Comparable Corpora Within and Across Languages, Word Frequency Lists and the KELLY Project

Adam Kilgarriff

Lexical Computing Ltd
Brighton, UK
adam@lexmasterclass.com

Abstract

Word frequency lists play a pivotal role as we explore and exploit comparable corpora. They form a compact summary of what is in a corpus. They also make it possible to assess how similar two corpora are, and how they contrast with each other. They are also widely used by educators, psychologists and publishers in their own right. In the recently-started EU project KELLY, we are exploring these issues across nine languages, including starting from loosely comparable corpora across languages. The paper describes how word frequency lists can be developed from corpora, and how they might be used, complete with plans and experiences from Kelly.

1. Measuring comparability

What makes comparable corpora 'comparable'? They should have roughly the same text type(s), covering the same subject matter, in the same proportions. Given that definition, comparable corpora may be of the same or different languages.

In 2003 Maia could not help but conclude that "comparability is in the eye of the beholder" (Maia, 2003). This is not a satisfactory state of affairs: we do not want the sampling for the datasets underlying our scientific endeavour to be subjective. We could avoid subjectivity if we could make measurements. We would like to be able to measure how comparable, or similar, two corpora are.

Then it becomes useful that the definition of comparability (or, hereafter, similarity) relates equally to same-language and different-language corpora. It gives us a reference point: any corpus is entirely similar to itself. It also gives us some history in quantitative comparison of same-language corpora.

(Biber, 1988) opened the field, showing how corpus counts could be used for the systematic study of contrasts between language varieties. For him the object of the study was the differences between the text types, rather than calibrating differences between corpora. I explored the calibration question in (Kilgarriff, 2001). At the time the issue was of largely theoretical interest as corpus users tended to be beggars not choosers: most corpus users were using any corpus of approximately the right type that they could lay their hands on, options being few and far between.

Since then we have had BootCaT and the 'web as corpus' strategy, making it possible to quickly and cheaply build a corpus to a specification (Baroni and Bernardini, 2006). In that model, once you have built a corpus the overriding questions are "is it what I wanted? What kind of a corpus (in terms of text types, subject matter, proportions) is it?" The collection strategy may have been more, or less, succesful in gathering what was wanted, and will probably have picked up some things that were not wanted along the way, so the builder wants to evaluate the corpus.

The simplest place to start is a word frequency list.

2. Word Frequency Lists

Word frequency lists can be seen from several perspectives. For computational linguistics or information theory, they are also called unigram lists and can be seen as a compact representation of a corpus, lacking much of the information in the corpus but small and easily tractable.

Psychologists exploring language production, understanding, and acquisition are interested in word frequency, as a word's frequency is related to the speed with which it is understood or learned so frequency needs to be allowed for in choosing words to use in psycholinguistic experiments. Educationalists are interested too, so frequency can guide the curriculum for learning to read and similar. To these ends, Thorndike and Lorge prepared *The Teacher's WordBook of 30,000 words* in 1944 by counting words in a corpus, creating reference set used for many studies for many years (Thorndike and Lorge, 1944). It made its way into English Language Teaching via West's General Service List (West, 1953) which was a key resource for choosing which words to use in the English Language Teaching curriculum until the British National Corpus[1] replaced it in the 1990s.

In language teaching word frequency lists are used for:

- defining a syllabus
- deciding which words are used in
  - learning-to-read books for children
  - textbooks for non-native learners
  - dictionaries
  - language tests for non-native learners.

2.1. Creating word frequency lists

There are three ways to get a word list, for purposes as above

- copy
- guess

[1] http://natcorp.ox.ac.uk

- count

Most word lists for most languages have used the first and the second. Where there are no corpora available, this is forgiveable. In 2010, this is no longer an excuse for any medium-sized or larger language. Principled word lists must be based on corpora.

Following on from Thorndike and Lorge, in the 1960s Kučera and Francis developed the Brown Corpus, a carefully compiled selection of current American English of a million words drawn from a wide variety of sources (Francis and Kučera, 1982). They undertook a number of analyses of it, touching on linguistics, psychology, statistics, and sociology. The corpus has been very widely used in-all of these fields. The Brown Corpus is the first modern English-language corpus, and a useful reference as a starting-point for the sub-discipline of corpus linguistics (from an English-language perspective).

While the Brown Corpus was being prepared in the USA, in London the Survey of English Usage was under way, collecting and transcribing conversations as well as gathering written material. It was used in the research for the Quirk *et al* Grammar of Contemporary English (Quirk et al., 1972), and was eventually published in the 1980s as the London-Lund Corpus, an early example of a spoken corpus.

My personal involvement in word lists came about when, in 1994 and 1995 I counted the words in the (then new) British National Corpus, the first time for inclusion in LDOCE3 (LDOCE3, 1995; Kilgarriff, 1997), and the second, for the world at large, putting them on the web. The web version has been used and used, for example as the source of the JCET 8000 which defines the English syllabus in Japan. So people have come to think of me as an expert on word lists. The work described below is an attempt to live up to that cheaply-earned reputation.

There are various steps in getting from corpus to high-quality word list, as spelt out below.

2.2. Core and sublanguage

A language consists of core vocabulary and sublanguages. Core vocabulary is used across the board, sublanguage vocabulary changes according to what is being talked about (and in what genre) so will be different from corpus to corpus. My suspicion is that the core is quite small. When preparing word frequency lists, one strategy is to, firstly, identify the core, and secondly, decide which sublanguages are privileged, in the context of, for example, language learners: perhaps sublanguages like family relationships (*brother, sister, uncle aunt* etc) and body parts (*eye ear nose throat wrist shoulder* etc).

2.3. What is a word

In English, a (textual) word is, to a first approximation, an item found between spaces comprising a-z characters. English is a particularly easy language here. Chinese does not have spaces between the words at all, Arabic (and, to a lesser extent Italian) often incorporates pronouns, articles and other grammatical items into the same space-delimited object. Swedish, Norwegian, German and Dutch have compounding and separable verbs.

2.4. Words and lemmas

In texts we find word forms (*invade invading invades invaded*) whereas in dictionary we find lemmas, also called dictionary headwords: just *invade*. Word lists for educators should be lists of lemmas. To get from word forms to lemmas is the process of lemmatisation: not needed at all for Chinese (which has no inflections), simple for English, middling for Italian, Greek, Norwegian and Swedish, and very complex for Russian, Polish and Arabic.

2.5. Grammatical classes

English *brush* can be a noun or a verb. Should the noun and the verb be counted as separate for purposes of the word list, or as a single item? Some dictionaries treat them as separate headwords, others as the same. Languages also vary: Chinese has a weak sense of word class so for Chinese, giving different noun and verb entries is less appealing as it may force decisions as to whether a word is a noun, a verb, or both. English has a lot of freedom for using nouns as verbs and *vice versa*, but, in context, there is usually a right answer as to whether a word is being used as a noun or verb (or adjective; for *-ed* and *-ing* forms this becomes difficult).

If the word list is to distinguish different word classes, we shall need a taxonomy of word classes for the language. It is desirable that this is the same for each language except where there is a good linguistic reason why it cannot be. The work done in EAGLES and associated projects presents an approach for this task (EAGLES, 1996).

2.6. Non-central word types

There are various marginal classes of word:

- numbers, ordinals, fractions
- names (of people, places of various kinds, organisations)
- countries, currencies, nationalities, languages, ethnic groups, religions and philosophies and their adherents (nouns and adjectives)
- days of week, months, decades, festivals
- abbreviations, initials, acronyms
- informal, slang, offensive language
- dialect words, regional variants

Decisions will be required on what to include.

2.7. Multiwords

English *according to* is, from a linguistic point of view, a word, but is written with a space. Let us call all such items multiwords. (This does not relate to Chinese or Japanese as they are not written with spaces between words at all.) Big classes of multiwords for English are phrasal verbs, compound prepositions and compound nominals. Linguistically, word lists should contain multiwords but, unlike simple words, we cannot easily count them. If we count all two-word strings in an English corpus the commonest is *of the* but no-one wants that in their wordlist. Very many

common two word strings are not multiwords. So, if we use a direct strategy for including multiwords in a wordlist, we are back to copying or guessing.

2.8. Homonymy

The English noun *bank* can be the side of a river or a financial institution. Should these count as two separate items in a frequency list?
Every different dictionary makes different decisions about what is to count as a separate meaning so if we try to build homonymy into word lists, we shall introduce some arbitrariness.

3. Contrasting corpora

Word frequency lists as compact representations of corpora, and word lists for use by educators may seem very different things, but if the latter do not in some way come from the former we are either copying or guessing. A word frequency list is only of value for educators if it is based on 'the right corpus', which throws us back on the question of how we might assess corpora.
We assess a corpus by comparing its word frequency list with the list from another corpus. While other approaches are possible (for example, measuring cross-entropy between the corpora) it is harder to interpret their outcomes.
The simplest strategy is to compare the top ten, or top twenty, words in the two lists. Often, many of them are the same, and it is not clear whether there are interesting differences between the words that are in a different position in the two lists.
A better method is to identify the words that are most different in their frequencies between the two corpora: the keywords of each with respect to the other. To do this we

- normalise frequencies to per-million
- for each word, calculate the ratio between normalised frequencies in the two corpora
- sort by ratios
- the top and bottom items are the keywords (of the first corpus versus the second, and vice versa).

We can make the scheme more flexible, and address the fact that we cannot compute a ratio against zero, by adding a constant to all normalised counts before computing ratios. The higher the constant, the more the frequency list will focus on higher-frequency items, as shown in (Kilgarriff, 2009).
Provided the lists are prepared in uniform ways in realation to tokenization, lemmatisation etc., an examination of the keywords will allow us to rapidly identify the main contrasts between two corpora. We used this method to compare a Dutch web corpus, NlWaC, with the ANW corpus, a balanced corpus of 100 million words built to support the lexicography for the ANW, a major new dictionary of Dutch. It comprises: present-day literary texts (20%), texts containing neologisms (5%), texts of various domains in the Netherlands and Flanders (32%) and newspaper texts (40%).
The twenty highest-scoring (ANW) keywords and the twenty lowest-scoring (NlWaC) keywords, with English glosses and clustered by themes, are given in *Table* 1.
The classification into themes was undertaken by checking where and how the words were being used. The analysis shows that these two large, general corpora of Dutch have different strengths and weaknesses, and different areas that might be interpreted as over-representation or under-representation. The ANW has a much stronger representation of Flemish (the variety of Dutch spoken in Belgium). It has 20% fiction: *keek* (looked, watched) is used almost exclusively in fiction. It is 40% newspaper and newspapers talk at length about money (which also interacts with time and place: franks were the Belgian currency until 1999; also the units were small so sums in franks were often in millions or even billions). There is a particularly large chunk from the Meppel local newspaper. Most occurrences of *foto* were in "Photo by" or "Photo from" and of *auteur*, in newspaper by-lines, which might ideally have been filtered out. Daily newspapers habitually talk about what happened the day before, hence *gisteren*.
NlWaC has a large contingent of religious texts. It is based on Web texts, some of which could have been more rigorously cleaned to remove non-continuous-text and other non-words like URL components *www, http, nl*. The English might appear to be because we had gathered mixed-language or English pages but when we investigated, we found most of the instances of *and* and *the* were in titles and names, for example "The Good, the Bad and the Ugly", where the film was being discussed in Dutch but with the title left in English.
This analysis (also in (Kilgarriff et al., 2010)) is presented here to illustrate how we can assess how 'comparable' same-language corpora are.[2]

4. The KELLY Project

KELLY is an EU Lifelong Learning Project with the goal of developing language-learning cards, with a word in one language on one side and its translation on the other. The languages involved are Arabic, Chinese, English, Greek, Italian, Norwegian, Polish and Swedish. In the past, tools of this kind have rarely been corpus-based, or even corpus-informed. In Kelly we hope to be able to prepare high-quality lists which are fully corpus-based.
The method is as follows ('lempos' is shorthand for lemma plus part of speech; our lists will be lists of lemposes):

- prepare (tokenised, lemmatised, POS-tagged) corpora
- Generate lempos-lists (call these M1 lists, for monolingual first-stage lists)
- Study keywords lists from different corpora; review and fix anomalies to give M2 lists
- Translate into all eight other languages, to give T1 (first Translated) lists

[2] See (Kilgarriff, 2001) for global figures of how similar two corpora are; a drawback of these figures is that they can only be used to compare similarity scores between two or more pairs of corpora, and cannot be interpreted in isolation.

- Review candidate additions to lists
- Review and finalise monolingual lists and bilingual lists for word cards (M3, T2 lists)

We hope that omissions and failings of the M2 list for a language might be rectified by the set of translations of lists from eight other languages into that language. In particular, although the M2 list will not include multiwords, multiwords are, by definition, akin to a single word linguistically so one can expect them to have single-word equivalents in other languages, so they are likely to feature as translations. We expect to acquire many items to add to M2 lists in this way, to give M3 lists.

At time of writing M2 lists are being finalised.

We wished to use comparable corpora for each language for preparing M1 lists. The only type of large, general corpus that we could obtain for all languages was a BootCaT-style web corpus. (For Swedish, where we did not know of any such corpus, we prepared one (Kilgarriff et al., 2010).)

To get from M1 lists to M2 lists, which can reasonably be presented to translators, a gamut of issues have been encountered. Junk needed deleting. POS-taggers and lemmatisers made many errors. The most heated debates at our initial project meeting related to multiwords and homonymy, with the one argument being that lists including multiwords and homonymy decisions would include a large dose of arbitrariness, and the counter-argument being that the eight translators-out-of-English needed guidance, to know, for example, that the English noun *mean* occurred in the M2 list because of its occurrence in *by means of*. For homonyms, how were the eight translators to know whether to translate *money bank* or *river bank*? Consortium members for different languages have adopted slightly different strategies on these issues, each according to their own perspective.

A further problem relates simply to the text type mix of web corpora. A recent email thread was titled *alphabet, orange, banana and elbow*: Swedish equivalents of these words were not in the top-6000 list, yet they were basic vocabulary. Responses have included:

- for English and Norwegian, corpora of conversational speech were available and have been used as comparison corpora, so words such as these have entered the M2 lists via that route if not otherwise
- if the words are there in the M2 list for one language, it is likely they will percolate across to all languages
- we may do further checking of lists against textbook vocabularies
- we may allow addition of items simply because the person preparing the list knows they should be there!
- I am not sure that *elbow* is such a common word in any text type, but there is nonetheless an argument for including it as a body-part term: as noted above, some domains are privileged from a language-learning perspective. (It is part of the project's agenda to relate word cards to the language levels as defined in the Common European Framework (CEF, 2010). The CEF makes explicit reference to some thematic areas including food and drink, and health and body care.)

The list of words added to the English M2 list from the English conversational speech corpus started with *yeah mum dad okay sorry hello dear*. We fear we underestimated the mismatch between web corpus frequencies and frequencies from everyday language use, as required by a learner.

4.1. Frequencies and points

Where the person looking at keywords lists decides that a word needs adding to the list, or a word has too high or too low a score, how should they implement it? The word cards are, in due course, to be divided into six levels (of 1500 words each) so we need to retain order information. The list is, initally, a frequency list, so should the person make up a frequency that puts it in a position that they judge to be appropriate?

Making up frequencies feels monstrous. We have a slightly less bad variant: first translate frequencies into points, then promote or demote words by adding or subtracting points.

The initial list is of 6000 items: the top 500 get twelve points, the next 500, eleven, the next 500, ten, and so on. When a word is introduced into the list from the top of a spoken conversation list, it will be introduced with twelve points; ones introduced from lower down the spoken list may be introduced with a smaller number of points. Words found to be entirely absent in the spoken corpus can be demoted, say, four points.

We begin with each band containing 500 items, but that will not stay true. If, at some point, we need to specify the top 1500 items, we can use frequency in the web corpus as a second level of sorting for words with the same number of points.

While the strategy makes no claim to objectivity, it provides a framework for systematic amendment of a starter list.

4.2. The Translations database

All translations will be entered in a database. With translations of 6000 items for each of nine languages into all eight other languages, it will be a large and rich resource.

There are just 6000 words in M2 lists as against 9000 word cards for each language pair eventually required. We anticipate making up the difference from "back translations": words and multiwords which were not in M2 but do occur as translations from other languages. In addition to the 6000 M2 words for a language, there will be up to $6000 \times 8 = 48,000$ additional items: most will overlap with the M2 list and each other, and it remains to be seen how many are useful. We envisage adding items according to rules such as:

> if a multiword or word not in M2 occurs more than once as a translation (either as the translation of equivalent terms from two different other languages, or otherwise) then it is a candidate for inclusion.

5. Summary

Word frequency lists play a pivotal role as we explore and exploit comparable corpora. They form a compact summary of what is in a corpus, and make it possible to assess where two corpora of the same language are comparable, and how they contrast with each other. They are also widely used by educators, psychologists and publishers in their own right. In the recently-started EU project KELLY, we are exploring the preparation of word lists from corpora across nine languages, including starting from loosely comparable corpora across languages and the large-scale translation of lists. We hope to shed light on how we might measure comparability between corpora across, as well as within, languages in due course.

Acknowledgments

This work was supported by the EU Lifelong Learning Project KELLY. Much of the work descussed has been undertaken together with members of the KELLY team.

6. References

Marco Baroni and Silvia Bernardini, editors. 2006. *Wacky! Working Papers on the Web as Corpus*. Gedit, Bologna.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press.

CEF. 2010. Common european framework of reference for languages. Technical report, Council of Europe.

EAGLES. 1996. Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. a common proposal and applications to european languages. Technical Report EAG-CLWG-Morphsyn/R, ILC-CNR, Pisa.

N. Francis and H. Kučera. 1982. *Frequency Analysis of English Usage*. Houghton Mifflin, Boston.

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and Avinesh PVS. 2010. A corpus factory for many languages. In *Proc. LREC*, Malta.

Adam Kilgarriff. 1997. Putting frequencies in the dictionary. *Int Jnl Lexicography*, 10(2):135–155.

Adam Kilgarriff. 2001. Comparing corpora. *Int Jnl Corpus Linguistics*, 6(1):1–37.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proc. Corpus Linguistics*, Liverpool, UK.

LDOCE3. 1995. *Longman Dictionary of Contemporary English*. Longman, 3rd edition.

Belinda Maia. 2003. What are comparable corpora? In *Multilingual Corpora: Linguistic Requirements and Technical Perspectives. A Workshop on the Corpus Linguistics Conference*, Lancaster, UK.

Randolph Quirk, Sydney Greenbaum, Geoff Leech, and Jan Svartvik. 1972. *A Grammar of Contemporary English*. Longman.

E. L. Thorndike and I. Lorge. 1944. *The Teachers Word-Book of 30,000 words*.

Michael West. 1953. *A General Service List of English Words*. Longman.

ANW			NlWaC		
Theme	**Word**	**English gloss**	**Theme**	**Word**	**English gloss**
Belgian	Brussel	(city)	Religion	God	
	Belgische	Belgian		Jezus	
	Vlaamse	Flemish		Christus	
Fiction	Keek	Looked/watched		Gods	
Newspapers	vorig	previous	Web	http	
	kreek	watched/looked		Geplaatst	posted
	procent	Percent		Nl	(Web domain)
	miljoen	million		Bewerk	edited
	miljard	billion		Reacties	Replies
	frank	(Belgian) Franc		www	
	Zei	said	English	And	In book/film/song titles, names etc
	aldus	thus		The	
	Meppel	City with local newsp	History	Arbeiders	workers
	gisteren	yesterday		Dus	thus
	Foto	Photo		Macht	power
	Auteur	Author		Oorlog	war
	Van	(in names)		Volk	people
Pronouns	Hij	Him/he	Pronouns	We	we
	haar	She/her(/hair)		Ons	us
	Ze	(They/them)		Jullie	you

Table 1: Keywords in ANW and NlWaC