Getting to know your corpus

Adam Kilgarriff

Lexical Computing Ltd., Brighton, UK
http://www.sketchengine.co.uk

Abstract. Corpora are not easy to get a handle on. The usual way of getting to grips with text is to read it, but corpora are mostly too big to read (and not designed to be read). We show, with examples, how keyword lists (of one corpus *vs.* another) are a direct, practical and fascinating way to explore the characteristics of corpora, and of text types. Our method is to classify the top one hundred keywords of corpus1 *vs.* corpus2, and corpus2 *vs.* corpus1. This promptly reveals a range of contrasts between all the pairs of corpora we apply it to. We also present improved maths for keywords, and briefly discuss quantitative comparisons between corpora. All the methods discussed (and almost all of the corpora) are available in the Sketch Engine, a leading corpus query tool.

Keywords: corpora, corpus similarity, keywords, keyword lists, Sketch Engine

1 Spot the difference

> accord actually amendment among bad because behavior believe bill blog ca center citizen color defense determine do dollar earth effort election even evil fact faculty favor favorite federal foreign forth guess guy he her him himself his honor human kid kill kind know labor law let liberal like man maybe me military movie my nation never nor not nothing official oh organization percent political post president pretty professor program realize recognize say shall she sin soul speak state suppose tell terrorist that thing think thou thy toward true truth unto upon violation vote voter war what while why woman yes

> accommodation achieve advice aim area assessment available band behaviour building centre charity click client club colour consultation contact council delivery detail develop development disabled email enable enquiry ensure event excellent facility favourite full further garden guidance guide holiday improve information insurance join link local main manage management match mm nd offer opportunity organisation organise page partnership please pm poker pp programme project pub pupil quality range rd realise recognise road route scheme sector service shop site skill specialist st staff stage suitable telephone th top tour training transport uk undertake venue village visit visitor website welcome whilst wide workshop www

These two lists are the keywords we see when we compare one web-crawled English corpus (UKWaC, [1]) with another (enTenTen, [2]).

It does not take long to spot recurring themes: one classification (with each word assigned to one and only one category) is shown in Table 1.

enTenTen keywords	***UKWaC keywords***
American spellings: among behavior center color defense favor favorite honor labor organization program realize recognize toward while **American politics:** amendment bill blog citizen election federal law liberal nation official president state vote voter **Bible:** believe evil forth nor sin soul speak thou thy true truth unto upon **Informal:** guy kid oh pretty yes **Core verbs:** be determine do guess know let say shall suppose tell think **War and terrorists:** foreign military kill terrorist violation war **Pronouns:** he her him his me my she **Negatives:** never nor not nothing **Other adverbs:** even actually even maybe like pretty **Other grammatical words:** because that what why **Academic:** faculty professor **Core nouns:** kind thing fact effort man woman human **Other:** accord bad ca dollar earth movie percent post	**British spelling:** behaviour centre colour favourite organise organisation programme realise recognise whilst **Schools, training:** assessment council guidance local pupil scheme skill training workshop **Business:** achieve advice aim building client consultation develop development facility improve information manage management offer opportunity partnership project quality sector service specialist staff undertake **Furniture of web pages:** available contact click enquiry detail email further guide join link page please site telephone visit visitor website welcome www **British lexical variants:** garden holiday shop transport (American equivalents: yard vacation store transportation) **British culture:** pub village **Music:** band event stage tour venue **Addresses:** rd road route st **Nonwords:** th pm uk nd mm pp **Adjectives:** main suitable excellent full wide **Other:** accommodation area charity club disabled enable ensure insurance main match poker range top

Table 1. Keywords; enTenTen *vs.* UKWaC

So: enTenTen has more American, more politics, more informal material, more war and terrorism, plus seams of biblical and academic material. UKWaC has a corresponding set of Britishisms and more on schools and training, business and music.

At all times we should note that the terms on one list were not missing from the other – American spellings are found in large numbers in UKWaC too – just that the balance is different. Needless to say many words might belong in multiple categories (shouldn't *believe* go with verbs?) and the classification requires some guessing about dominant meanings and word classes of polysemous words: I think it will be the adverbial *pretty* ("a pretty good idea") , not the adjectival ("a pretty dress"). There is nothing intrinsically business-y about words like *development, project, opportunity* but it is my hunch, on seeing them all in the same list, that their prevalence is due to their appearance in texts where companies are giving an account of all the good work they do.

We can attempt to check which words belong where by looking at concordances, but this turns out to be hard. Typically many patterns of use will be common in both corpora and, other than the plain statistic, there is no obvious way to summarise contrasts. For cases like *pretty*, where different meanings are for different word classes, we can see how the word classes differ (and the evidence from automatic pos-tagging confirms my guess: the adverb is an order of magnitude more frequent than the adjective (in both corpora) and dominates the frequencies). However that only helps in a limited range of cases (and part-of-speech tagging makes many mistakes with ambiguous words).

The lists are sorted by ratio of normalised frequencies (after addition of the simplemaths parameter, see below) and the lists are then simply the top 100 items, with no manual editing. Other settings were:

- Lemmas (as opposed to word forms).
  - This does some generalising over, for example, singular and plural form of the same noun.
  - It is only possible when the same lemmatisation procedure has been applied to both corpora. Otherwise, even if the differences seem minimal, the top of the keyword list will be dominated by the cases that were handled differently.
- Simplemaths parameter: 100 (see below)
- Only items containing exclusively lowercase a-to-z characters were included
- A minimum of two characters

Varying the setting gives other perspectives. Looking at capitalised items, reducing simplemaths parameter to ten and setting minimum length to five, we find the top items in enTenTen are *Obama Clinton Hillary McCain*, and for UKWaC *Centre Leeds Manchester Edinburgh*. (I changed the parameters to get longer and potentially lower-frequency items. Otherwise the lists had many acronyms and abbreviations: I wanted to see names.) enTenTen was collected in the run-up to the US Presidential election, which also explains the ‘political’ cluster. In UKWaC we have many places. *rd* and *st* are abbreviations, as usually used, in addresses, for ‘road’ and ‘street’.

The settings I most often use are: simplemaths 100, exclusively lowercase words, of at least three characters. This usually gives a set of core-vocabulary words with minimal noise. Shorter items (one and two characters) are often not words For the lists above I used two characters, thereby including the two-letter words *be do he me my* and non-words *oh ca rd st th nd pm uk mm pp*. While my usual preference is for words, these all tell their story too, with, for example, *oh* vouching for informality and *mm* for the preference for the metric system in the UK (the USA more often uses inches).

If there are differences in how the data was prepared, they tend to dominate keyword lists. Between these two corpora there were not many differences - but the technology available for ‘cleaning up’ and removing duplicated material from web datasets has improved (thanks particularly to the work of Jan Pomikalek, whose tools were used for enTenTen). The ‘furniture of web pages’ cluster in UKWaC is probably there because, in 2009, we removed repeated material from web pages more effectively than in 2006.

1.1 Formality

Whereas lower-frequency items will support an understanding of the differences of content between the two corpora, as linguists we are also interested in differences of register. As Biber shows, the dominant dimension according to which text varies, across a wide range of text types and also languages, is from formal to informal, or to use his more specific terms, from interactional (for which everyday conversation is the prototype) to informational (with an academic paper as an extreme case) [3, 4]. There are many features of text that vary according to where it sits on this dimension. Ones that are easily counted include word class: interactional language uses more verbs, personal pronouns and adverbs, informational uses more nouns, articles and adjectives [5].

We can see that there is a higher proportion of less formal material in enTenTen from the categories pronouns, core verbs, adverbs as well as the one marked informal, and the adjectives category in UKWaC is perhaps an indicator of higher formality. Both corpora have been tagged by TreeTagger[1] and we can investigate further by looking at a keyword list, not of words or lemmas, but of word classes.

The word classes with a ratio between relative frequencies of 1.2 or greater are:

enTenTen key word classes	***UKWaC key word classes***
PP, PP$ personal pronoun (regular, possessive) **VVD, VVP** lexical verb (past, present tense) **IN/that** that as subordinator **WP, WDT** wh-pronoun, wh-determiner **UH** interjection **VHD, VH** the verb have, base form and past tense **RB** adverb	**POS** Possessive ending **NP** Proper noun

Table 2. Key word classes, enTenTen $vs.$ UKWaC

This confirms the greater formality (on average) of UKWaC.

2 Simple maths for keywords

The statistics used here for identifying keywords improve on those used elsewhere.

"This word is twice as common here as there." This is the simplest way to make a comparison between a word's frequency in one text type and its frequency in another. "Twice as common" means the word's frequency (per thousand or million words) in the first corpus is twice its frequency in the second. We count occurrences in each corpus, divide each number by the number of words in that corpus, optionally multiply by 1,000 or 1,000,000 to give frequencies per thousand or million, and divide the first number by the second to give a ratio. (Since the thousands or millions cancel out when we carry out the division, it makes no difference whether we use thousands or millions.)

If we find the ratio for all words, and sort by the ratio, we find the words that are most associated with each corpus as against the other. This will give a first pass at two

[1] See http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/; for the tagset used, see https://trac.sketchengine.co.uk/wiki/tagsets/penn

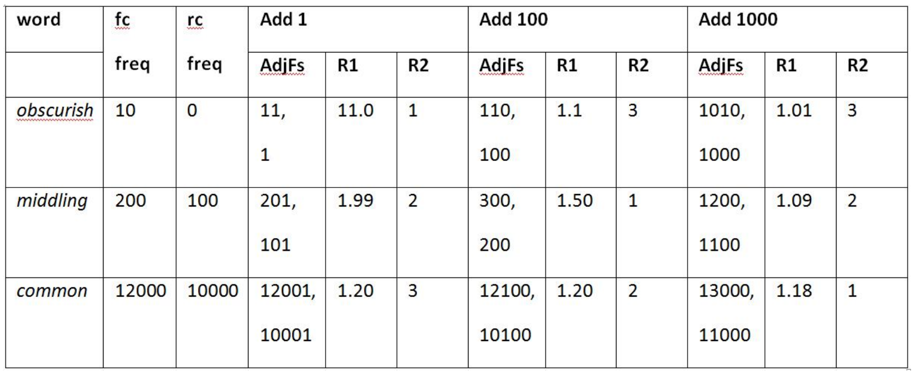
keyword lists, one (taken from the top of the sorted list) of corpus1 $vs.$ corpus2, and the other, taken from the bottom of the list (with scores below 1 and getting close to 0), for corpus2 $vs.$ corpus1. (In the discussion below we will refer to the two corpora as the focus corpus or fc, for which we want to find keywords, and the reference corpus or rc. We divide relative frequency in the fc by relative frequency in the rc and are interested in the high-scoring words.)

One problem with preparing keyword lists in this way is that you can't divide by zero, so it is not clear what to do about words which are present in the fc but absent in the rc.

A second problem is that, even setting aside the cases of zero occurrences, the list will be dominated by words with very few occurrences in the rc. There is nothing very surprising about a contrast between 10 in fc and 1 in rc, giving a ratio of 10, and we expect to find many such cases; but we would be very surprised to find words with 10,000 hits in fc and only 1,000 in rc, even though that also gives a ratio of 10. Simple ratios will give a list dominated by rare words.

A common solution to the zeros problem is 'add one'. If we add one to all the frequencies, including those for words which were present in fc but absent in rc, then we have no zeros and can compute a ratio for all words. A word with 10 hits in fc and none in rc gets a ratio of 11:1 (as we add 1 to 10 and to 0) or 11. "Add one" is widely used as a solution to a range of problems associated with low and zero frequency counts, in language technology and elsewhere [6].

This suggests a solution to the second problem. Consider what happens when we add 1, 100, or 1000 to all counts from both corpora. The results, for the three words *obscurish, middling,* and *common,* in two hypothetical corpora, are presented in Figure 1.



word	fc	rc	Add 1			Add 100			Add 1000		
	freq	freq	AdjFs	R1	R2	AdjFs	R1	R2	AdjFs	R1	R2
obscurish	10	0	11, 1	11.0	1	110, 100	1.1	3	1010, 1000	1.01	3
middling	200	100	201, 101	1.99	2	300, 200	1.50	1	1200, 1100	1.09	2
common	12000	10000	12001, 10001	1.20	3	12100, 10100	1.20	2	13000, 11000	1.18	1

Fig. 1. Frequencies, adjusted frequencies (AdjFs), ratio (R1), and keyword rank (R2), for three Simplemaths parameter settings, for rare, medium, and common words.

All three words are notably more common in fc than rc, so all are candidates for the keyword list, but they are in different frequency ranges.

- When we add 1, $obscurish$ comes highest on the keyword list, with $middling$ second, and $common$ last.
- When we add 100, the order is $middling, common, obscurish.$
- When we add 1000, it is $common, middling, obscurish.$

Different values for the 'add-N' or 'simplemaths' parameter give prominence to different frequency ranges. For some purposes a keyword list with commoner words is desirable; for others, we would want more rarer words. Our model lets the user specify the keyword list they want by adjusting the parameter. The model provides a way of identifying keywords without unwarranted mathematical sophistication, and reflects the fact that there is no one-size-fits-all list and different lists are wanted for different research questions.

The model is called 'simple maths' in contrast to other methods for keyword extraction, several of which use a hypothesis testing approach to see by what margin the null hypthesis is disproved. Such approaches both have much more complex maths, and are built on a flawed analysis of corpus statistics: the case is presented in full in [7].

3 Comparing Corpora of known, different genres

Our first test case –UKWaC *vs.* enTenTen– was one in which we did not, at the outset, know what the differences were between the two corpora. The same method can be used where we know the differences of text type, which are there by design, and then we can use the keyword lists to find out more about the distinctions between the two text types, and also to find other, possibly unintended, contrasts between the two corpora.

We compared BAWE (British Academic Written English [8]) with SiBol/Port (comprising British broadsheet national newspapers [9]) and classified the top hundred words (word forms, with simplemaths 100, at least three letters, all lowercase) as follows.

A side-effect of using word forms rather than lemmas is that we see, in many cases, multiple forms of the same lemma (*factor factors, theory theories, use used using, played player players playing* etc.) While in one way this means we have had to waste time on multiple copies of the same word, in another it is reassuring: it shows how systematic the process is, where, of all the tens of thousands of English words that could have appeared in these top-100 lists, the words that do are so often different forms of the same lemma.

Much could be said about the analyses above, and what they tell us about academic writing, journalism, and the contrasts between them. A few brief comments:

1. Academic writing is more formal. The BAWE list is mostly nouns, with some adjectives and prepositions. The verbs that do appear are mostly past or past participle forms, with some (*associated, cited*) that rarely occur except in the passive. By contrast the SiBol/Port list has many pronouns and verbs.
2. Discourse structure is a central theme for academic prose, and discourse markers appear in the BAWE list.
3. The nouns listed under 'theory' for BAWE are a set of highly general and polysemous words, most of which have concrete meanings as well as abstract ones, so defy easy classification: a *solution* can be a solution in water as well as a solution of a problem, *development* can be what a plant does, or what a society does.

BAWE keywords	***SIBOL/Port keywords***
Nouns: **Experimental method:** analysis control data error equation factor factors graph model method output sample variables **Theory:** behaviour characteristics concept context development differences effect effects extent function information individual individuals knowledge nature states social systems process product products results theories theory type value values **Not-quite-so-general:** cell cells communication environment gender human labour language learning meaning protein species temperature **Academic process:** eds essay program project research section study **Verbs:** associated cited considered defined increase increased occur required shown use used using **Adjectives:** different important negative significant specific various **Prepositions:** between upon within **Discourse connectives:** due *(to)* hence therefore these thus whilst	**Time:** ago day days former last latest minutes month months next never night now season summer week weekend year years yesterday **Money, numbers:** billion cent five million per pounds shares six **Bosses:** chairman chief director executive head minister secretary spokesman **Sport:** ball club football game games hit manager match played player players playing team top victory win won **Verbs:** announced came come get going got had say says said think told took want went **Pronouns:** him his you your she who (there may well have been more but for the three-letter minimum) **Prepositions/particles:** about ahead back down like round off **News/politics:** cut died election news party police **Adjectives:** big young **Non-time adverbs:** just really **Other:** bit com home house music thing television www

Table 3. Keywords; BAWE $vs.$ SiBol/Port

4. Newspapers are very interested in time (and money).
5. Sport forms a substantial component of SiBol/Port.

Both journalism and academic writing have been objects of extended study, with corpus work including, for journalism, [10, 9], and for academic writing, [11–13]. Our current goal is simply to show how keyword methods can very quickly and efficiently contribute to those areas of research, as well as highlighting aspects of contrasting datasets that researchers might not have considered before.

4 Designed corpora and crawled corpora

Two contrasting approaches to corpus-building are:

Design: Start from a design specification and select what goes into the corpus accordingly

Crawl: Crawl the web, and put whatever you find into the corpus.

The British National Corpus is a model designed corpus. UKWaC and enTenTen are both crawled.

The relative merit of the two approaches is a live topic [14–16, 1]. Crawling is very appealing, since it involves no expert linguist input, is fast and cheap, and can be used to prepare vast corpora. But can we trust a crawled corpus? How do we know what is in it, or if it does a good job of representing the language?

4.1 BNC *vs.* enTenTen

enTenTen keywords	***BNC keywords***
Pronouns: our your **Encoding:** don percent **Web:** com site email request server internet comments click website online posted web list access data search www files file blog address page **University:** article campus faculty graduate information project projects read research science student students **American spelling :** behavior center color defense favor favorite labor organizations program programs toward **Bible:** believe evil faith forth sin soul thee thou thy unto upon **Politics:** current federal global laws nation president security world **Creative industries:** author content create digital film game images media movie review story technology **Informal:** folks guess guy guys kids **Language change:** issues **Other:** code efforts entire focus human include including human located mission persons prior provides	**Pronouns:** he herself her **Encoding** **Speech transcription:** cos cent erm gon per pound pounds **Numbers:** eight fifty five forty four half hundred nine nineteen seven six ten thorty three twenty two **British spelling:** behaviour centre colour defence favour labour programme round towards **British lexical variants:** bloody pupils shop **Past tense verbs:** got felt turned smiled sat looked stood was said been seemed had went were knew put thought **Particles:** away back down off **Local government:** council firm hospital local industrial police social speaker **Household nouns:** bed car door eyes face garden girl hair house kitchen mother room tea **Informal:** alright mean quite perhaps sort yeah yes **Language change:** chairman **Other:** although club considerable could head know main manager night there studio yesterday

Table 4. Keywords; enTenTen $vs.$ BNC

Here there is no simple story to tell regarding formality. Both lists include pronouns: the BNC has three third person singular feminine pronouns, whereas enTenTen has a first and a second person one. This, along with the 'informal' cluster, suggests enTenTen has more interactional material. It is the BNC that has the verbs but they are all in the past tense. Biber shows that narrative is a central dimension of variation in language. The cluster of features associated with narrative includes past tense verbs and third person pronouns. The BNC has 16% fiction, and also a large quantity of newspaper, where the ‘story’ is central, so it seems that these two components place the BNC further along the narrative dimension than enTenTen. The daily newspaper material accounts for an abundance of *yesterday*, and the fiction, the ‘household nouns’ cluster.

don (in enTenTen) and *gon* (in BNC) arise from tokenisation issues: *don't* and *gonna* (‘going to’) both have different possible tokenisations, and different choices were made in the processing of the two corpora. Also there are different conventions on spelling out ‘%’.

10% of the BNC is transcribed speech. The BNC transcription manual specified that *erm* (pause filler), *cos* (spoken variant of *because*) and *gonna* (again) should be transcribed as *erm, cos* and *gonna*, and that *pound(s)* should be spelt out. So should numbers: hence the numbers cluster.

Whereas enTenTen has a biblical seam, the BNC has a local government one.

Language has changed in the two decades separating the two corpora, with *chairman* becoming less politically acceptable and *issues* acquiring a popular new sense, as in “we have some issues with that”. And the world has changed: the web was unknown outside academia at the time of the BNC. Hence the web cluster.

We now have two comparisons involving enTenTen that we can compare. Some of the clusters (bible, American spellings) are much the same in both cases but most are quite different. Both tell us about enTenTen, but from different vantage points. The more corpora we compare our corpus with, the better we will get to know it.

4.2 Czech: CNC *vs.* czTenTen

The Czech National Corpus, as used in this study, comprises three ‘balanced’ 100m-word components (from 1990-99, 2000-04 and 2005-09) and one billion words of newspapers and magazines (1989-2007) [17]. czTenTen is a web corpus crawled in 2011. Here there were no constraints on case or item-length, the simplemaths parameter was again 100, and there was a little manual editing to remove tokenization anomalies, words with missing diacritics, and Slovak words, and to merge multiple forms of the same lemma.

As with enTenTen *vs.* BNC, the web corpus is more interactional, with many first and second person forms of verbs, and 2nd person personal pronoun. As for English, there is a web cluster. With a large part of CNC being newspaper, it shares narrative characteristics like past tense reporting verbs with the BNC but also with SiBol/Port, with which it also shares politics, economics, sport and time.

5 Quantitative approaches: measuring distances between corpora

In this paper we have presented a keywords-based method of comparing corpora. This is just one method, and a qualitative one, empoying skills typically taught in humanities departments. enTenTen is more similar to UKWaC than the BNC, but this fact has not been foregrounded in the keyword-list analysis. A complementary approach is a quantitative one, in which we measure distances between corpora. [18] makes the case for corpus distance measures (and the closely related case for homogeneity/heterogeneity

CczTenTen keywords	*CNC keywords*
informal: taky, teda, moc, sem, fakt, dneska, taky, sme, zas, dost, můžu, tak, takže, nějak, prostě, ahoj, tohle, super, jinak, fotky, jak, takže, holky, takhle, fajn, doufám **verbs in 1st or 2nd person:** jsi, můžete, najdete, mám, děkuji, bych, budete, máte, máš, nevím, prosím, děkuji, myslím, jsem, budu, díky, chci, naleznete, nemám, ráda, budeš, nejsem, vím, chcete **pronouns** (half are forms of second person plural): Vám, Vás, Vaše, vám, moje, ten, Váš, něco, nějaký, tebe, ono, vás, toto, nějaké, toho **adverbs:** trošku, naprosto, opravdu, akorát, docela, bohužel, trochu, krásně, jinak, pěkně, tam **web, computing:** web, aplikace, stránky, Windows, verze, video, online, odkaz, server, nastavení **other:** dobrý, dle, jestli, článek, pokud, zeptat, použití, nachází, pomocí, snad, jelikož, napsat, odpověď, den, nebo, přeci, týče	**politics/functions/institutions:** policie, starosta, ODS, unie, radnice, ČSSD, ředitel, úřad, policisté, policejní, nemocnice, klub, předseda, šéf, vedoucí, USA, ministr, prezident, banka, vláda **economics/mostly numerals:** koruna, tisíc, procento, milion, pět, čtyři, miliarda, tři, deset, šest, sto, osm, sedm, dvacet, dolar, padesát **spokesman-related words** (told, stated, said, spokesman, explained): uvedl, řekl, mluvčí, dodal, tvrdí, uvedla, prohlásil, říká, vysvětlil, řekla, sdělil **sports:** trenér, utkání, domácí, liga, kouč, vítězství, soutěž **names:** Jiří, Josef, Jan, Jaroslav, Vladimír, Pavel, Petr, Miroslav, Václav, Zdeněk, František, Karel, Milan **places:** Praha, ulici, náměstí, Brno, Ústí, Plzeň, Králové, město, České, Ostrava, Hradec, Liberec **time:** včera, letos, hodina, sobota, loni, pondělí, neděle, dosud, nyní, zatím, víkend, úterý **other:** výstava, expozice, však, totiž, Právo, muž, například, zřejmě, zhruba, lidé, uskuteční

Table 5. Keywords; czTenTen $vs.$ CNC

measures) and makes some proposals. Using a variant of the method found to work best there, we computed the distances between the five English corpora.
The most similar two corpora are indeed enTenTen and UKWaC, although enTenTen and BNC are only slightly further apart. Of the five, BAWE, comprising exclusively academic prose, is the outlier.

A careful comparison between two corpora generally requires both quantitative and qualitative approaches.

6 Functionality in the Sketch Engine

The Sketch Engine is a leading corpus query tool, in use for lexicography at Oxford University Press, Cambridge University Press, Collins, Macmillan, Cornelsen, Le Robert, and ten national language institutes (including those for Czech and Slovak), and for teaching and research at over one hundred universities worldwide. The Sketch Engine

	BNC	enTenTen	SiBol/Port	UKWaC
BAWE	2.15	1.98	2.39	1.92
BNC		1.51	1.64	1.63
enTenTen			1.75	1.42
SiBol/Port				1.74

Table 6. Distances between English corpora.

website has, already installed in the Sketch Engine and accessible to all users, large corpora for over sixty languages. For English there are many others as well. Users can install their own corpora and make comparisons between it and any other corpus (or subcorpus) of the same language.

The Sketch Engine provides functions for generating a range of lists, including all the keyword lists used in this paper. The interface for specifying a list (which may be a simple list, or a contrastive ‘keyword’ one) is shown in Figure 2.

Until recently one might have argued that, while the procedures outlined in this paper for getting to know your corpus were sensible and desirable, they were hard to do, and unreasonable to expect of busy researchers, particularly those without programming skills. As it is now straightforward to use the Sketch Engine to prepare the lists, this argument is no longer valid.

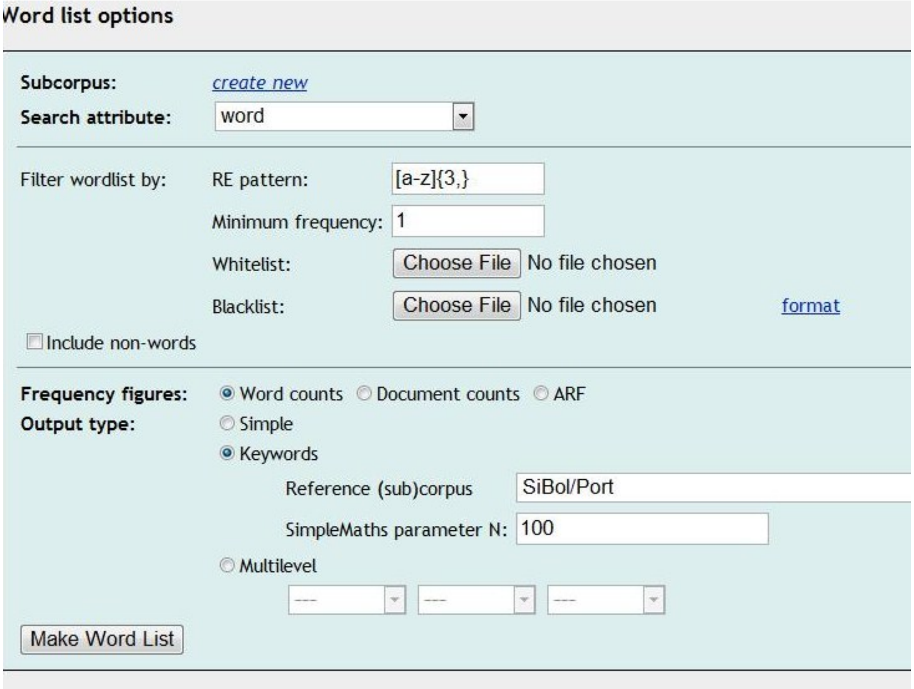
7 The Bigger Picture

Corpora are not easy to get a handle on. The usual way of engaging with text is to read it, but corpora are mostly too big to read (and are not designed to be read). So, to get to grips with a corpus, we need some other strategy: perhaps a summary. A summary in isolation is unlikely to be helpful, because we do not know what we expect a corpus summary to look like. The summary only becomes useful when we can compare it with a summary for another corpus. A keyword list does this in the most straightforward way: it takes frequency lists as summaries of the two corpora, and shows us the most contrasting items.

Corpora are usually mixtures, and any two corpora vary in a multitude of ways, according to what their components are, and in what proportions. Any large, general corpus will have components that we do not expect, and maybe do not want. Keyword lists are a methodology for finding what they might be.

Keyword lists are an approach for all three of:

- General comparison of two corpora with unknown differences
- Quality control: identifying pre-processing errors, unwanted content, and other anomalies
- Comparing and contrasting different text types, varying, for example, according to:
  - Register, genre
  - Domain, subject area
  - Time, for studies of language change
  - Region



Word list options

Subcorpus: create new
Search attribute: word

Filter wordlist by:
RE pattern: [a-z]{3,}
Minimum frequency: 1
Whitelist: Choose File No file chosen
Blacklist: Choose File No file chosen format
□ Include non-words

Frequency figures: ● Word counts ○ Document counts ○ ARF
Output type:
○ Simple
● Keywords
Reference (sub)corpus SiBol/Port
SimpleMaths parameter N: 100
○ Multilevel
--- --- ---

Make Word List

Fig. 2. Sketch Engine's form for specifying a word list, including specifying whether the list should be of word forms, lemmas, word classes etc., any pattern that should be matched, and whether the list should be a simple list or a keyword list. For English there are twenty corpora, installed and available, that one might choose to make comparisons with.

7.1 The moral of the story

My title is "Getting to know your corpus". You should.

If you publish results when you have not, it is like a drug company publishing and saying "use this drug" although they have not noticed that the group of subjects who they tested the drug on were largely under 25, with a big cluster who had travelled round South America, and none of them were pregnant. We need to guard against such bad science, and, if we intend to continue to be empiricist, and to work with data samples – corpora – in linguistics, we need to get to know our corpora.

The Sketch Engine does the grunt work. What remains is the interesting bit. Do it.

Acknowledgements With thanks to Vit Suchomel for the analysis for Czech.

References

1. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The wacky wide web: a collection of very large linguistically processed web-crawled corpora. Corpora **43**(3) (2009) 209–226

2. Pomikàlek, J., Rychlý, P., Kilgarriff, A.: Scaling to billion-plus word corpora. Advances in Computational Linguistics. Special Issue or Research in Computer Science **41** (2009)
3. Biber, D.: Variation across speech and writing. Cambridge University Press (1988)
4. Biber, D.: Dimensions of Register Variation: a cross-linguistic study. Cambridge University Press (2006)
5. Heylighen, F., Dewaele, J.M.: Formality of language: definition, measurement and behavioral determinants. Technical report, Free University of Brussels (1999)
6. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press (1999)
7. Kilgarriff, A.: Language is never ever ever random. Corpus Linguistics and Linguistic Theory **1**(2) (2005) 263–276
8. Heuboeck, A., Holmes, J., Nesi, H.: The BAWE corpus manual. Technical report, Universities of Warwick, Coventry and Reading (2007)
9. Partington, A.: Modern diachronic corpus-assisted discourse studies MD-CADS on UK newspapers: an overview of the project. Corpora **5**(2) (2010) 83–108
10. Baker, P., Gabrielatos, C., McEnery, T.: Discourse Analysis and Media Bias: The representation of Islam in the British Press. Cambridge University Press (2012)
11. Biber, D.: University Language: A corpus-based study of spoken and written registers. John Benjamins (2006)
12. Paquot, M.: Academic Vocabulary in Learner Writing. Continuum (2010)
13. Kosem, I.: Designing a model for a corpus-driven dictionary of Academic English. PhD thesis, Aston University, UK (2010)
14. Keller, F., Lapata, M.: Using the web to obtain frequencies for unseen bigrams. Computational Linguistics **29**(3) (2003) 459–484
15. Sharoff, S.: Creating general-purpose corpora using automated search engine queries. In Baroni, M., Bernardini, S., eds.: WaCky! Working papers on the Web as Corpus. Gedit, Bologna (2006)
16. Leech, G.: New resources, or just better old ones? the holy grail of representativeness. In Hundt, M., Nesselehauf, N., Biewer, C., eds.: Corpus Linguistics and the Web. Rodopi, Amsterdam (2007) 133–149
17. Čermák, F., Schmiedtová, V., Křen, M.: Czech national corpus - syn. Technical report, Institute of the Czech National Corpus, Prague, Czech Republic Accessible at `http://www.korpus.cz`. Accessed on 2012-06-08.
18. Kilgarriff, A.: Comparing corpora. Int. Jnl. Corpus Linguistics **6**(1) (2001) 263–276