Using corpora [and the web] as data sources for dictionaries

Adam Kilgarriff

1 Introduction

There are three ways to write a dictionary

- Copy
- Introspect
- Look at data.

The first has its place. Making checks against other dictionaries is a not-to-be-overlooked step in the lexicographic process. But if the dictionary is to be an original work, it never has more than a secondary role.

Introspection is central. The lexicographer always needs to ask themself, "what do I know about this word?" "how do I interpret this evidence?" "does that make sense?" But by itself, intuition does a limited and partial job. Asked "what is there to say about the verb *remember*?" we might come up with some facts about meaning, grammar and collocation – but there are many more we will miss, and some of our ideas may be wrong.

That leaves data, and for lexicography, the relevant kind of data is a large collection of text: a *corpus*. A corpus is just that: a collection of data – text and speech - when viewed from the perspective of language research.

A corpus supports many aspects of dictionary creation:

- headword list development
- for writing individual entries:
  - discovering the word senses and other lexical units (fixed phrases, compounds, etc.)
  - identifying the salient features of each of these lexical units
    - their syntactic behaviour
    - the collocations they participate in
    - any preferences they have for particular text-types or domains
  - providing examples
  - providing translations.

The chapter describes how a corpus can support each of these parts.

First, two apologies. First, I am a native speaker of English who has worked mostly on monolingual English dictionaries. Examples, and the experiences which inform the chapter, will largely be from English. Second, to illustrate and demonstrate how corpora support lexicography, one cannot go far without referring to the piece of software that is

the intermediary between corpus and lexicographer: the corpus query system. But this article is not a review of corpus query systems, so we simply use one – the one developed by the author and colleagues, the Sketch Engine (Kilgarriff et al 2004) – to illustrate the various ways in which the corpus can help the lexicographer. For a review of corpus query systems see Kilgarriff and Kosem (2012).

With ever growing quantities of text available online, faster computers, and progress in corpus and linguistic software, the field is changing all the time. By the time any practice is standard and widely-accepted, it will be well behind the latest developments, so if this article were to talk only of standard and widely-accepted practices it would run the risk of looking dated by the time it was published. Instead, I mainly describe recent and current work on projects I am involved in, arrogantly assuming that this coincides substantially with the leading edge of the use of corpora in lexicography.

2 Headword lists

Building a headword list is the most obvious way to use a corpus for making a dictionary. *Ceteris paribus,* if a dictionary is to have *N* words in it, they should be the *N* words from the top of the frequency list.

2.1 In search of the ideal corpus

It is never as simple as that, mainly because the corpus is never good enough. It will contain noise and biases. The noise is always evident within the first few thousand words of all the corpus frequency lists that I have ever looked at. In the British National Corpus[1] (BNC), for example, a large amount of data from a journal on gastro-uterine diseases presents noise in the form of words like *mucosa* – a term much-discussed in these specific documents, but otherwise rare and not known to most speakers of English.[2] Bias in the spoken BNC is illustrated by the very high frequencies for words like *plan, elect, councillor, statutory* and *occupational*: the corpus contains a quantity of material from local government meetings, so the vocabulary of this area is well represented. Thus keyword lists of the BNC in contrast to other large, general corpora show these words as particularly BNC-flavoured.

If we turn to UKWaC (the UK 'Web as Corpus', Baroni et al. 2009), a web-sourced corpus of around 1.6 billion words, we find other forms of noise and bias. The corpus contains a certain amount of web spam. We discovered that people advertising poker are skilled at producing vast quantities of 'word salad' which, at the time, escaped our automatic routines for filtering out bad material. Internet-related bias also shows up in the high frequencies for words like *browser* and *configure*. While noise is simply wrong, and its impact is progressively reduced as our technology for filtering it out improves, biases are more subtle in that they force questions about the sort of language to be covered in the dictionary, and in what proportions.[3]

2.2 Multiwords

Dictionaries have a range of entries for multiword items, typically including, for English, noun compounds (*credit crunch*, *disc jockey*), phrasal and prepositional verbs (*take after*, *set out*) and compound prepositions and conjunctions (*according to*, *in order to*). While corpus methods can straightforwardly find high-frequency single-word items and thereby provide a fair-quality first pass at a headword list for simple words, they cannot do the same for multiword items. Lists of high-frequency word-pairs in any English corpus are dominated by items which do not merit dictionary entries: the string *of the* usually tops the list of word-pairs, or bigrams.

The Sketch Engine has several strategies here: one is to view multiword headwords as collocations (see discussion below) and to find multiword headwords when working through the alphabet looking at each headword in turn.

Another is to use lists of translations. This was explored in the Kelly project (Kilgarriff et al 2012). The project worked on nine languages. First, we prepared and cleaned up a corpus headword list of around 6000 words for each language. Then, all the words on those lists were translated (by a professional translation agency) into each of the eight other languages, giving us a database with seventy-two directed language pairs.[4] We reasoned that where one language uses a multiword expression for a unitary concept (say, English *look for*) it was likely that other languages had a single word for the concept (*eg.*, French *chercher*, Italian *cercare*) and that when the Italian-to-English and French-to-English translators encountered *cercare* and *chercher*, they were likely to translate it as *look for*. So, although *look for* did not appear in the English source list, it appeared multiple times in the database as a translation. The strategy produced a modest number of multiword expressions.

2.3 Lemmatisation

The words we find in texts are inflected forms; the words we put in a headword list are lemmas. So, to use a corpus list as a dictionary headword, we need to map inflected forms to lemmas: we need to lemmatise.

English is not a difficult language to lemmatise as no lemma has more than eight inflectional variants (*be*, *am*, *is*, *are*, *was*, *were*, *been*, *being*), most nouns have just two (*apple*, *apples)* and most verbs, just four (*invade*, *invades*, *invading*, *invaded*). Most other languages present a substantially greater challenge. Yet even for English, automatic lemmatisation procedures are not without their problems. Consider the data in Table 1. To choose the correct rule we need an analysis of the orthography corresponding to phonological constraints on vowel type and consonant type, for both British and American English.[5]

Table 1: *Complexity in verb lemmatisation rules for English*

Lemma	***-ed, -s* forms**	**Rule**	***-ing* form**	**Rule**
Fix	*fixed, fixes*	delete *–ed, -es*	*fixing*	delete *–ing*

Care		*cared, cares*	delete *–d, -s*	*caring*	delete *-ing,* add *-e*
Hope		*hoped, hopes*	delete *–d, -s*	*hoping*	delete *-ing,* add *-e*
Hop		*hopped*	delete *–ed,* undouble consonant	*hopping*	delete *-ing,* undouble consonant
		hops	delete *–s*		
Fuse		*fused*	delete *–d*	*fusing*	delete *-ing,* add *-e*
Fuss		*fussed*	delete *–ed*	*fussing*	delete *–ing*
bus	AmE	*bussed, busses*??	delete *–ed/-s,* undouble consonant	*bussing*	delete *-ing,* undouble consonant
	BrE	*bused, bused*	delete *–ed*	*busing*	delete *–ing*

Even with state-of-the-art lemmatisation for English, an automatically extracted lemma list will contain some errors.

These and other issues in relating corpus lists to dictionary headword lists are described in detail in Kilgarriff (1997).

2.4 User profiles

Building a headword list for a new dictionary (or revising one for an existing title) has never been an exact science, and little has been written about it. Headword lists are typically extended in the course of a project and are only complete at the end. A good starting point is to have a clear idea of who will use your dictionary, and for what purpose: a 'user profile'. A user profile "seeks to characterise the typical user of the dictionary, and the uses to which the dictionary is likely to be put" (Atkins & Rundell 2008: 28). This is a manual task, but it provides filters with which to sift computer-generated wordlists.

2.5 New words

As everyone involved in commercial lexicography knows, neologisms punch far above their weight. They might not be very important for an objective description of the language but they are loved by marketing teams and reviewers. New words and phrases often mark the only obvious change in a new edition of a dictionary, and dominate the press releases.

Mapping language change has long been a central concern of corpus linguists and a longstanding vision is the 'monitor corpus', the moving corpus that lets the researcher explore language change objectively (Clear 1988, Janicivic & Walker 1997). The core method is to compare an older 'reference' corpus with an up-to-the-minute one to find words which are not already in the dictionary, and which are in the recent corpus but not in the older one. O'Donovan & O'Neill (2008) describe how this has been done at Chambers Harrap Publishers, and Fairon et al. (2008) describe a generic system in which users can specify the sources they wish to use and the terms they wish to trace.

The nature of the task is that the automatic process creates a list of candidates, and a lexicographer then goes through them to sort the wheat from the chaff. There is always far more chaff than wheat. The computational challenge is to cut out as much chaff as possible without losing the wheat – that is, the new words which the lexicography team have not yet logged but which should be included in the dictionary.

For many aspects of corpus processing, we can use statistics to distinguish signal from noise, on the basis that the phenomena we are interested in are common ones and occur repeatedly. But new words are usually rare, and by definition are not already known. Thus lemmatisation is particularly challenging since the lemmatiser cannot make use of a list of known words. So for example, in one list we found the 'word' *authore,* an incorrect but understandable lemmatisation of *authored,* past participle of the unknown verb *author*.

For new-word finding we will want to include items in a candidate list even though they occur just once or twice. Statistical filtering can therefore only be used minimally. We are exploring methods which require that a word that occurred a maximum of once or twice in the old material occurs in at least three or four documents in the new material, to make its way onto the candidate list. We use some statistical modulation to capture new words which are taking off in the new period, as well as the items that simply have occurred where they never did before. Many items that occur in the new words list are simply typing errors. This is another reason why it is desirable to set a threshold higher than one in the new corpus.

For English, we have found that almost all hyphenated words are chaff, and often relate to compounds which are already treated in the dictionary as 'solid' or as multiword items. English hyphenation rules are not fixed: most word pairs that we find hyphenated (*sand-box*) can also be found written as one word (*sandbox*), and as two (*sand box*). With this in mind, to minimise chaff, we take all hyphenated forms and two- and three-word items in the dictionary and 'squeeze' them so that the one-word version is included in the list of already-known items, and we subsequently ignore all the hyphenated forms in the corpus list.

Prefixes and suffixes present a further set of items. Derivational affixes include both the more syntactic (*-ly, -ness*) and the more semantic (*-ish, geo-, eco-*).[6] Most are chaff: we do not want *plumply* or *ecobuddy* or *gangsterish* in the dictionary, because, even though they all have google counts in the thousands, they are not lexicalised and there is nothing to say about them beyond what there is to say about the lemma, the affix and the affixation rule. The ratio of wheat to chaff is low, but amongst the nonce productions there are some which are becoming established and should be considered for the dictionary. So we prefer to leave the nonce formations in place for the lexicographer to run their eye over.

For the longer term, the biggest challenge is acquiring corpora for the two time periods which are sufficiently large and sufficiently well-matched. If the new corpus is not big enough, the new words will simply be missed, while if the reference corpus is not big

enough, the lists will be full of false positives. If the corpora are not well-matched but, for example, the new corpus contains a document on vulcanology and the reference corpus does not, the list will contain words which are specialist vocabulary rather than new, like *resistivity* and *tephrochronology*.

While vast quantities of data are available on the web, most of it does not come with reliable information on when the document was originally written. While we can say with confidence that a corpus collected from the web in 2009 represents, overall, a more recent phase of the language than one collected in 2008, when we move to words with small numbers of occurrences, we cannot trust that words from the 2009 corpus are from more recently-written documents than ones from the 2008 corpus. Two text types where date-of-writing is usually available are newspapers and blogs. Both of these have the added advantage that they tend to be about current topics and are relatively likely to use new vocabulary. My current strategy for new-word-detection involves large-scale gathering of newspaper and blog feeds every day.

3 Collocation and word sketches

The arrival of large corpora provided the empirical underpinning for a view of language associated with Firth and Sinclair in which the patterning of words in text was central: *collocation* came to the fore.

Since the beginning of corpus lexicography, the primary means of analysis has been the reading of concordances. Since the earliest days of the COBUILD project, the lexicographers scanned concordance lines – often in their thousands – to find all the collocations and all the patterns of meaning and use. The more lines were scanned, the more patterns and collocations were found (though with diminishing returns). This was good and objective, but also difficult and time-consuming. Dictionary publishers were always looking to save time, and hence cut costs.

Early efforts to offer computational support were based on finding frequently co-occurring words in a window surrounding the headword (Church & Hanks 1990). While these approaches had generated plenty of interest among university researchers, they were not taken up as routine processes by lexicographers: the ratio of noise to signal was high, the first impression of a collocation list was of a basket of earth with occasional glints of possible gems needing further exploration, and it took too long to use them for every word.

The 'word sketch' is a response to this problem. A word sketch is a one-page, corpus-based summary of a word's grammatical and collocational behaviour, as illustrated in Figure 1. It uses a parser to identify all verb-object pairs, subject-verb pairs, modifier-modifiee pairs and so on, and then applies statistical filtering to give a fairly clean list, as proposed by Tapanainen & Järvinen (1998, and for the statistics, Rychly 2008). Word sketches need very large, part-of-speech-tagged corpora: in the late 1990s this had recently become available for general English in the form of the British National Corpus, and the first edition of word sketches were prepared to support a new, 'from scratch'

dictionary for advanced learners of English, the Macmillan English Dictionary for Advanced Learners (MEDAL, Rundell 2001).

Figure 1: Word sketch for *baby* (from enTenTen12, a very large 2012 web corpus)

As the lexicographers became familiar with the software, it became apparent that word sketches did the job they were designed to do. Each headword's collocations could be listed exhaustively, to a far greater degree than was possible before. That was the immediate goal. But analysis of a word's sketch also tended to show, through its collocations, a wide range of the patterns of meaning and usage that it entered into. In most cases, each of a word's different meanings is associated with particular collocations, so the collocates listed in the word sketches provided valuable prompts in the key task of identifying and accounting for all the word's meanings in the entry. The word sketches functioned not only as a tool for finding collocations, but also as a useful guide to the distinct senses of a word – the analytical core of the lexicographer's job (Kilgarriff & Rundell 2002).

It became clear that the word sketches were more like a contents page than a basket of earth. They provided a neat summary of most of what the lexicographer was likely to find by the traditional means of scanning concordances. There was not too much noise. Using them saved time. It was more efficient to start from the word sketch than from the concordance.

Thus the unexpected consequence was that the lexicographer's methodology changed, from one where the technology merely supported the corpus-analysis process, to one where it pro-actively identified what was likely to be interesting and directed the lexicographer's attention to it. And whereas, for a human, the bigger the corpus, the greater the problem of how to manage the data, for the computer, the bigger the corpus, the better the analyses: the more data there is, the better the prospects for finding all salient patterns and for distinguishing signal from noise. Though originally seen as a useful supplementary tool, the sketches provided a compact and revealing snapshot of a word's behaviour and uses and became the preferred starting point in the process of analyzing complex headwords.

Since the first word sketches were used in the late 1990s, the Sketch Engine, the corpus query tool within which they are presented, has not stood still. Word sketches have been developed for a dozen languages (the list is steadily growing) and have been complemented by an automatic thesaurus (which identifies the words which are most similar, in terms of shared collocations, to a target word, see Fig. 2) and a range of other tools including 'sketch diff', for comparing and contrasting a word with synonyms or antonyms (see Fig. 3). There are also options such as clustering a word's collocates or its thesaurus entries. The largest corpus for which word sketches have been created so far contains seventy billion words (Pomikalek et al. 2012). In a quantitative evaluation, two thirds of the collocates in word sketches for four languages were found to be 'publishable quality': a lexicographer would want to include them in a published collocations dictionary for the language (Kilgarriff et al. 2010).

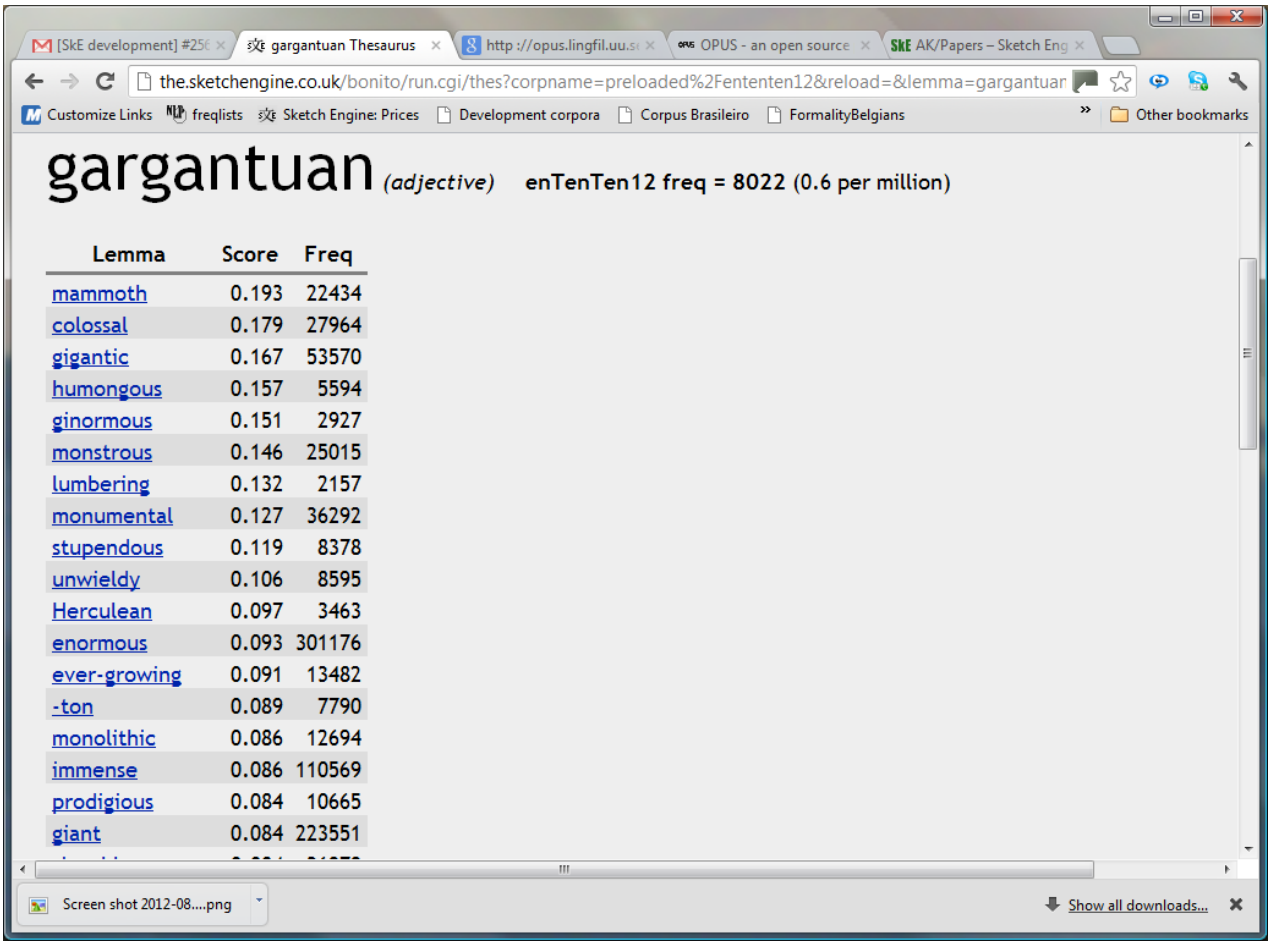

Figure 2: Thesaurus entry for *gargantuan*

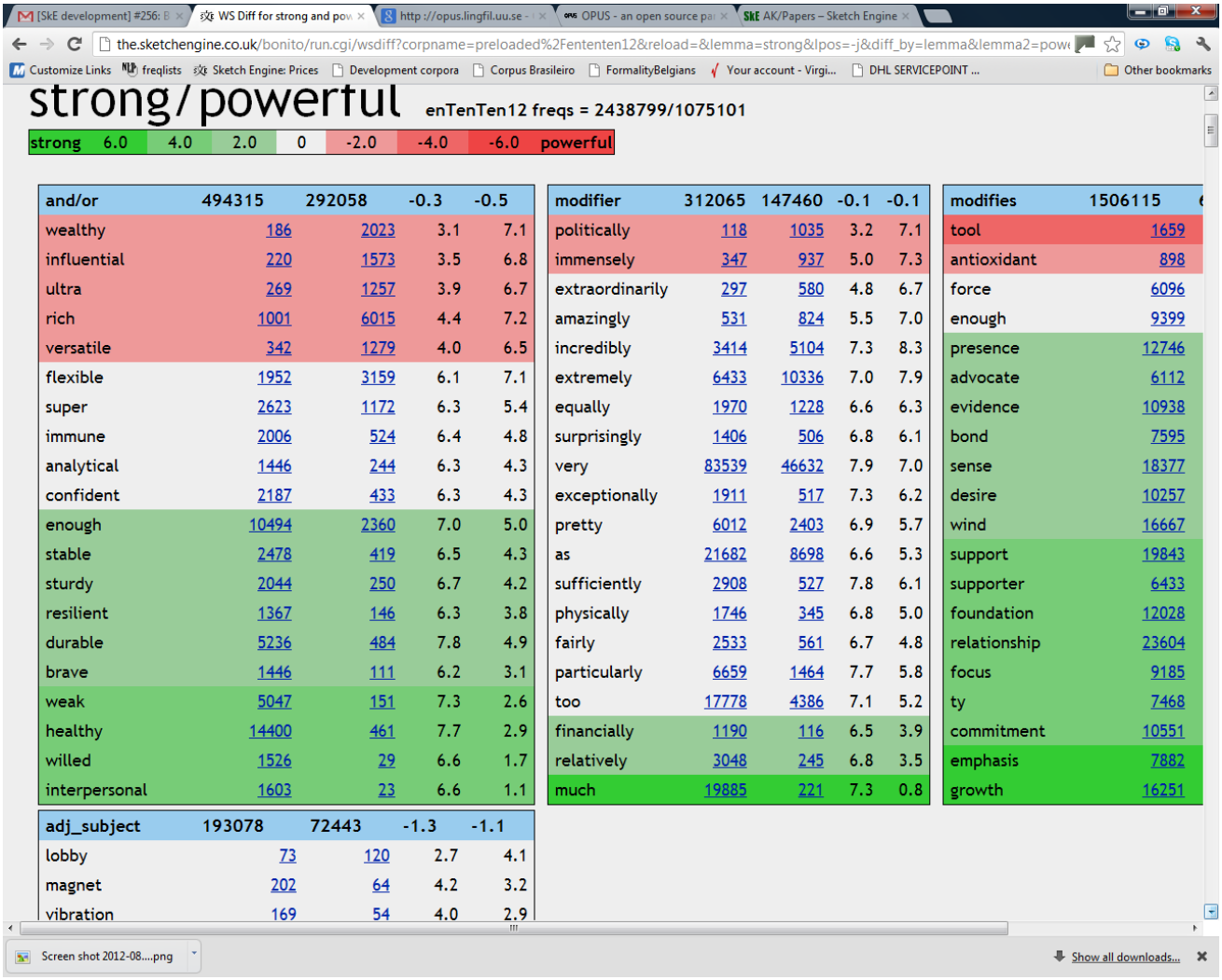

Figure 3: Sketch diff comparing *strong* and *powerful*

4 Labels

Dictionaries use a range of labels (such as *usu pl.*, *informal*, *Biology*, *AmE*) to mark words according to their grammatical, register, domain, and regional characteristics, whenever these deviate significantly from the (unmarked) norm. All of these are facts about a word's distribution, and all can, in principle, be gathered automatically from a corpus. In each of these four cases, computationalists are currently able to propose some labels to the lexicographer, though there remains much work to be done.

In each case the methodology is to:

- specify a set of hypotheses
    - there will usually be one hypothesis per label, so grammatical hypotheses for the category 'verb' may include:
        - is it often/usually/always passive
        - is it often/usually/always progressive

        - is it often/usually/always in the imperative
- for each word
    - test all relevant hypotheses
    - for all hypotheses that are confirmed, alert the lexicographer
        - (in the Sketch Engine, by adding the information to the word sketch).

Where no hypotheses are confirmed – when, in other words, there is nothing interesting to say, which will be the usual case – no alerts are given.

4.1 Grammatical labels*: usually plural, often passive*, etc.

To determine whether a noun should be marked as 'usually plural', we simply count the number of times the lemma occurs in the plural, and the number of times it occurs overall, and divide the second number by the first to find the proportion. Similarly, to discover how often a verb is passivized, we can count how often it is a past participle preceded by a form of the verb *be* (with possible intervening adverbs) and determine what fraction of the verb's overall frequency the passive forms represent. Given a lemmatised, part-of-speech-tagged corpus, this is straightforward. A large number of grammatical hypotheses can be handled in this way.

The next question is: when is the information interesting enough to merit a label in a dictionary? Should we, for example, label all verbs which are over 50% passive as *often passive*?

To assess this question, we want to know what the implications would be: we do not want to bombard the dictionary user with too many labels (or the lexicographer with too many candidate-labels). What percentage of English verbs occur in the passive over half of the time? Is it 20%, or 50%, or 80%? This question is also not in principle hard to answer: for each verb, we work out its percentage passive, and sort according to the percentage. We can then give a figure which is, for lexicographic purposes, probably more informative than 'the percentage passive': the percentile. The percentile indicates whether a verb is in the top 1%, or 2%, or 5%, or 10% of verbs from the point of view of how passive they are. We can prepare lists as in Table 2. This uses the methodology for finding the 'most passive' verbs (with frequency over 500) in the BNC. It shows that the most passive verb is *station*: people and things are often *stationed* in places, but there are far fewer cases where someone actively *stations* things. For *station*, 72.2% of its 557 occurrences are in the passive, and this puts it in the 0.2% 'most passive' verbs of English. At the other end of the table, *levy* is in the passive just over half the time, which puts it in the 1.9% most passive verbs. The approach is similar to the collostructional analysis of Gries & Stefanowitsch (2004).

Table 2: *The 'most passive' verbs in the BNC, for which a 'usually passive' label might be proposed.*

Percentile	Ratio	Lemma	Frequency
0.2	72.2	station	557
0.2	71.8	base	19201
0.3	71.1	destine	771
0.3	68.7	doom	520
0.4	66.3	poise	640
0.4	65.0	situate	2025
0.5	64.7	schedule	1602
0.5	64.1	associate	8094
0.6	63.2	embed	688
0.7	62.0	entitle	2669
0.8	59.8	couple	1421
0.9	58.1	jail	960
1.1	57.8	deem	1626
1.1	55.5	confine	2663
1.2	55.4	arm	1195
1.2	54.9	design	11662
1.3	53.9	convict	1298
1.5	53.1	clothe	749
1.5	52.8	dedicate	1291
1.5	52.4	compose	2391
1.6	51.5	flank	551
1.7	50.8	gear	733
1.9	50.1	levy	603

As can be seen from this sample, the information is lexicographically valid: all the verbs in the table would benefit from an *often passive* or *usually passive* label.

A table like this can be used by editorial policy-makers to determine a cut-off which is appropriate for a given project. For instance, what proportion of verbs should attract an *often passive* label? Perhaps the decision will be that users benefit most if the label is not overused, so just 4% of verbs would be thus labelled. The full version of the table in Figure 4 tells us what these verbs are. And now that we know precisely the hypothesis to use ("is the verb in the top 4% most-passive verbs?") and where the hypothesis is true, the label can be added into the word sketch. In this way, the element of chance – will the lexicographer notice whether a particular verb is typically passivized? – is eliminated, and the automation contributes to a consistent account of word behaviour.

4.2 Register Labels: formal, informal, etc.

Any corpus is a collection of texts. Register is in the first instance a classification that applies to texts rather than words. A word is informal (or formal) if it shows a clear tendency to occur in informal (or formal) texts. To label words according to register, we need a corpus in which the constituent texts are themselves labelled for register in the document header. Note that at this stage, we are not considering aspects of register other than formality.

One way to come by such a corpus is to gather texts from sources known to be formal or informal. In a corpus such as the BNC, each document is supplied with various text type classifications, so we can, for example, infer from the fact that a document is everyday conversation, that it is informal, or from the fact that it is an academic journal article, that it is formal.

The approach has potential, but also drawbacks. In particular, it is not possible to apply it to any corpus which does not come with text-type information. Web corpora do not. An alternative is to build a classifier which infers formality level on the basis of the vocabulary and other features of the text. There are classifiers available for this task: see for example Heylighen & Dewaele (1999), and Santini et al. (2009). Following this route, we have recently labelled all documents in a twelve billion word web corpus according to formality, so we are now in a position to order words from most to least formal. The next tasks will be to assess the accuracy of the classification, and to consider – just as was done for passives – the percentage of the lexicon we want to label for register.

The reasoning may seem circular: we use formal (or informal) vocabulary to find formal (or informal) vocabulary. But it is a spiral rather than a circle: each cycle has more information at its disposal than the previous one. We use our knowledge of the words that are formal or informal to identify documents that are formal or informal. That then gives us a richer dataset for identifying further words, phrases and constructions which tend to be formal or informal, and allows us to quantify the tendencies.

4.3 Domain Labels: Geol., Astron., etc

The issues are, in principle, the same as for register. The practical difference is that there are far more domains (and domain labels): even MEDAL, a general-purpose learner's dictionary, has eighteen of these; larger dictionaries typically have over one hundred. Collecting large corpora for each of these domains is a significant challenge.

It is tempting to gather a large quantity of, for example, geological texts from a particular source, perhaps an online geology journal. But rather than being a 'general geology' corpus, that subcorpus will be an 'academic-geology corpus', and the words which are particularly common in the subcorpus will include vocabulary typical of academic discourse in general, and vocabulary associated with the preferences and specialisms of that particular journal, as well as of the domain of geology. Ideally, each subcorpus will have the same proportions of different text-types as the whole corpus. None of this is technically or practically impossible, but the larger the number of subcorpora, the harder it is to achieve.

Once we have the corpora and counts for each word in each subcorpus, we need to use statistical measures for deciding which words are most distinctive of the subcorpus: which words are its 'keywords', the words for which there is the strongest case for labelling. The maths we use is based on a simple ratio between relative frequencies, as implemented in the Sketch Engine and presented in Kilgarriff (2009).

4.4 Region Labels: AmE, AustrE, etc

The issues concerning region labels are the same as for domains but in some ways a little simpler. The taxonomy of regions, at least from the point of view of labelling items used in different parts of the English-speaking world, is relatively limited, and a good deal less open-ended than the taxonomy of domains. In MEDAL, for example, it comprises just twelve varieties or dialects: American, Australian, British, Canadian, Carribean, Irish, New Zealand, and South/East/West African English.

5 Examples

Most dictionaries include example sentences. They are especially important in pedagogical dictionaries, where a carefully-selected set of examples can clarify meaning, illustrate a word's contextual and combinatorial behaviour, and serve as models for language production. The benefits for users are clear, and the shift from paper to electronic media means that we can now offer users far more examples. But this comes at a cost. Finding good examples in a mass of corpus data is labour-intensive. For all sorts of reasons, a majority of corpus sentences will not be suitable as they stand, so the lexicographer must either search out the best ones or modify corpus sentences which are promising but in some way flawed.

5.1 GDEX

In 2007, the requirement arose – in a project for Macmillan – for the addition of new examples for around 8,000 collocations. The options were to ask lexicographers to select and edit these in the 'traditional' way, or to see whether the example-finding process could be automated. Budgetary considerations favoured the latter approach, and subsequent discussions led to the GDEX ('good dictionary examples') algorithm, which is described in Kilgarriff et al. (2008).

The method is to score sentences, and only display the highest-scoring ones. A wide range of heuristics are used for scoring, including sentence length, the presence (or absence) of rare words or proper names, and the number of pronouns in the sentence. The system worked successfully on its first outing – not in the sense that every example it identified was immediately usable, but in the sense that it streamlined the lexicographer's task.

GDEX continues to be refined, as more selection criteria are added and the weightings of the different filters adjusted, for English and for other languages. The lexicographer can scan a short list until they find a suitable example for whatever feature is being illustrated, and GDEX means they are likely to find what they are looking for in the top five examples, rather than, on average, within the top twenty to thirty.

6 Translations

The corpora that help most for finding translations are parallel corpora : corpora comprising pairs of texts that are translations of each other. Parallel corpora are the fuel that Google Translate feeds on, and 'statistical machine translation', of which Google Translate is the highest-profile example, is a great success story of language technology and the use of corpora.

Parallel corpora are of most use if they are aligned: that is, for each sentence, or word, in the one text, the computer knows what the corresponding item is in the other. Where the text is a straightforward literal translation, sentence alignment can now be performed with high accuracy. Of course, some sentences are not one-to-one, and some sections may exist in only one language. Working solutions have been found for identifying and handling these cases, which are all on a cline of how closely the translation follows the original. Throughout parallel corpus work, text-pairs which are literal translations are easiest to work with, whereas free translations of novels offer much less.

Word alignment is intrinsically a trickier concept than sentence alignment. Firstly, very often, an individual word is not translated by a single-word. Secondly, items often do not stay in the same order. One can expect the sentences and their translations to be in the same order as each other, but one cannot expect the words and their translations to be in the same order in source and target text.

There are two ways for lexicographers to use parallel corpora: parallel concordances, and summaries. The first is simpler and is based only on sentence alignments. The lexicographer searches for a word or phrase on one side of the corpus, and sees pairs of sentences which are translations of each other. The website http://www.linguee.com offers exactly this, for the big European languages, and since its arrival in 2009.

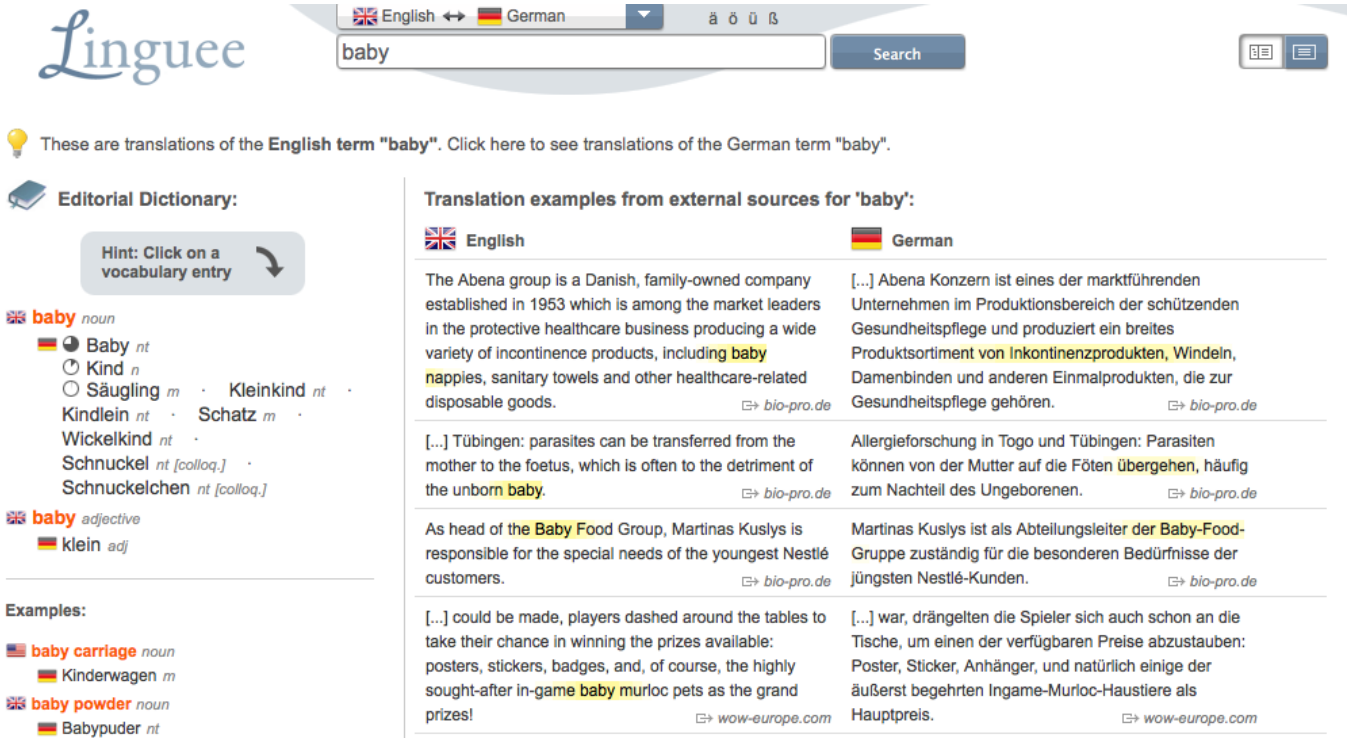

Figure 4: Screenshot from Linguee.com for English search term *baby*, language pair English-German.

It has rapidly become a translator's and favourite. Its display, for the English search term *baby* and the language pair English-German, is shown in Figure 4.

As with examples in general, people (translators, lexicographers and other users) find these example pairs very useful and easy to use. They will often remind a lexicographer of ways of translating a word or phrase that should be included in the dictionary entry, and will supply example sentence pairs to be included (usually after some editing).

The 'summaries' approach for using parallel corpora only applies when the corpora are large, and for words where there are many sentence pairs. Then, it will not be possible for the lexicographer to read all the sentence pairs, and it should be possible for the computer to summarise what it finds in them. This is a bilingual version of the reasoning that led to word sketches. In a process closely related to methods for word alignment, the computer can find the other-language words that occur with particularly high frequency in the node word's aligned sentences. The process can also be applied to find candidate collocations as translations of the node word's collocations.

A first version of a bilingual word sketch based on a parallel corpus, for *red* for the language pair English-French, is shown in Fig. 5.

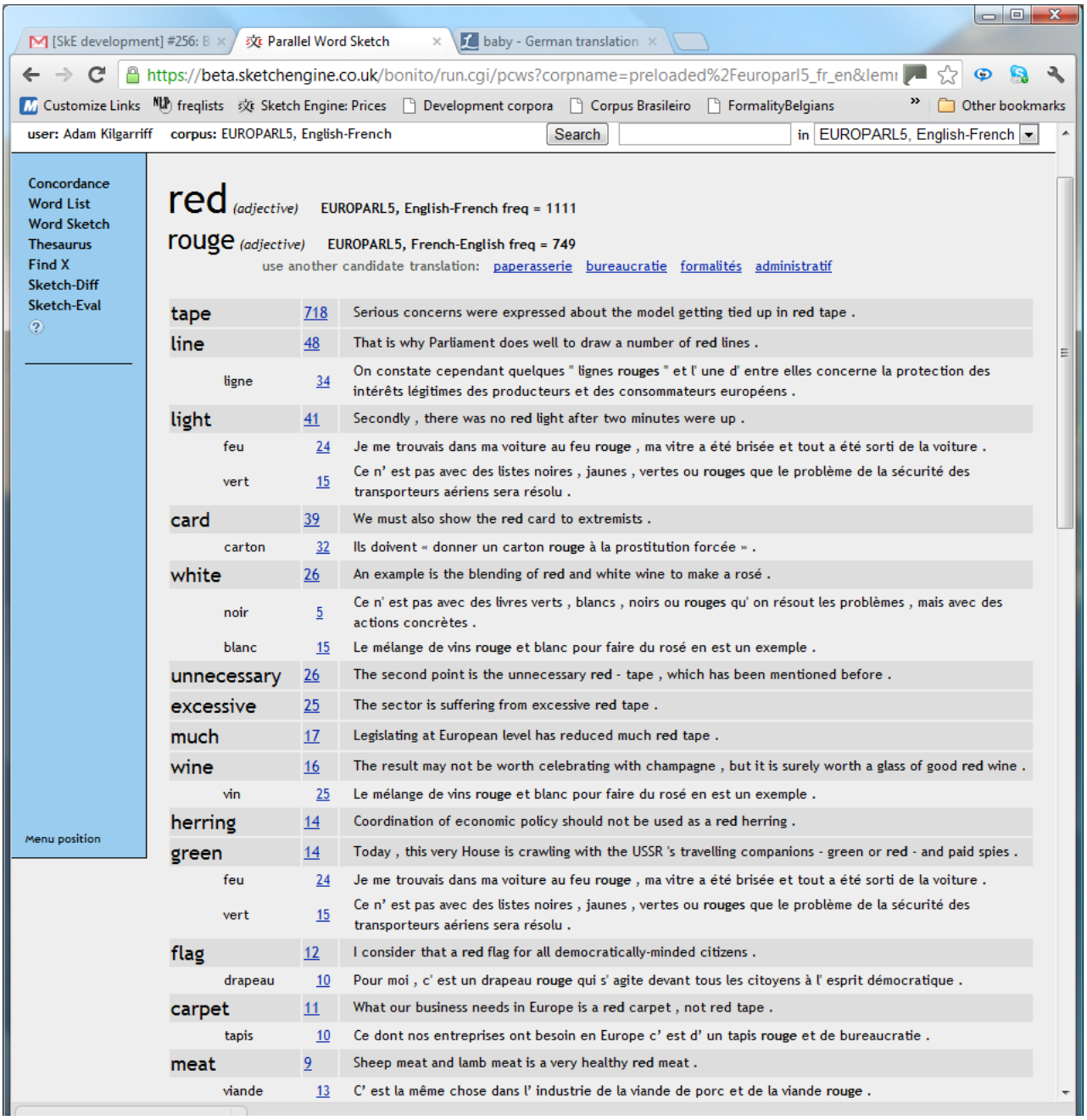

Figure 5: Bilingual word sketch based on a parallel corpus, for *red* for the language pair English-French.

A limiting factor for parallel-corpus work is the availability of a parallel corpus, for the language pair in question. The early work in the field was based on the Canadian parliamentary proceedings, 'Canadian Hansard', which were available in English and French and were fairly literal professional translations of each other. Other sources frequently used, for the languages of the EU, are the European parliamentary proceedings and other documents from the EU (as used for the screenshot). Other text types where parallel data is often available include software documentation, documentation for

vehicles and machinery, and film transcripts. A large and well-maintained collection of parallel data is available at the OPUS website.[17] For any particular language pair, some text types will be available, others will not.

7 Summary

Corpora can make dictionary-making more accurate, efficient, complete and consistent. They can deliver a candidate headword list, and, where the corpora are developed with care with neologism-finding in mind, can identify candidate neologisms.

There are many ways in which they can support entry-writing. They can provide a wide range of clues to the lexicographer for analysing the word's range of meaning into distinct senses. In combination with a suitable corpus query system they can find the idioms, phrases and collocations for a word. They can identify if a word has noteworthy behaviour in relation to grammar, domain, region and register. They can do the preparatory work for finding good example sentences, and translations.

Corpora have been used in these ways in a range of dictionary projects, and the chapter has described how this has worked, with reference to a particular corpus query tool, the Sketch Engine. Over the last two decades, the lexicographer's role has been more and more often, checking and confirming or editing the corpus tool's work, where earlier it would have been 'writing from scratch'.

In the early twenty-first century, with the advent of the web and many and varied online resources, much is changing in the world of dictionary-making, and many things are uncertain. Quite what the role of the lexicographer will be, in ten years' time, is far from clear, but I am confident that the role of the corpus will grow, with the line between dictionary and corpus blurring, and the lexicographer operating at that interface.

References

Atkins, S. & Rundell, M. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation Journal* 43(3): 209-226.

Church, K. & Hanks, P. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics* 16:22–29.

Clear, J. 1988. The Monitor Corpus. In *ZüriLEX '86 Proceedings*, M. Snell-Hornby (ed.), 383-389. Tübingen: Francke Verlag.

Fairon, C., Macé, K., & Naets, H. 2008. GlossaNet2: a linguistic search engine for RSS-based corpora. *Proceedings, Web As Corpus Workshop (WAC4),* S. Evert, A. Kilgarriff & S. Sharoff (eds), 34-39. Marrakech.

Gries, S. Th. & Stefanowitsch, A. 2004. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9(1): 97-129.

Heylighen F. & Dewaele, J.-M. 1999. Formality of Language: Definition, measurement and behavioural determinants. Internal Report, Free University Brussels, http://pespmc1.vub.ac.be/Papers/Formality.pdf

Janicivic, T. & Walker, D. 1997. NeoloSearch: Automatic Detection of Neologisms in French Internet Documents. *Proceedings of ACH/ALLC'97*: 93-94. Queen's University, Ontario, Canada.

Kilgarriff, A. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography* 10(2): 135-155.

Kilgarriff, A. 2009. Simple maths for keywords. *Proceedings, Corpus Linguistics.* M. Mahlberg, V. González-Díaz & C. Smith (eds). Liverpool; online at http://ucrel.lancs.ac.uk/publications/cl2009/.

Kilgarriff, A. & Kosem, I. 2012. Corpus Tools for Lexicographers. In *Electronic Lexicography.* S. Granger & M. Paquot (eds). Oxford University Press.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In *Proceedings of the XIII Euralex Congress*, E. Bernal & J. DeCesaris (eds), 425-431. Barcelona: Universitat Pompeu Fabra.

Kilgarriff, A., Kovář, V. Krek, S. Srdanović, I., & Tiberius, C. 2010. A quantitative evaluation of word sketches. *Proceedings of 14th EURALEX International Congress*, A. Dykstra & T. Schoonheim (eds). Leeuwarden, The Netherlands.

Kilgarriff, A. & Rundell, M. 2002. Lexical Profiling Software and its Lexicographic Applications: A Case Study. In *Proceedings of the Tenth Euralex Congress*, A. Braasch & C. Povlsen (eds), 807-818. Copenhagen: University of Copenhagen.

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. 2004. The Sketch Engine In *Proceedings of the Eleventh Euralex Congress*, G. Williams & S. Vessier (eds), 105-116. Lorient, France: UBS.

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Bondi Johannessen, J., Khalil, S., Johansson Kokkinakis, S., Robert Lew, R., Sharoff, S., Vadlapudi, R. & Volodina, E. 2012. Corpus-Based Vocabulary lists for Language Learners for Nine Languages. *Language Resources and Evaluation Journal.* To appear.

O'Donovan, R. & O'Neill, M. 2008. A Systematic Approach to the Selection of Neologisms for Inclusion in a Large Monolingual Dictionary. In *Proceedings of the XIII Euralex Congress*, E. Bernal & J. DeCesaris (eds), 571-579. Barcelona: Universitat Pompeu Fabra.

Pomikálek, J., Jakubíček, M. & and Rychlý, P. 2012. Building a 70 billion word corpus of English from ClueWeb. *Proc LREC.* Istanbul.

Rundell, M. (ed.). 2001. *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan Education.

Rychlý, P. 2008. A Lexicographer-Friendly Association Score. *Proc. 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing*. Sojka P., Horák A. (Eds). Brno : Masaryk University.

Santini M., Rehm, G., Sharoff, S., & Mehler, A. (eds). 2009. Introduction, *Journal for Language Technology and Computational Linguistics*, Special Issue on Automatic Genre Identification: Issues and Prospects. 24(1):129-145.

Tapanainen, P. & Järvinen, T. 1998. Dependency Concordances. International *Journal of Lexicography* 11(3):187-203.

Notes

[1] The website for the BNC is http://natcorp.ox.ac.uk
[2] In the BNC *mucosa* is marginally more frequent than *spontaneous* and *enjoyment*, but appears in far fewer corpus documents.
[3] As is now generally recognised, the notion of 'representativeness' is problematical with regard to general-purpose corpora like BNC and UKWaC, and there is no 'scientific' way of achieving it: see e.g. Atkins & Rundell (2008: 66).
[4] The database can be explored online at http://kelly.sketchengine.co.uk
[5] The issue came to our attention when an early version of the BNC frequency list gave undue prominence to verbal *car*.
[6] Here we exclude inflectional morphemes, addressed under lemmatisation above: in English a distinction between inflectional and derivational morphology is easily made for most cases.
[7] http://opus.lingfil.uu.se/