

Comparing Corpora using Frequency Profiling

Paul RAYSON

Computing Department,
Lancaster University
Lancaster, UK,
paul@comp.lancs.ac.uk

Roger GARSIDE

Computing Department,
Lancaster University
Lancaster, UK,
rgg@comp.lancs.ac.uk

Abstract

This paper describes a method of comparing corpora which uses frequency profiling. The method can be used to discover key words in the corpora which differentiate one corpus from another. Using annotated corpora, it can be applied to discover key grammatical or word-sense categories. This can be used as a quick way in to find the differences between the corpora and is shown to have applications in the study of social differentiation in the use of English vocabulary, profiling of learner English and document analysis in the software engineering process.

1 Introduction

Corpus-based techniques have increasingly been used to compare language usage in recent years. One of the largest early studies was the comparison of one million words of American English (the Brown corpus) with one million words of British English (the LOB corpus) by Hofland and Johansson (1982). A difference coefficient defined by Yule (1944) showed the relative frequency of a word in the two corpora. A statistical goodness-of-fit test, the Chi-squared test, was also used to compare word frequencies across the two corpora. They noted any resulting chi-squared values which indicated that a statistically significant difference at the 5%, 1%, or 0.1% level had been detected between the frequency of a word in American English and in British English. The null hypothesis of the test is that there is no difference between the observed frequencies.

More recently, this size of corpus comparison is becoming the standard even for postgraduate

studies with the increasing availability of corpora and reasoning that one million words gives sufficient evidence for higher frequency words. However, with the production of large corpora such as the British National Corpus (BNC) containing one hundred million words (Aston & Burnard, 1998), frequency comparisons are available across millions of words of text. There are two main types of corpus comparison:

- comparison of a sample corpus to a large(r) corpus
- comparison of two (roughly-) equal sized corpora

In the first type, we refer to the large(r) corpus as a 'normative' corpus since it provides a text norm (or standard) against which we can compare. These two main types of comparison can be extended to the comparison of more than two corpora. For example, we may compare one normative corpus to several smaller corpora at the same time, or compare three or more equal sized corpora to each other. In general, however, this makes the results more difficult to interpret.

There are also a number of issues which need to be considered when comparing two (or more) corpora:

- representativeness
- homogeneity within the corpora
- comparability of the corpora
- reliability of statistical tests (for different sized corpora and other factors)

Representativeness (Biber, 1993) is a particularly important attribute for a normative corpus when comparing a sample corpus to a large normative corpus (such as the BNC) which contains sections from many different text types

and domains. To be representative a corpus should contain samples of all major text types (Leech, 1993) and if possible in some way proportional to their usage in 'every day language' (Clear, 1992). This first type of comparison is intended to discover features in the sample corpus with significantly different usage (i.e. frequency) to that found in 'general' language.

The second type of comparison is one that views corpora as equals (as in the Brown and LOB comparison). It aims to discover features in the corpora that distinguish one from another. Homogeneity within each of the corpora is important here since we may find that the results reflect sections within one of the corpora which are unlike other sections in either of the corpora under consideration (Kilgarriff 1997). Comparability is of interest too, since the corpora should have been sampled for in the same way. In other words, the corpora should have been built using the same stratified sampling method and with, if possible, randomised methods of sample selection. This is the case with Brown and LOB, since LOB was designed to be comparable to the Brown corpus.

The final issue, which has been addressed elsewhere, is the one regarding the reliability of the statistical tests in relation to the size of the corpora under consideration. Kilgarriff (1996) points out that in the Brown versus LOB comparison many common words are marked as having significant chi-squared values, and that because words are not selected at random in language we will always see a large number of differences in two such text collections. He selects the Mann-Whitney test that uses ranks of frequency data rather than the frequency values themselves to compute the statistic. However, he observes that even with the new test 60% of words are marked as significant. Ignoring the actual frequency of occurrence as in the Mann-Whitney test discards most of the evidence we have about the distribution of words. The test is often used when comparing ordinal rating scales (Oakes 1998: 17).

Dunning (1993) reports that we should not rely on the assumption of a normal distribution when performing statistical text analysis and suggests

that parametric analysis based on the binomial or multinomial distributions is a better alternative for smaller texts. The chi-squared value becomes unreliable when the expected frequency is less than 5 and possibly overestimates with high frequency words and when comparing a relatively small corpus to a much larger one. He proposes the log-likelihood ratio as an alternative to Pearson's chi-squared test. For this reason, we chose to use the log-likelihood ratio in our work as described in the next section. In fact, Cressie and Read (1984) show that Pearson's X^2 (chi-squared) and the likelihood ratio G^2 (Dunning's log-likelihood) are two statistics in a continuum defined by the power-divergence family of statistics. They go on to describe this family in later work (1988, 1989) where they also make reference to the long and continuing discussion of the normal and chi-squared approximations for X^2 and G^2 .

We have applied the goodness-of-fit test for comparison of linguistically annotated corpora. The frequency distributions of part-of-speech and semantic tags are sharply different to words. In these comparisons, we are unlikely to observe rare events such as tags occurring once. However, much higher frequencies will occur and so the log-likelihood test is less likely to overestimate significance in these cases.

2 Methodology

The method is fairly simple and straightforward to apply. Given two corpora we wish to compare, we produce a frequency list for each corpus. Normally, this would be a word frequency list, but as described above and as with examples in the following application section, it can be a part-of-speech (POS) or semantic tag frequency list. However, let us assume for now that we are performing a comparison at the word level¹. For each word in the two frequency lists we calculate the log-likelihood (henceforth LL) statistic. This is performed by constructing a contingency table as in Table 1.

¹ The application of this technique to POS or semantic tag frequency lists is achieved by constructing the contingency table with tag rather than word frequencies.

Table 1 Contingency table for word frequencies

	CORPUS ONE	CORPUS TWO	TOTAL
Freq of word	a	b	a+b
Freq of other words	c-a	d-b	c+d-a-b
TOTAL	c	d	c+d

Note that the value ‘c’ corresponds to the number of words in corpus one, and ‘d’ corresponds to the number of words in corpus two (N values). The values ‘a’ and ‘b’ are called the observed values (O). We need to calculate the expected values (E) according to the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

In our case $N_1 = c$, and $N_2 = d$. So, for this word, $E_1 = c*(a+b) / (c+d)$ and $E_2 = d*(a+b) / (c+d)$. The calculation for the expected values takes account of the size of the two corpora, so we do not need to normalise the figures before applying the formula. We can then calculate the log-likelihood value according to this formula:

$$-2 \ln \lambda = 2 \sum_i O_i \ln \left(\frac{O_i}{E_i} \right)$$

This equates to calculating LL as follows:
 $LL = 2*((a*\log(a/E_1)) + (b*\log(b/E_2)))$

The word frequency list is then sorted by the resulting LL values. This gives the effect of placing the largest LL value at the top of the list representing the word which has the most significant relative frequency difference between the two corpora. In this way, we can see the words most indicative (or characteristic) of one corpus, as compared to the other corpus, at the top of the list. The words which appear with roughly similar relative frequencies in the two corpora appear lower down the list. Note that we do not use the hypothesis-test by comparing the LL values to a chi-squared distribution table. As Kilgarriff & Rose (1998) note, even Pearson’s X^2 is suitable without the ‘hypothesis-testing

link’. Given the non-random nature of words in a text, we are always likely to find frequencies of words which differ across any two texts, and the higher the frequencies, the more information the statistical test has to work with. Hence, it is at this point that the researcher must intervene and qualitatively examine examples of the significant words highlighted by this technique. We are not proposing a completely automated approach.

3 Applications

This method has already been applied to study social differentiation in the use of English vocabulary and profiling of learner English. In Rayson et al (1997), selective quantitative analyses of the demographically sampled spoken English component of the BNC were carried out. This is a subcorpus of circa 4.5 million words, in which speakers and respondents are identified by such factors as gender, age, social group and geographical region. Using the method, a comparison was performed of the vocabulary of speakers, highlighting those differences which are marked by a very high value of significant difference between different sectors of the corpus according to gender, age and social group.

In Granger and Rayson (1998), two similar-sized corpora of native and non-native writing were compared at the lexical level. The corpora were analysed by a part-of-speech tagger, and this permitted a comparison at the major word-class level. The patterns of significant overuse and underuse for POS categories demonstrated that the learner data displayed many of the stylistic features of spoken rather than written English.

The same technique has more recently been applied to compare corpora analysed at the semantic level in a systems engineering domain and this is the main focus of this section. The motivation for this work is that despite natural language’s well-documented shortcomings as a medium for precise technical description, its use in software-intensive systems engineering remains inescapable. This poses many problems for engineers who must derive problem understanding and synthesise precise solution descriptions from free text. This is true both for

the largely unstructured textual descriptions from which system requirements are derived, and for more formal documents, such as standards, which impose requirements on system development processes. We describe an experiment that has been carried out in the REVERE project (Rayson et al, 2000) to investigate the use of probabilistic natural language processing techniques to provide systems engineering support.

The target documents are field reports of a series of ethnographic studies at an air traffic control (ATC) centre. This formed part of a study of ATC as an example of a system that supports collaborative user tasks (Bentley et al, 1992). The documents consist of both the verbatim transcripts of the ethnographer's observations and interviews with controllers, and of reports compiled by the ethnographer for later analysis by a multi-disciplinary team of social scientists and systems engineers. The field reports form an interesting study because they exhibit many characteristics typical of documents seen by a systems engineer. The volume of the information is fairly high (103 pages) and the documents are not structured in a way designed to help the extraction of requirements (say around business processes or system architecture).

The text is analysed by a part-of-speech tagger, CLAWS (Garside and Smith, 1997), and a semantic analyser (Rayson and Wilson, 1996) which assigns semantic tags that represent the semantic field (word-sense) of words from a lexicon of single words and an idiom list of multi-word combinations (e.g. 'as a rule'). These resources contain approximately 52,000 words and idioms.

The normative corpus that we used was a 2.3 million-word subset of the BNC derived from the transcripts of spoken English. Using this corpus, the most over-represented semantic categories in the ATC field reports are shown in Table 2. The log-likelihood test is applied as described in the previous section and represents the semantic tag's frequency deviation from the normative corpus. The higher the figure, the greater the deviation.

Table 2. Over-represented categories in ATC field reports

Log-likelihood	Tag	Word sense (examples from the text)
3366	S7.1	power, organising ('controller', 'chief')
2578	M5	flying ('plane', 'flight', 'airport')
988	O2	general objects ('strip', 'holder', 'rack')
643	O3	electrical equipment ('radar', 'blip')
535	Y1	science and technology ('PH')
449	W3	geographical terms ('Pole Hill', 'Dish Sea')
432	Q1.2	paper documents and writing ('writing', 'written', 'notes')
372	N3.7	measurement ('length', 'height', 'distance', 'levels', '1000ft')
318	L1	life and living things ('live')
310	A10	indicating actions ('pointing', 'indicating', 'display')
306	X4.2	mental objects ('systems', 'approach', 'mode', 'tactical', 'procedure')
290	A4.1	kinds, groups ('sector', 'sectors')

With the exception of Y1 (an anomaly caused by an interviewee's initials being mistaken for the PH unit of acidity), all of these semantic categories include important objects, roles, functions, etc. in the ATC domain. The frequency with which some of these occur, such as M5 (flying), are unsurprising. Others are more revealing about the domain of ATC. Figure 1 shows some of the occurrences of the semantic category O2 (general objects). The important information extracted here is the importance of 'strips' (formally, 'flight strips'). These are small pieces of cardboard with printed flight details that are the most fundamental artefact used by the air traffic controllers to manage their air space. Examination of other words in this category also shows that flight

```

to write '1260L' in red on a strip , whilst at the same time instru
he Isle of Man ... " This strip was towards ' the bottom of one
cated by the beacon printed in box ' B ' of the strip ( second lef
on printed in box ' B ' of the strip ( second left ) Strips seemed br
arrival time over that beacon ( box ' A ' ) This was obviously only
viously only approximate- some strips were out of position , and I got
al line near the call sign on a strip to indicate an unusual speed . <
med much busier . There were 16 strips in one of his racks . <BR> A ;
rewere 16 strips in one of his racks . <BR> A ; ' <BR> c&lt;Tide &gt
sy , that talking and using an input device might also be , but that the pr
hat talking and using an input device might also be , but that the pr
: " the nice thing about strips is their flexibility . " &quot; a

```

Figure 1. Browsing the semantic category O2

strips are held in 'racks' to organise them according to (for example) aircraft time-of-arrival.

Similarly, browsing the context for Q1.2 (paper documents and writing) would allow us to discover that controllers annotate flight strips to record deviations from flight plans, and L1 (life, living things) would reveal that some strips are 'live', that is, they refer to aircraft currently traversing the controller's sector. Notice also that the semantic categories' deviation from the normative corpus can also be expected to reveal domain roles (actors). In this example, the frequency of S7.1 (power, organising) shows the importance of the roles of 'controllers' and 'chiefs'.

Using the frequency profiling method does not automate the task of identifying abstractions, much less does it produce fully formed requirements that can be pasted into a specification document. Instead, it helps the engineer quickly isolate potentially significant domain abstractions that require closer analysis.

4 Conclusions

This paper has described a method of comparing corpora which uses frequency profiling. The method has been shown to discover key items in the corpora which differentiate one corpus from another. It has been applied at the word level, part-of-speech tag level, and semantic tag level. It can be used as a quick way in to find the differences between the corpora and is shown to have applications in the study of social differentiation in the use of English vocabulary, profiling of learner English and document analysis in the software engineering process.

Future directions in which we aim to research include a more precise specification of the

reliability of the statistical tests (LL, Pearson's X^2 and others) under the effects of corpus size, ratio of the corpora being compared and word (or tag) frequency.

We do not propose a completely automated approach. The tools suggest a group of key items by decreasing order of significance which distinguish one corpus from another. It is then that the researcher should investigate occurrences of the significant items in the corpora using standard corpus techniques such as KWIC (key-word in context). The reasons behind their significance can be discovered and explanations sought for the patterns displayed. By this process, we can compare the corpora under investigation and make hypotheses about the language use they represent.

Acknowledgements

Our thanks go to Geoffrey Leech and the anonymous reviewers who commented on earlier versions of this paper. The REVERE project is supported under the EPSRC Systems Engineering for Business Process Change (SEBPC) programme, project number GR/MO4846.

References

- Aston, G. and Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press.
- Bentley R., Rodden T., Sawyer P., Sommerville I, Hughes J., Randall D., Shapiro D. (1992). *Ethnographically-informed systems design for air traffic control*, In Proceedings of Computer-Supported Cooperative Work (CSCW) '92, Toronto, November 1992.
- Biber, D. (1993). *Representativeness in Corpus Design*. *Literary and Linguistic Computing*, 8, Issue 4, Oxford University Press, pp. 243-257.

- Clear, J. (1992). *Corpus sampling*. In G. Leitner (ed.) *New directions in English language corpora*. Mouton-de-Gruyter, Berlin, pp. 21 - 31.
- Cressie, N. and Read, T. R. C. (1984) *Multinomial Goodness-of-Fit Tests*. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 46, No. 3, pp. 440 - 464.
- Cressie, N. and Read, T. R. C. (1989). *Pearson's X^2 and the Loglikelihood Ratio Statistic G^2 : A comparative review*. *International Statistical Review*, 57, 1, Belfast University Press, N.I., pp. 19-43.
- Dunning, T. (1993). *Accurate Methods for the Statistics of Surprise and Coincidence*. *Computational Linguistics*, 19, 1, March 1993, pp. 61-74.
- Garside, R. and Smith, N. (1997). *A Hybrid Grammatical Tagger: CLAWS4*, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Longman, London.
- Granger, S. and Rayson, P. (1998). *Automatic profiling of learner texts*. In S. Granger (ed.) *Learner English on Computer*. Longman, London and New York, pp. 119-131.
- Hofland, K. and Johansson, S. (1982). *Word frequencies in British and American English*. The Norwegian Computing Centre for the Humanities, Bergen, Norway.
- Kilgarriff, A. (1996) *Why chi-square doesn't work, and an improved LOB-Brown comparison*. ALLC-ACH Conference, June 1996, Bergen, Norway.
- Kilgarriff, A. (1997). *Using word frequency lists to measure corpus homogeneity and similarity between corpora*. *Proceedings 5th ACL workshop on very large corpora*. Beijing and Hong Kong.
- Kilgarriff, A. and Rose, T. (1998). *Measures for corpus similarity and homogeneity*. In proceedings of the 3rd conference on Empirical Methods in Natural Language Processing, Granada, Spain, pp. 46 - 52.
- Leech, G. (1993). *100 million words of English: a description of the background, nature and prospects of the British National Corpus project*. *English Today* 33, Vol. 9, No. 1, Cambridge University Press.
- Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh.
- Rayson, P., and Wilson, A. (1996). *The ACAMRIT semantic tagging system: progress report*, In L. J. Evett, and T. G. Rose (eds.) *Language Engineering for Document Analysis and Recognition, LEDAR, AISB96 Workshop proceedings*, pp 13-20. Brighton, England.
- Rayson, P., Leech, G., and Hodges, M. (1997). *Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus*. *International Journal of Corpus Linguistics*. 2 (1). pp. 133 - 152. John Benjamins, Amsterdam/Philadelphia.
- Rayson, P., Garside, R., and Sawyer, P. (2000). *Assisting requirements engineering with semantic document analysis*. In *Proceedings of RIAO 2000 (Recherche d'Informations Assistie par Ordinateur, Computer-Assisted Information Retrieval) International Conference, Collège de France, Paris, France, April 12-14, 2000*. C.I.D., Paris, pp. 1363 - 1371.
- Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. Springer series in statistics. Springer-Verlag, New York.
- Yule, G. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press.