Francis and Kučera 1964/1979). Counting words was still an intensive use of resources, even with computerisation, and it took a million-dollar IBM 7070

mainframe computer with 50Kb of RAM '14 hours of continuous dedicated processing with the aid of six tape drives to construct the first word list' (Kučera 2002: 307). Today's researchers, by contrast, have a tremendous amount of computing power on hand, either on their desktop or in the cloud.

In the decades since the Brown Corpus, corpora have grown ever larger, from the COBUILD Bank of English (Sinclair 1987) of 18 million tokens, the British National Corpus (1994) of 100 million tokens, the Corpus of Contemporary American English (Davies 2008-) of 1 billion tokens, to corpora taken automatically from webpages such as the English Web 2020 corpus of 38 billion tokens on *Sketch Engine* (Kilgarriff *et al.* 2004, 2014). The majority of these resources are accessible to researchers online, usually by subscription, and all of the observations about lexis in this chapter can be replicated by the reader.

A wide variety of corpus-derived wordlists has since been compiled in this pre-corpus tradition, although the tendency of list makers to include the frequent senses of each word in the list in the manner of Thorndike, Lorge, and West has fallen away. It is now customary to use wordlists as a way of evaluating the words which occur in a text or texts. List makers now refer to how many of the word forms in a text or corpus are matched by the word forms in a particular wordlist, known as the "coverage" of a text by a list, with no information about the different meanings of these word forms (e.g. Nation 2013: 16). Information about word sense frequencies is now instead to be found in language learner dictionaries, which, since the introduction of the COBUILD dictionary in 1987, have used corpora as a basis for their definitions and list the different meanings of a headword in order of their frequency in the corpus (see Chapter 28, this volume).

The most frequent words in any wordlist from a corpus of written English are the "grammatical" or "function" or "closed-class" words such as *the*, *to*, and *of*. These have little meaning in themselves in a wordlist but are essential for building meaning by combination with other words. It can be seen from Table 14.1 that the top ten most frequent words in the 12 stories by Arthur Conan Doyle comprising *The Adventures of Sherlock Holmes* (Conan Doyle 1892) are grammatical words. The list was obtained from *Sketch Engine* (op. cit.) (see Chapter 9, this volume for more on software). It can be seen that, in addition to the usual common grammatical words, the wordlist contains *be* and *have.* This is because in this instance the tool has been instructed to count lemmas, and so the 4,566 occurrences for *be* also include other forms of the lemma such as *is*, *was*, *were*, *are*, *been* and *being*. Rows 211–218 from lower down the frequency list show how much less common lexical words are such as *wife* and *mind*. Indeed, word frequencies in different corpora, whatever size, display similar distributions: A very small number of words occur very frequently, a lot more words occur infrequently and around half the words occur only once.

Word frequency lists can be displayed graphically instead of as a table, and Figure 14.1 shows the full wordlist for *The Adventures of Sherlock Holmes* displayed in different visualisations. The word clouds in Figure 14.1a and b reveal how much grammatical words dominate the wordlist. Word clouds, in which the size of the word is proportional to its frequency of occurrence, can display word frequency information quickly in a small amount of space; frequent words are those which are legible, while infrequent words are displayed but are too small to be read. Figure 14.1a includes grammatical words, which are the most visible in the cloud, whereas Figure 14.1b omits

Table 14.1 Excerpts from word frequency lists for *The Adventures of Sherlock Holmes* in *Sketch Engine*

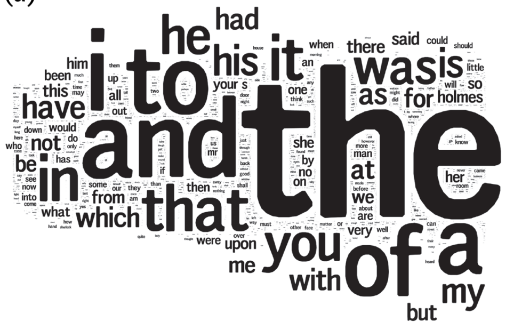
	Sketch Engine	
Rank	Word	Frequency
1	the	5601
2	be	4566
3	i	3032
4	and	3001
5	to	2681
6	of	2645
7	a	2623
8	have	2125
9	in	1757
10	that	1743
–	–	–
211	lie	61
212	mind	61
213	between	60
214	fact	60
215	run	60
216	wife	60
217	chair	59
218	course	59
219	drive	59
220	meet	59

grammatical words, and so far fewer words are immediately legible, and the differences between the size of these words and that of less frequent words is noticeably smaller.

The other way of visually displaying a wordlist is through frequency charts. Figure 14.1c shows the word frequencies in two ways: A line graph shows the cumulative percentage of the words in the corpus at each point in the frequency list. Thus *give* is the 213th word form in the list and at that point these 213 word forms – known as types – account for 67 per cent of all the words – known as tokens – in the corpus. Figure 14.1c also shows a bar chart of the word frequencies, but the precipitous fall in frequencies at the top of the list, and the very large number of words which occur only once in the corpus, means that the bars are difficult to see. Word frequencies are therefore often plotted on a log scale, shown in Figure 14.1d, which plots a word's frequency against its rank in the word list. The log scale, in this case to the power of 10, allows both large and small frequencies to be visible. The most frequent word in the list, *the*, has a rank of 1 and occurs 5,601 times, while the least frequent words in the list, each of which occurs only once, comprise 44 per cent of the list. The roughly straight line made by the points on the chart suggest a constant relationship between a word's frequency in a list and its rank in that list, a property first observed in written corpora in the pre-computer era in the study by Zipf in 1935.

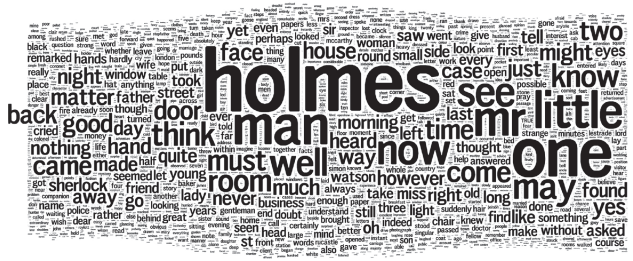
Frequency lists for modern large corpora provide more information about the words in the corpus. The word frequency list for the 1-billion-token *Corpus of Contemporary American English* (COCA) in Table 14.2 shows the frequencies of words in the corpus as

(a)



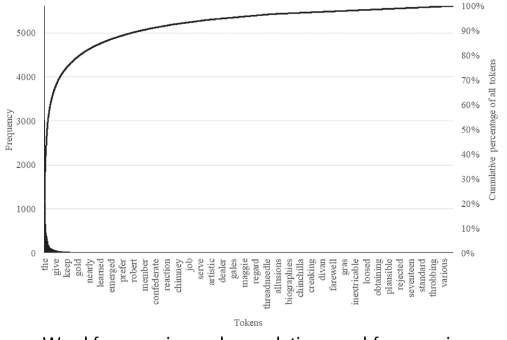
Word cloud for *The Adventures of Sherlock Holmes* with grammatical words included

(b)



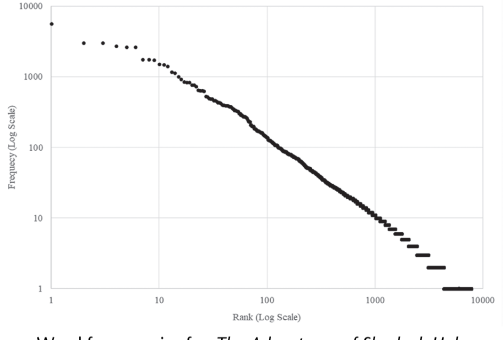
Word cloud for *The Adventures of Sherlock Holmes* with grammatical words excluded

(c)



Word frequencies and cumulative word frequencies for *The Adventures of Sherlock Holmes*

(d)



Word frequencies for *The Adventures of Sherlock Holmes* on log scale

Figure 14.1 Different ways of graphically presenting a word frequency list for *The Adventures of Sherlock Holmes*

Table 14.2 Sample from the word frequency list for the *Corpus of Contemporary American English* (COCA)

Rank	*Word*	*Freq*	*#texts*	*%caps*	*Blog*	*Web*	*TVM*	*Spok*	*Fic*	*Mag*	*News*	*Acad*
2585	partners	35,578	20,671	13	4414	4790	2578	2538	1523	5919	6370	7446
2595	closely	35,323	28,239	1	3881	4767	1051	3976	3566	5285	4582	8215
2605	notion	35,179	23,089	0	4815	4811	756	3872	2258	4580	3452	10635
2615	turkey	34,960	12,175	58	3455	4613	2816	4164	1956	6206	6002	5748
2625	joint	34,802	20,524	23	2775	4531	2645	3715	2441	4900	5265	8529
2635	flowers	34,622	17,489	11	1932	2463	4747	2039	8473	9461	3660	1845
2645	refused	34,450	25,654	1	3813	4461	1703	3940	5874	4684	6288	3685
2655	figured	34,340	25,587	3	4108	3252	9668	2871	7401	3421	2751	867
2665	shots	34,212	19,679	6	4973	3679	4137	3585	3512	6471	6862	993
2675	ford	34,111	13,688	98	3203	2881	1991	5206	2559	7018	8983	2180

a whole and also in each of the different genres or varieties of English collected in the corpus: blogs, websites, TV programmes, spoken language, fiction, magazines, newspaper and academic prose.

The list shows that 58 per cent of occurrences of the word form *turkey* are capitalised and thus likely to refer to the country, while the other 42 per cent of occurrences are likely to refer to the bird. Similarly the vast majority (98 per cent) of occurrences of *ford* are capitalised and thus likely to refer to the American motor company rather than to a river crossing.

One of the strengths of COCA for the study of lexis is that it contains over 127 million tokens of unscripted speech. Spoken data is more difficult to collect than written data in terms of transcription costs and privacy concerns, and the analysis of large amounts of spoken language use in corpora is relatively recent. Studies of conversations in English corpora have revealed that certain words are used much more in speech than in writing, and important lexical differences between speech and writing have quickly become clear. Buttery and McCarthy (2012: 288), for example, in a comparison of word frequency lists from the spoken and written fiction sections of the *British National Corpus*, found that only 65 per cent of words occurred on both lists and many words that were frequent on one list were much less frequent in the other list. Evaluative adjectives ending in -y, such as *yucky*, *stroppy*, *comfy* and *grumpy*, were much more common in speech than in writing; nouns indicating facial expressions like *grimace*, *scowl*, *smirk* and *pout*, on the other hand, occurred much more frequently in writing than in speech. Words used for keeping a conversation going, response tokens such as *right*, *yeah* and so on, are also higher in frequency lists for speaking than in those for writing.

3 Words in context: concordance lines

Corpus analysis software has adapted another method from the pre-corpus linguistics era mentioned earlier, that of concordancing. Instead of an index of the location of each word in a corpus, concordancing software presents lines of text showing the search word in the centre of the screen and a pre-set number of characters surrounding the search term; these concordance lines are also known as Key Word-In-Context (KWIC) lines (see Chapter 9, this volume, for more on concordancing software). Concordance lines

show a word in its immediate lexical and grammatical environment and are ideal for investigating the relations between words, their forms and meanings.

Concordance lines offer a fascinating snapshot of a word's lexical and grammatical behaviour. The software drags snippets of speech and writing into alignment, regardless of context, so that at times the researcher feels like they are performing the equivalent of eavesdropping on private conversations or rifling through private correspondence. The context on either side of the aligned search word offers clues to the situation of language use it was used in and the register or variety constituted by that situation. Whether written or spoken, fiction or real life, the word is taken by the software and impartially aligned for inspection. The juxtaposition, for example, of daytime TV chat show, impassioned movie dialogue, dull instruction booklet or parliamentary debate can be intriguing.

Figure 14.2 shows a random sample of ten concordance lines for a search for the word form *remote* in COCA. Generally the context around *remote* in these lines is enough to reveal its part of speech and its meaning. It is used as part of a compound noun *remote control,* and as an attributive adjective is used to modify nouns like *area* or *areas*, *islands* and *places*. Words immediately to the right of *remote* are sorted so that words with initial letters nearer the beginning of the alphabet are higher up the screen. This helps the corpus linguist identify words that are used repeatedly with *remote*, such as *control* and *area.*

In the Collins COBUILD dictionary, based on analysis of vastly more examples of *remote* than in these few lines, the "far away" sense of *remote* is the most frequent meaning found:

> Remote areas are far away from cities and places where most people live, and are therefore difficult to get to.
>
> *(Collins COBUILD Advanced Learner's Dictionary 2018: 1270)*

and the nouns *area* and *places* seen in the surrounding context are used in the definition itself to reinforce their importance to this meaning of *remote*. Early corpus lexicographers manually read through concordance lines like these to identify the meanings of the words from the context. This was only possible, however, because the number of occurrences of less frequent words in these early corpora was low. Moon (2007: 166), for example, described how the Collins COBUILD dictionary definition for *skate* was based on the 35 examples found in the first Bank of English corpus of 7.5 million tokens. Modern corpora are much larger and contain many, many more occurrences of even low-frequency words; *skate*, for example, occurs nearly 300,000 times in the English Web 2020 corpus. It would of course be impossible for a corpus linguist to read all the concordance lines for the word and make systematic judgements on its meaning from the surrounding context in each case, and so most concordancing software allows a limited number of randomly sampled lines to be displayed. Word profiling software which generalises patterns from concordances, such as *Sketch Engine*, is normally used for lexicographical work on dictionaries and will be discussed more in Section 4 on collocation.

When corpus linguists began to look at large numbers of concordance lines, they noticed that words were repeatedly used in certain grammatical configurations. Patterns became immediately visible from the regularities in concordance lines arising from the font in which they were presented. Figure 14.3 shows concordance lines for *want*; words immediately to the right of *want* are sorted so that words nearer the beginning of the alphabet are higher up the screen. The repeated infinitive complementation of *want* by *talk*, followed by

bodies of two young men were found in this remote area called the Geronimo Trail outside Douglas, A
ust pack up the ole minivan and drive into remote areas and survive off the land are living in a fa
ese. Any American who has ever commanded a remote control knows the unavoidable truth about this na
 the space with fumes. The operator uses a remote control to keep a safe distance. Happily missing
lent and yet full of dimly heard echoes, a remote disturbance of mumbling voices, swept into town p
y because their home is on one of the most remote islands in the Pacific. It is also because they h
s like this. Listen. Mark grabs the STEREO REMOTE off the kitchen counter and turns up the volume t
ve comes first, always, drives me to these remote places. She didn't have the heart for it. So tell
nd that they needed to be redistributed to remote rural areas and re-educated. Ms-UNG: Yes. SIMON:
at the further away the story is, the more remote the population, the less interest the press has.

Figure 14.2 Concordance lines for *remote* in COCA

es and a data stick. "This is what they want to stop."# "Wait, "Virginia said.
thing very bad happened to her. I don't want to surround myself with that. So you wan
ompany, but before I accept the badge I want to take a moment to think through the
  "# Angela Rogers added: "I just don't want to take a risk on Obama. We just can't take
It means, what do you have that I might want to take from you in a bet? What do I have
 impeachment proceedings in Congress. I want to take that now to Democratic Congressman Eric
hout registering, which I really didn't want to take the time to do right now, but does what
llance and other things that he did not want to talk about, particularly because its name, Dirty
u changed your mind. You probably don't want to talk about it. You're doing the right thing.
s information is being compromised. You want to talk about the 1800s Erica, maybe you should go bac
ey. I heard about what happened. Do you want to talk about it? That's cool. Do you mind
     He's... He's not... I don't really want to talk about Glen. Okay. What about other family?
 if he did acid. Bradley said he didn't want to talk about that stuff. # Howard said he doesn't
en by your own definition of what women want to talk about, you'd pass it. According to wikipedia
     she said angrily. Versus: "I don't want to talk about it," she said and smacked her hand

Figure 14.3 Sample concordance lines for *want* in COCA sorted one, two and three words to the right of the search word

post-modifying prepositional phrases beginning with *about,* causes vertical white lines to be visible to the right of *want* which are not observable on the left of the word.

Patterns in concordance lines like this revealed a strong relationship between the meaning of a word and the grammatical pattern it was used in and led to the notion of pattern grammar (Hunston and Francis 2000) in which words are used in similar grammatical configurations when used with similar meanings. Generalisations from concordance lines like *want to talk* led to the formulation of grammar patterns, such as **V to-inf**, which were added to COBUILD dictionary definitions. Grammar patterns have similarities to the concept of "construction", a concept from cognitive linguistics where lexis takes on meaning by being used in a particular grammatical sequence (Hunston and Su 2017: 570), and are an example of how corpus linguistics has blurred the boundaries between traditional grammar and lexis.

4 Collocation and semantic prosody

Corpus linguistics has had a lasting effect on the study of collocation, an important area of lexis also identified in the pre-corpus linguistics era. It has long been known that particular words have a tendency to combine more often with some words rather than with others. The twentieth-century British linguist J. R. Firth's famous example was *strong tea* (Firth 1957). He pointed out that while words in theory can be combined in many different ways, allowing for grammatical constraints, language users largely prefer to use particular combinations more than others, and supposedly synonymous words can sound "odd", particularly to native speakers of the language, when combined. The word *powerful* is regarded as a synonym of *strong*, yet the collocation *powerful tea* seems at the very least to mean something different than *strong tea*. Corpus linguistics has enabled researchers to quantify these typical combinations and calculate the probability of their co-occurrences in a corpus so that the likelihood of a particular word being followed by another particular word can be predicted.

Frequent collocations of a word can easily be observed from the surrounding context in concordance lines. In Figure 14.2, as we have already seen, concordance lines show that the word form *remote* occurs with *control*, *area, areas*, *islands* and *places*. In larger corpora, word profiling software such as *Sketch Engine* summarises the many hundreds of thousands of concordance lines for occurrences of words like *remote* to produce a behavioural profile known as a "word sketch". Table 14.3 shows part of a word sketch for *remote*.

Sketch Engine lists collocations in terms of their typicality rather than their absolute frequency, here based on how often *remote* collocates with a word rather than with other words. The adjective *remote* collocates with nouns like *control*, *villages*, *locations* and *server* much more frequently than it does with other nouns.

The learning of collocations has long been seen as crucial to attaining proficiency when learning a language (Palmer 1933; Sinclair *et al.* 1970/2004; Lewis 1993; Nesselhauf 2003; Oakey 2010; Szudarski 2018). Through the *English Vocabulary Profile* site (Capel 2015) it is possible to see the collocations used by English language learners at different proficiency levels, as measured against the *Common European Framework of Reference for Languages* (CEFR) (Council of Europe 2021). The site summarises results from the Cambridge Learner Corpus, a collection of hundreds of thousands of exam papers by English learners from the lowest proficiency level, A1, to the highest, C2. In the case of *remote*, the corpus shows that learners at level B2 are able to use the word with its most frequent sense of "far away," collocating with nouns such as *area*. More proficient learners at level C2, in addition, can

Table 14.3 Excerpt from the word sketch for *remote* based on the English Web 2020 corpus in *Sketch Engine*

Modifiers of remote	*Nouns modified by remote*	*Verbs complemented by remote*
geographically	control	shake
geographically remote	*remote control*	*shake the Wii Remote*
impossibly	location	program
impossibly remote	*remote locations*	*program a firestick remote*
partially	server	use
Full-time, partially remote position	*a remote server*	*using the Wii remote*
infinitely	attacker	swing
infinitely remote	*allows remote attackers to*	*swing the Wii Remote*
however	village	grab
however remote	*remote villages*	*grabbed the TV remote*
live	monitoring	point
a live remote from	*remote monitoring*	*pointing the Wii Remote*
exceedingly	access	tilt
exceedingly remote	*remote access to*	*by tilting the Wii Remote*
extremely	island	pair
extremely remote	*a remote island*	*pair a different firestick remote*
inconceivably	viewing	programme
inconceivably remote	*remote viewing*	*programming your Harmony remote*
relatively	computer	connect
a relatively remote	*the remote computer*	*Connect Virtual Technician remote*
fairly	working	replace
a fairly remote	*remote working*	*replaces the original Wii Remote*

use *remote* with its less frequent senses of "far away in time", collocating *remote* with *past*, and "far away from reality", collocating *remote* with *possibility*.

Corpus linguistics studies language in use as a moving target. The COVID-19 pandemic offers many examples of the kind of real-world lexical shift which corpus linguists are well placed to study. In February 2020 the disease caused by the SARS-CoV-2 virus was officially named by the World Health Organization as *COVID-19* (WHO 2020). It then began to be written and spoken about in the news and on social media, and this writing and speech began to be archived in online corpora; in turn, results from corpus-informed studies into the lexis of COVID-19 in English swiftly appeared. The Oxford English Dictionary (OED 2020) provided empirical proof, from their 8-billion-token monitor corpus of web-based news which is collected almost in real time, of the vertiginous increase in use of the words *coronavirus* and *COVID-19* between December 2019 and March 2020 and listed frequent collocations such as *outbreak, infection, spread* and *fear*. Also prior to December 2019, similar to the other corpora surveyed here, the OED corpus showed that the most common collocates of *remote* were familiar nouns like *control*, *island* and *village*. However, restrictions on office working imposed as a result of the COVID-19 pandemic led to many employees having to work from home. The proportion of people working remotely rose from 5.7 per cent of workers in January/February 2020 to 43.1 per cent in April 2020 (Felstead and Reuschke 2020). By the end of 2020, consequently, the most frequent collocates of *remote* had suddenly become *learning*, *working* and *work force* (Schuessler 2020).

Evidence on collocations from corpus linguistics has blurred traditional category boundaries in lexicology. Both synonymy and antonymy are paradigmatic lexical relations between words: When synonyms are substituted, there is no change in the propositional meaning of the sentence as a whole (Carter 2012: 34), whereas when antonyms are substituted, the propositional meaning of a sentence becomes opposite. Corpus evidence, however, has shown that antonyms are actually used together in sentences rather than instead of each other; antonyms like *right and wrong* and *high and low* co-occur much more often than by chance (Jones *et al.* 2012), and evidence of such collocations shows that syntagmatic relations are also important in understanding their use.

The traditional view of synonymy has similarly been challenged by corpus linguistic work on collocations. Word processing software often has thesaurus tools which offer synonyms to help writers choose alternative words. The synonyms for *strong* listed on the software on my PC include *robust*, *sturdy* and *solid*, all of which would sound odd when used to describe tea. Use of such thesaurus synonym tools can be risky: Extensive word substitution by student writers using thesaurus suggestions has been termed *roget-ing* (Grove 2014: 7) and has been seen to result in meaningless combinations when collocational relations are unwittingly broken. Grove (ibid) gives the example of *left behind* transformed through roget-ing into *sinister buttocks* and *powerful personalised services* into *Herculean personalised liturgies*.

Liberman (2012) illustrates how unfamiliar collocations can sound "odd" using a statement by Mitt Romney, a presidential candidate in the 2012 US presidential election, who said in a speech that 'I was a severely conservative governor.' Liberman lists puzzled reactions to this statement from US political commentators and quotes Molly Ball of *The Atlantic* magazine as saying that the statement 'described conservatism as if it were a disease'. Liberman points out that reference to a corpus can help explain how the collocation "severely conservative" had such a poor reaction. The word cloud from COCA and the word sketch visualisation from the English Web 2020 in Figure 14.4 show the adjectives occurring one word to the right of *severely*; both images are very revealing about the lexical environment brought about by the use of the word.

The corpus evidence in Figure 14.4 shows the overwhelming tendency of *severely* to collocate with words with negative meanings. This is a fact about language use which is replicable: It can be observed by looking in both COCA and Sketch Engine, two independently collected corpora. This aspect of collocation has been found to be so widespread that it has been termed *semantic prosody*: 'the consistent aura of meaning with which a form is imbued by its collocates' (Louw 1993: 57). A word, like *conservative*, can take on a negative meaning purely by reason of its collocation with *severely*. The negative semantic prosody of *severely* in Liberman's example meant that Mitt Romney, while clearly intending to say something positive, instead was understood to be saying that he thought *conservative* was a negative quality.

An example of positive semantic prosody can be seen from the verb collocates of the noun *diversity,* as shown in Figure 14.5. The examples in the corpus show that diversity is seen as something to be *respected*, *promoted*, *celebrated*, *valued* and *embraced*, for example.

There may be language users for whom *diversity* has a negative semantic prosody, e.g. for ideological reasons, but they are either too few in the corpus or the users do not speak or write about it by using the word *diversity* in a way that can be picked up by corpus linguistic analysis.

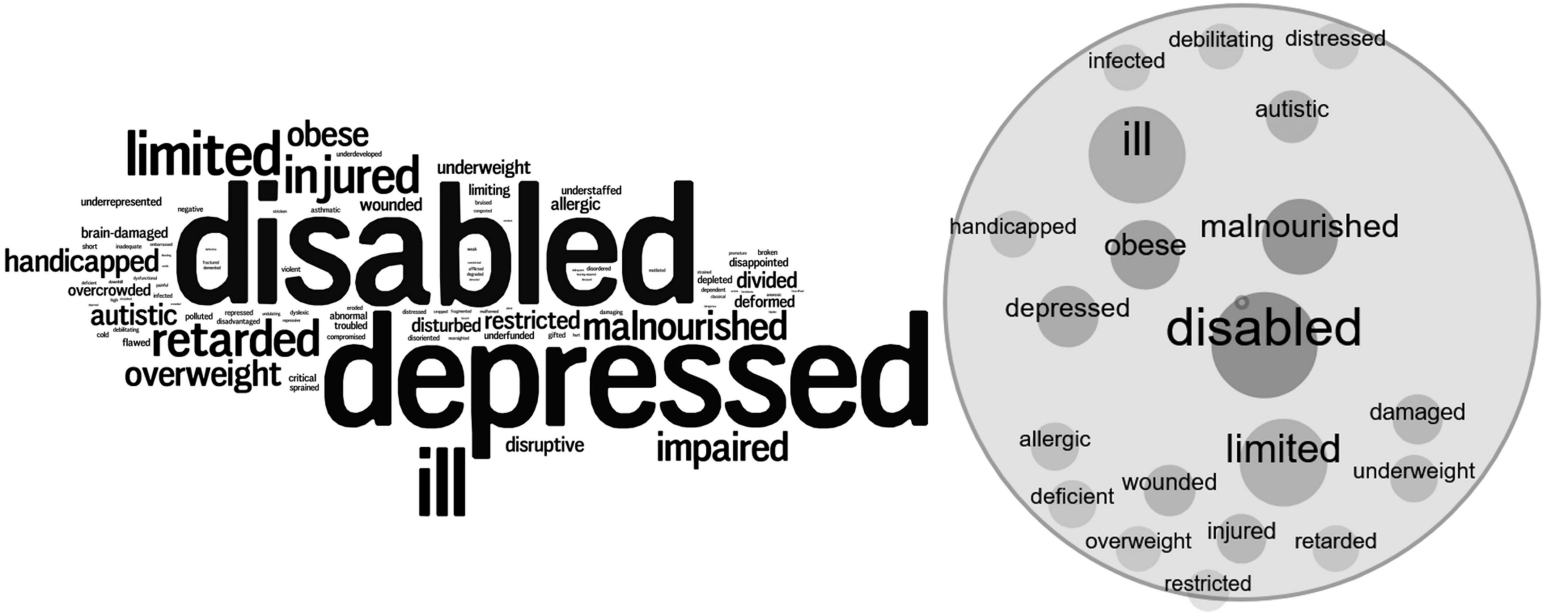


Figure 14.4 Adjectives modified by *severely* in COCA and English Web 2020 on Sketch Engine

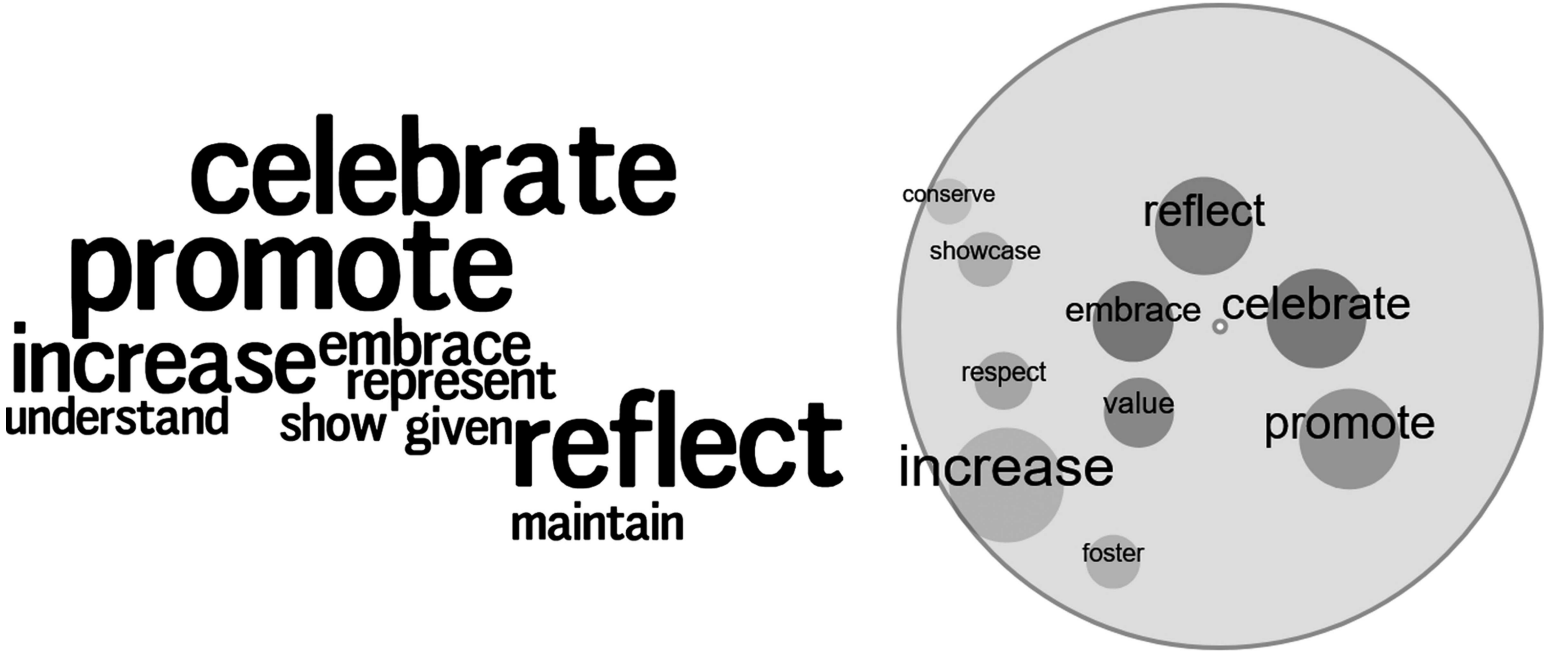


Figure 14.5 Verbs for which *diversity* is the object in COCA and English Web 2020 on Sketch Engine

The idea of semantic prosody is similar to the traditional lexical relation of connotation, which involves the positive or negative associations a word conjures up (Carter 2012: 36), but semantic prosody is less subjective, since it can be quantified through examining a word's collocations. Semantic prosody can also be criticised (Hanks 2013: 124), since it is implausible that every word can be divided semantically into objectively determined categories of positive and negative polarity. But corpus analysis shows how a word tends to be used, and the researcher can better determine the speaker's or writer's attitude to what they are talking about in specific contexts. Semantic prosody is another example of how corpus linguistics has blurred the boundaries between traditional grammar and lexis. Books on both subjects, in fact, deal with the concept in more detail: i.e. *Corpus Linguistics for Grammar* (Jones and Waller 2015) and *Corpus Linguistics for Vocabulary* (Szudarski 2018).

5 Metaphor

The patterns of lexis observed using corpus linguistics methods have been influential in revealing more about metaphor. Word meaning can be seen as literal and metaphorical, where features of one domain have been transferred to another target domain. The literal sense of the noun *crescendo*, for example, refers to sound becoming louder and more intense. The most frequent verb with *crescendo* as its object in COCA is *reach*, and Figure 14.6 shows example concordance lines for this collocation. It can be seen that only four lines (4, 5, 9 and 10) refer to uses of the word in its musical sense, in the form of *applause*, *cheering*, *music* and *musketry*.

Even then, the collocation of *crescendo* with *reach* means its original sense has changed to one of an endpoint, or climax, which is the kind of change which is resisted by some language users. The music critic Miles Hoffman (2013: 17) is adamant that 'you cannot "reach" a crescendo … even if you make the most enormous crescendo in the world, you will not have "reached" anything until you get to the top.' He believes that musicians will never accept that 'a word that for centuries has had one and only one precise meaning will, through repeated flagrant misuse, come to mean something else' (ibid). This example shows where a corpus linguist can document the process of changing use through which a word evolves new meanings according the lexical preferences of its users. The meanings of *crescendo* in the other lines are metaphorical, transferring this new climactic aspect of the meaning of *crescendo* to other things that can grow more intense, such as *outcry, speculation, attacks, sensation, concerns* and *hostility*.

Metaphorical as opposed to literal language has traditionally been seen as belonging to the domain of literature, but more recent work has instead argued that metaphors used in language reflect the way people think and that studying related metaphors in language can reveal conceptual networks (Deignan 2005: 4). A distinction has thus arisen between linguistic and conceptual metaphors, with the former providing evidence for the existence of the latter (Deignan 2017). Corpus evidence about lexis and meaning is thus becoming more relevant to the E-language and I-language dialectic mentioned in Section 1.

An area for future corpus linguistic work on metaphor is the COVID-19 coronavirus, which has already been shown to have greatly impacted lexical behaviour. Hunston (2020), using the 10-billion-token *News on the Web* corpus (Davies 2016-), described how language users have extended the meanings of words and combinations such as *isolation*, *social distance* and *lockdown* in order to accommodate new COVID-19–related meanings

1. as the outcry from animal rights activists reached a crescendo. JEFFREY-KAHN: We did acknowled
2. e speculation gained strength all fall and reached a crescendo when Lloyd Carr made his
3. list fervor. The attacks began in 1990 and reached a crescendo the following year in
4. st free throw attempt missed, the applause reached a crescendo, climaxing as the second shot
5. their way up Crescent Street. The cheering reached a crescendo. " I wouldn't ever try to do
6. 't it obvious? " The sensation in my chest reached a crescendo, and I felt sick. I leaped up
7. n trial just before public safety concerns reached a crescendo. In June 1992, public
8. # SOVIET HOSTILITY to the West over Dnitz reached a crescendo in the week that followed,
9. Connie. Then you can collapse. The music reached a tumultuous crescendo, and she was off the
10. Wayne's support as the Colonials' musketry reached a crescendo. Monckton fell so close to the

Figure 14.6 Concordance lines for *crescendo* as the object of *reach* in COCA

and pointed out that 'our thoughts have been guided by wartime metaphors' (Hunston 2020: 1). This fact is reflected in the observable use around the term *coronavirus* of collocations like *fight, battle* and *combat,* revealing that the virus is seen as an enemy that needs to be fought. In this respect the linguistic and conceptual metaphors around the coronavirus are similar to that observed by corpus linguists in reporting of the SARS outbreak in 2003 (Wallis and Nerlich, 2005), and in public discourse in general (Flusberg *et al.* 2018). If people talk about the response to the coronavirus using militaristic linguistic metaphors, it is because they think of it in similar terms.

This chapter has given a brief overview of the wide range of areas where corpus linguistics can offer the researcher lexical insights. With access to a corpus and basic software, the reader can follow up by counting words, identifying repeated patterns and investigating the relations between words, their forms and meanings and their use.

Further reading

Deignan, A. H. (2017) 'From Linguistic to Conceptual Metaphors,' in E. Semino and Z. Demjen (eds) *The Routledge Handbook of Metaphor and Language*, London: Routledge, pp. 102–17. (This chapter is a very readable 'way in' to using corpora to study metaphor and makes a convincing case for the necessity of using corpus data to inform research into this important area of lexis at the interface of E-language and I-language.)

Flowerdew, L. (2012) *Corpora and Language Education*, New York: Palgrave Macmillan. (A wide-ranging survey of applications of corpus linguistics to language teaching and learning, this book contains insights on features blurring the boundaries between lexis and grammar and their implications for learners of English.)

Hanks, P. (2013) *Lexical Analysis*, Cambridge, MA: MIT Press. (This book propounds a lexically driven theory of language reflecting the tendency of users to choose certain ways of expressing themselves. A fascinating overview of words and meanings and how their use changes over time.)

Hasselgård, H., Ebeling, J. and Oksefjell Ebeling, S. (eds) (2013) *Corpus Perspectives on Patterns of Lexis*, Amsterdam: John Benjamins. (This collection of papers illustrates pertinent questions of lexis that can be investigated by a corpus linguistic approach.)

Murphy, M. L. (2010) *Lexical Meaning*, Cambridge: Cambridge University Press. (This book is a readable survey of traditional concerns in the study of lexis and is useful for benchmarking corpus linguistic studies.)

References

British National Corpus (Version 1.0) (1994), Oxford: Oxford University Computing Service.

Buttery, P. and McCarthy, M. J. (2012) 'Lexis in Spoken Discourse', in J. P. Gee and M. Handford (eds) *The Routledge Handbook of Discourse Analysis*, London: Routledge, pp. 285–300.

Capel, A. (2015) 'The English Vocabulary Profile', in J. Harrison and F. Barker (eds) *English Profile in Practice*, Cambridge: UCLES/Cambridge University Press, pp. 9–27.

Council of Europe (2021) *Common European Framework of Reference for Languages* (CEFR), Available online at https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions.

Carter, R. (2012) *Vocabulary: Applied Linguistic Perspectives*, 2nd edn, London: Routledge.

Collins *COBUILD Advanced Learner's Dictionary*, 9th edn, (2018), London: Collins.

Conan Doyle, A. (1892) *The Adventures of Sherlock Holmes*, London: George Newnes.

Davies, M. (2008-) *The Corpus of Contemporary American English* (COCA), 1 Billion Words, 1990–2019. Available online at https://www.english-corpora.org/coca/.

Davies, M. (2016-) *Corpus of News on the Web* (NOW), *10 Billion Words from 20 Countries*. Available online at https://www-english-corpora.org/now/.
Deignan, A. (2005) *Metaphor and Corpus Linguistics*, Amsterdam: John Benjamins.
Deignan, A. H. (2017) 'From Linguistic to Conceptual Metaphors', in E. Semino and Z. Demjen (eds) *The Routledge Handbook of Metaphor and Language*, London: Routledge, pp. 102–17.
Felstead, A. and Reuschke, D. (2020) 'Homeworking in the UK: Before and During the 2020 Lockdown', WISERD Report, Cardiff: Wales Institute of Social and Economic Research. Accessed on 5 February 2021 from: https://wiserd.ac.uk/publications/homeworking-uk-and-during-2020-lockdown.
Firth, J. R. (1957) 'A Synopsis of Linguistic Theory 1930-1955', in J. R. Firth (ed.) *Studies in Linguistic Analysis*, Oxford: Basil Blackwell, pp. 1–32.
Flusberg, S. J., Matlock, T. and Thibodeau, P. H. (2018) 'War Metaphors in Public Discourse', *Metaphor and Symbol* 33(1): 1–18.
Francis, W. N. and Kučera, H. (1964/1979) *Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*, Providence, RI: Brown University Department of Linguistics.
Gilner, L. (2011) 'A Primer on the General Service List', *Reading in a Foreign Language* 23: 65–83.
Grove, J. (2014) 'A Crafty Cheek in Sinister Buttocks', *Times Higher Education*, 7 August 2014, p. 7.
Halliday, M. A. K. (1991) 'Corpus Studies and Probabilistic Grammar', in K. Aijmer and B. Altenberg (eds) *English Corpus Linguistics,* London: Longman, pp. 30–43.
Halliday, M. A. K. (1966) 'Lexis as a Linguistic Level', in C. E. Bazell, J. C. Catford, M. A. K. Halliday and R. H. Robins (eds) *In Memory of J. R. Firth*, London: Longmans, pp. 148–62.
Hanks, P. (2013) *Lexical Analysis*, Cambridge, MA: MIT Press.
Hoey, M. (2005) *Lexical Priming: A New Theory of Words and Language*, London: Routledge.
Hoffman, M. (2013) 'A Crescendo of Errors', *New York Times*, July 29, 2013. Accessed on 23 February 2020 from: https://www.nytimes.com/2013/07/29/opinion/a-crescendo-of-errors.html.
Hunston, S. E. (2020) 'Changing Language in Unprecedented Times', *University of Birmingham, Department of English News*, 8th April 2020. Accessed on 23 April 2020 from: https://www.birmingham.ac.uk/schools/edacs/departments/englishlanguage/news/2020/changing-language.aspx.
Hunston, S. E., and Francis, G. (2000) *Pattern Grammar: A Corpus Driven Approach to the Lexical Grammar of English*, Amsterdam: Benjamins.
Hunston, S. and Su, H. (2017) 'Patterns, Constructions, and Local Grammar: A Case Study of 'Evaluation'', *Applied Linguistics* 40(4): 567–93.
Jones, S., Murphy, M. L., Paradis, C. and Willners, C. (2012) *Antonyms in English: Construals, Constructions and Canonicity*, Cambridge: Cambridge University Press.
Jones, C. and Waller, D. (2015) *Corpus Linguistics for Grammar: A Guide for Research*, London: Routledge.
Kilgarriff, A., Rychlý, P., Smrž, P. and Tugwell, D. (2004) The Sketch Engine. *Information Technology*, Available at https://www.sketchengine.eu/wp-content/uploads/The_Sketch_Engine_2004.pdf.
Kilgarriff, A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P. and Suchomel V. (2014) 'The Sketch Engine: Ten Years On', *Lexicography* 1(1): 7–36.
Kučera, H. (2002) 'Obituary for W. Nelson Francis', *Journal of English Linguistics* 30: 306–9.
Lewis, M. (1993) *The Lexical Approach: The State of ELT and a Way Forward*, Hove: Language Teaching Publications.
Liberman, M. (2012) 'Severely X', University of Pennsylvania Language Log, 11th February 2012. Accessed 2 February 2020 from: http://languagelog.ldc.upenn.edu/nll/?p=3762.
Lorge, I. and Thorndike, E. L. (1938) *A Semantic Count of English Words*, New York: Institute of Educational Research, Teachers College, Columbia University.
Louw, B. (1993) 'Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies', in M. Baker, G. Francis and E. Tognini-Bonelli (eds) *Text and Technology: In Honour of John Sinclair*, Amsterdam: Benjamins, pp. 157–76.
Maverick, G. V. (1969) 'Review of "Computational Analysis of Present-Day American English" by Henry Kučera and W. Nelson Francis', *International Journal of American Linguistics* 35: 71–75.

Moon, R. (2007) 'Sinclair, Lexicography, and the COBUILD Project: The Application of Theory', *International Journal of Corpus Linguistics* 12(2): 159–81.
Murphy, M. L. (2010) *Lexical Meaning*, Cambridge: Cambridge University Press.
Nation, I. S. P. (2013) *Learning Vocabulary in Another Language*, 2nd edn, Cambridge: Cambridge University Press.
Nesselhauf, N. (2003) 'The Use of Collocations by Advanced Learners of English and Some Implications for Teaching', *Applied Linguistics* 24(2): 223–42.
Oakey, D. J. (2010) 'English Vocabulary and Collocation', in S. Hunston and D. J. Oakey (eds) *Introducing Applied Linguistics: Concepts and Skills*, London: Routledge, pp. 14–23.
Oakey, D. J. (2020) 'Phrases in EAP Academic Writing Pedagogy: Illuminating Halliday's Influence on Research and Practice', *Journal of English for Academic Purposes* 44: 1–16.
OED (2020) 'Corpus Analysis of the Language of COVID-19', *OED Blog*, 14th April 2020. Accessed on23 April 2020 from https://public.oed.com/blog/corpus-analysis-of-the-language-of-covid-19/.
Palmer, H. E. (1933) *Second Interim Report on English Collocations*: *Submitted to the Tenth Annual Conference of English Teachers*, Tokyo: The Institute for Research in English Teaching.
Schuessler, J. (2020) 'Oxford's 2020 Word of the Year? It's too Hard to Isolate', *New York Times*, 22nd November 2020. Accessed on 23 November 2020 from: https://www.nytimes.com/2020/11/22/arts/oxford-word-of-the-year-coronavirus.html.
Sinclair, J. M. (1987) *Looking Up: An Account of the COBUILD Project in Lexical Computing*, London: Collins ELT.
Sinclair, J. M., Jones, S. and Daley, R. (1970/2004) *English Collocation Studies: The OSTI Report*, London: Continuum.
Szudarski, P. (2018) *Corpus Linguistics for Vocabulary: A Guide for Research*, London: Routledge.
Taylor, J. R. (2012) *The Mental Corpus: How Language Is Represented in the Mind*, Oxford: Oxford University Press.
Wallis, P. and Nerlich, B. (2005) 'Disease Metaphors in New Epidemics: The UK Media Framing of the 2003 SARS Epidemic', *Social Science and Medicine* 60(11): 2629–39.
West, M. (1953) *A General Service List of English Words*, New York: Longmans, Green and Co.
World Health Organisation (WHO) (2020) 'Naming the Coronavirus Disease (COVID-19) and the Virus that Causes it', Accessed on 5 February 2021 from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it.
Zipf, G. K. (1935) *The Psycho-Biology of Language: An Introduction to Dynamic Philology*, Cambridge, MA: MIT Press.