

# Genre in a frequency dictionary

Adam Kilgarriff  
Lexical Computing  
Ltd., UK

adam@lexmasterclass.com

Carole Tiberius  
INL  
Leiden, NL

carole.tiberius@inl.nl

## 1 Routledge Frequency Dictionaries

The Routledge Frequency Dictionary series is now well-established with titles available for eight languages. They give the five thousand commonest words for their language, with indexes for access alphabetically or by frequency, and additional features varying from dictionary to dictionary including English translations, example sentences and their translations, listings by word class, collocations, and tables for the vocabulary of various domains. We are currently preparing a Dutch title to add to the series.

## 2 Genre in dictionaries

There is some labelling in existing titles in the series for genre, region and dialect, as there is in traditional dictionaries, but the core of the dictionary is based on a single language-wide, genre-blind corpus-based list. While it has always been a premise of dictionary-making that one could work with ‘the language in general’, and there is just a small minority of words that are marked for genre, it leaves those of us who work with corpora uncomfortable. In any corpus-based dictionary, one has to choose what text types to include in the corpus, and those choices will determine the outlook on the language that the dictionary presents. In a frequency dictionary, the whole issue is staring the compilers in the face, since the dictionary content is directly determined by the corpus.

For our dictionary, we decided to foreground genre in a way that other dictionaries in the series had not, by setting aside the ‘one list’ approach and presenting a number of lists, some of which would be genre-specific. This paper is a discussion of the issues that arise in that context, and our responses to them.

## 3 Corpus construction

We chose to use four different kinds of text – hereafter, ‘genres’ – for our corpus: spoken conversation, fiction, newspaper and web. Of course there are arguments for going for a finer-grained set of genres, like Brown, but we restrained ourselves to genres where we had access to large numbers of texts, and to a modest number of genres

so that the complexities of analysis were manageable.

Three of our genres are three of the four used by Biber and colleagues in much of their work. Their fourth is ‘academic’; our fourth is ‘web’, in response to the growing importance of the web in our lives since Biber’s research programme began, and also because of the heterogeneity of the web. It is less likely that we shall miss common words that mainly occur in genres other than conversation, fiction and newspaper.

For **conversation** we used the Spoken Dutch Corpus (CGN – *Corpus Gesproken Nederlands*). The corpus contains 900 hours (ca. 9 million words) of Standard Dutch spoken by Flemish and Dutch adults. We used only those parts where full transcription was available.

For **newspaper** we used the newspaper material which is included in the SONAR<sup>1</sup> corpus, a large reference corpus of written Dutch.

For **fiction**, 25 books per year were available for each year from 1970 to 2009, including essays, romans, plays and stories.

For the **web** we used data from the SONAR corpus including material from blogs, discussion lists, e-magazines, press releases, websites and wikipedia.

## 4 A fixed-sample-size corpus

In the Brown corpus, all samples are 2000 words long. While the strategy of truncating samples has its detractors – notably John Sinclair, who insisted that corpora should comprise complete texts – and we would not dispute that it makes a corpus unsuitable for some research questions, it also has many advantages. If all samples are 2000 words long, then we immediately know that any results will not be distorted by different sample lengths. A frequency of ten for a word in one sample will have the same weight as its frequency of ten in another.

Interpreting statistics where sample sizes vary often becomes complex and subtle, and we suspect one of the reasons for the ongoing success of the Brown model is its fixed sample size.

One of the central problems in preparing frequency lists is the whelks problem: if there is a text about whelks (a variety of mollusc) then the word *whelk* will probably occur many times. We would rather not give all of those occurrences equal weight in our word frequency list. (Gries 2008 presents a review of methods used to address the issue, and the Routledge dictionaries use a range of mathematical devices.) One simple and appealing strategy is not to count the number of occurrences of each word, but the number of samples the word

<sup>1</sup> <http://lands.let.ru.nl/projects/SoNaR/>

occurs in. Then, however many times whelks are mentioned in a sample, it will just count as one sample. If samples are different sizes – particularly if, as often happens in corpus-building, some are hundreds or thousands of times as long as others – this is problematic and figures are hard to interpret. But if all samples are the same length, there are no such complications and it is a straightforward response to the whelks problem.

For our dictionary, frequencies will always be numbers of samples the word occurs in.

In homage to Brown, the fixed length for our samples is 2000 words. We first truncated any very long texts at 40,000 words, so we did not have too many samples from any single text, and then simply concatenated all the text of each genre and cut it into 2000-word slices. We considered more sophisticated strategies which paid heed to the beginnings and ends of texts, perhaps only taking one sample from each text (but then, for fiction, we would not have many samples) or not using short fragments (but most of the spoken material was in short fragments). We doubt that our crude strategy will have diminished the integrity of the resulting lists, though of course this is an empirical question.<sup>2</sup>

### The six lists

A multi-part list, as we propose, is a more complex object than a single list. The question, “what is frequent enough to include?” is no longer straightforward. We shall have six main lists, as follows.

### Core Vocab

The ‘core vocab’ list is simplest. A word is in this list if it is in over  $x\%$  of samples in each genre. Table 1 gives the number of words in over  $x\%$  of samples, for various values of  $x$ .

$x$	90	50	30	10	5	4	3
#	37	106	182	471	844	1025	1334

Table 1: # words in over  $x\%$  of samples

We used the 10% mark to identify core-vocab, giving us 471 core-vocab words. These words were then set aside and do not feature in any other lists.

### The ‘genre’ lists

The base method is, for each genre

- list the words according to frequency
- include the top items

The complication is that some words will occur in two, three or four of the lists generated in this way,

and for such cases we have to decide whether they go in:

- just one list
- more than one list
- the general list.

Our strategy is to say there should be some cases of each. To make these decisions we start from a table with the frequency (expressed as percentage of samples that it occurs in) of each word for each genre. The list is for the 5506 noncore words where the sum of these four frequencies is over 5. Then:

- if highest frequency is at least double the next highest, list in that genre only
- else: if two are high and two are low, that is, the first- and second-highest, and both more than double the third and fourth, list in both the top two
- else list in general.

We acknowledge that this is more complex than one might have hoped.

Applying this algorithm gave us the following counts:

	This genre only	This genre and one other	Total
Conversatn	81	131	212
Newsp	1220	726	1946
Fiction	882	319	1201
Web	74	496	570

Table 2: # words to go in each of the genre-lists.

We note the familiar finding that written texts use more different words than spoken (so a larger proportion of tokens in spoken material will comprise core words) and that the web is a mixture, sharing some characteristics of conversation but also sharing vocabulary with news and fiction.

### The general list

This leaves 2413 words for the general list.

At time of writing these lists have just been prepared. They will now be examined and reviewed prior to being fixed for the preparation of the dictionary.

## 5 Conclusion

The question, “what genres should I include in my dictionary” is always a delicate one, and the more corpus-based we are, the more directly we must address it. For a new Dutch frequency dictionary, we are addressing it by basing the headword selection not on one corpus list, as is normal, but on four, for four main genres. This raises a number of

<sup>2</sup> The full paper will consider further implications and also the relation to the Average Reduced Frequency measure.

questions such as “what is the threshold for a word being ‘core’ and what is the threshold for it being specialist’, and ‘under what circumstances (if any) should a word feature in more than one genre list?’ We have given our tentative answers.

## **Reference**

Gries, Stefan. 2008. Dispersions and adjusted frequencies in corpora. *Int Jnl Corpus Linguistics* 13 (4), 403-437.