WORD FREQUENCY AND KEYWORD EXTRACTION
AHRC ICT Methods Network Expert Seminar on Linguistics
8 September 2006, Lancaster University, UK

In Search of a Bad Reference Corpus

Mike Scott, *University of Liverpool, UK.*

Keywords
keywords, reference corpus, node text, aboutness

> They hunted till darkness came on, but they found
> Not a button, or feather, or mark,
> By which they could tell that they stood on the ground
> Where the Baker had met with the Snark.
>
> In the midst of the word he was trying to say,
> In the midst of his laughter and glee,
> He had softly and suddenly vanished away -- -
> For the Snark *was* a Boojum, you see.

Last two stanzas of *The Hunting of the Snark* by Lewis Carroll

Abstract
What are the tolerable limits of similarity between a reference corpus and a node text for the generation of a useful set of keywords? There is of course considerable subjectivity in the notion of usefulness, which will vary according to research goals which cannot in general be predicted with certainty. Nevertheless, the aim here is to explore the ways in which the similarity between reference corpus and node text varies on certain important dimensions, such as size in tokens, similarity of text-type, similarity of historical period, similarity of subject-matter.

This paper starts from the formula proposed by Berber Sardinha (2004: 101-3) which suggests that the larger the reference corpus, the more keywords will be detected, and his formula for predicting the number of keywords produced with a given text and reference corpus. It also considers his recommendation that a reference corpus should be about five times the size of the node text.

Using a series of reference corpora, the paper explores keywords results in relation to specific texts. The aim is to identify not, as one might imagine, the characteristics of the good reference corpus, but the limits defining a poor one, since in many cases, e.g. the analysis of a dead language or a restricted corpus, the chance of accessing a good reference corpus is slim. The study represents work in progress and much further work needs to be done to confirm and develop its preliminary findings.

Introduction

Snark or Boojum?

We are here hunting a Snark. For those who have not read the poem, a Snark is a mysterious creature sought by some people aboard a ship in Lewis Carroll's poem of 1876. It is never very clear

what a Snark is, or a Boojum, or why a Snark might turn out to be a Boojum. The ship is captained by the Bellman.

> He had bought a large map representing the sea,
> Without the least vestige of land:
> And the crew were much pleased when they found it to be
> A map they could all understand.
>
> "What's the good of Mercator's North Poles and Equators,
> Tropics, Zones, and Meridian Lines?"
> So the Bellman would cry: and the crew would reply
> "They are merely conventional signs!"

Likewise here it is not clear what a really bad reference corpus is and what may happen if we should meet up with one. We are dealing with conventional signs too, and our principles of navigation are also somewhat unclear.

There may be a variety of methods for identifying keywords in texts, methods which rely on word frequency alone, excluding function words via a stop list, or on human identification, ones which access a previously identified semantic word-bank, or ones which rely on a combination of these. The procedure for identifying keywords under discussion in the present paper is, however, that devised for use in WordSmith Tools (Scott 1996 with numerous subsequent versions) which essentially compares a wordlist based on the text in question and a wordlist based on a reference corpus. The idea is quite simple: by comparing the frequency of each item in turn with a known reference, one may identify those items which occur unusually frequently. This is done without any attempt to identify or match up the semantics or pragmatics, and is based on a simple verbatim comparison, without even necessarily lumping lemma variants together.

It is important to stress that no claim should be made that a set of keywords thus identified a. will match a set of human-generated keywords, b. is significant *as a set* even if each individual comparison reaches statistical significance. The main utility of the procedure has instead been in identifying items which are likely to be of linguistic interest in terms of the text's aboutness and structuring, and which can be expected to repay further study e.g. through concordancing to investigate collocation etc.[1]

The method will clearly achieve results which are largely dependent on the qualities of the reference corpus itself.

An analogy will help to make this clear. Suppose one wishes to identify and evaluate the qualities of a given automobile. If it is compared with all existing cars of the same category, such as family sedans, comparative features such as price, safety, speed and comfort will be used. The whole set of family sedan cars made by the world's auto manufacturers will probably be used as the 'reference corpus'. The mere facts that the motor is made of an alloy, or that the tyres are made of rubber, or that the engine burns fuel are not relevant to the comparison, if all such cars burn fuel, have alloy engines and rubber tyres. The amount of fuel consumed would come into the comparison, but not the fact that fuel is burned. But it would be possible to compare the same family saloon car instead to all means of transport available for a given purpose, e.g. getting to Barcelona for a family holiday. In that case, the comparison will involve different criteria, such as convenience, expense, opportunity to take and bring back luggage, impact on the environment, etc. The reference corpus is now the set of {car, train, ferry, taxi, plane etc.}. For different research purposes different reference comparisons are needed.

In general, then, claims can be made a. that the choice of reference corpus will affect the results, b. that features (such as rubber tyres) which are similar in the reference corpus and the node text itself will not surface in the comparison, but c. only features where there is significant departure from the reference corpus norm will become prominent for inspection.

The question then raises its head: how much difference, in the case of words and text, does it make if a somewhat imperfect reference corpus is used? In the real world, it might be hard to obtain a large, perfectly-matched reference for some comparisons. For example, the BNC might be considered a good reference for texts in English, but despite its numerous positive features and the

enormous effort that went into constructing it, it is still only based on 100 million words – any search of the Internet suggests that the amount of text in English on the Internet far exceeds this – and on a sampling procedure which gives about ten times as much weight to written than spoken English. In the case of the analysis of Sumerian or other dead languages, there is only one body of texts to be used. Suppose one wished to compare a text in Sumerian which belonged to the genre poetry, one might find that a reference corpus of Sumerian poetry was extremely slim indeed and question whether the results would be useful.

Size is not the only condition. Another criterion is date. If one were comparing a corpus of seventeenth-century sermons in English, would it be acceptable to use the 1990s BNC as a reference corpus? Or would that constitute a bad reference corpus? Hence the title of this paper. Furthermore, what the texts in a reference corpus are about is presumably critical, unless we use so many texts that what they are each about gets drowned out in the whole. A text about scoring goals in a football match will resonate with lots of others in newspaper texts, but one about the culture and customs of a small tribe may not.

A further consideration is whether one is comparing a single node text with a reference, or a whole set of texts (e.g. comprising a sub-genre) with the reference corpus. For most of this paper we shall be considering only the comparison of one text at a time with a reference corpus.

Berber Sardinha's Formula

Berber Sardinha (2004) discusses the tendency of a reference corpus which is similar to the node text to 'filter out' genre features common to both, and thereby derives a suggestion that a reference corpus which contains several different genres of text in it would be the non-marked choice. In general, he claims that 'critical reference corpus sizes are 2,3 and 5 times that of the node text' (2004:102) and presents a formula for calculating the number of keywords likely to be obtained when comparing two corpora.

Figure 1. Berber Sardinha's formula

KWs = 249.837059 - 0.00002734 * ref corpus tokens + 0.00886787 * ref corpus types + 0.00137131 * node corpus tokens

(Berber Sardinha, 2004: 102)

Thus if we have a reference corpus of 100 million tokens and 400,000 types, and a node corpus of 5,000 tokens, we should get 1,070 keywords. The formula works by computing a regression line. It works best with relatively small corpora, up to about 5 million running words (Berber Sardinha, personal communication)

The formula may be useful for predicting the number of keywords which can be expected to be found using a given reference corpus – but it will not tell us what sorts of keywords are likely to be generated, which is why the exploratory study described below was carried out.

Study

The present investigation was thus designed to study the influence on the keywords which would be generated using one and the same routine and settings, but varying the number and kinds of texts comprising the reference corpus (RC). The study was carried out in three stages.
The research questions were:

1. What distribution is obtained if a set of keyword calculations is made using RCs of different sizes, using randomly selected BNC texts regardless of genre? For example as the size of the RC increases does the quality of the keyword results increase? If so, is there any noticeable threshold below which the quality is unacceptable?
2. What sort of keyword results obtain if a deliberately strange RC is used, one which has little or no relation to the text in question apart from being based on the same language? Is the quality of results (un)acceptable?

3. What quality of keyword results obtains if genre is included as a variable, so that BNC texts are compared by genre with the source texts?

Materials and Methods

The software for all three research questions was WordSmith Tools version 4.0 (Scott, 2004). Two source texts were used as a sample for comparison with the various RCs. These were BNC text A6L, a book profile of leaders of commerce, about 46,000 words in length, and text KNG, a spoken text of only 615 words, between a doctor and a patient. Fragments of these are reproduced without BNC tags in the Appendix.

For research question 1 above, RC texts were chosen from the 4,054 BNC texts (spoken plus written) using a randomising function, so that 22 different sizes of RC were selected, comprising the following numbers of texts, without consideration of genre (i.e. mixed genres):

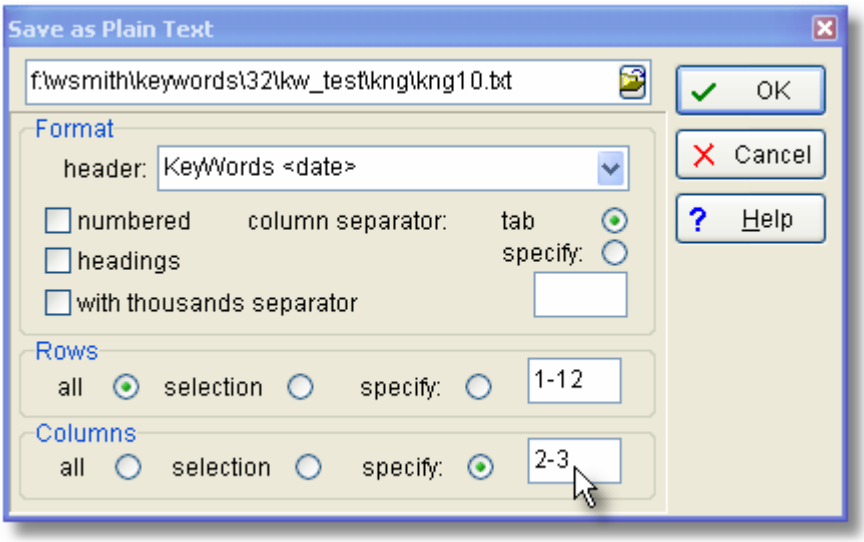
Table 1. Numbers of texts in each RC

10	50	250	2000
15	60	300	2500
20	75	400	3000
25	100	500	4000
30	150	1000	
40	200	1500	

Two source texts were then compared with these different RCs. These were BNC text A6L, a book profile of leaders of commerce, about 46,000 words in length, and text KNG, a spoken text of only 615 words, between a doctor and a patient.

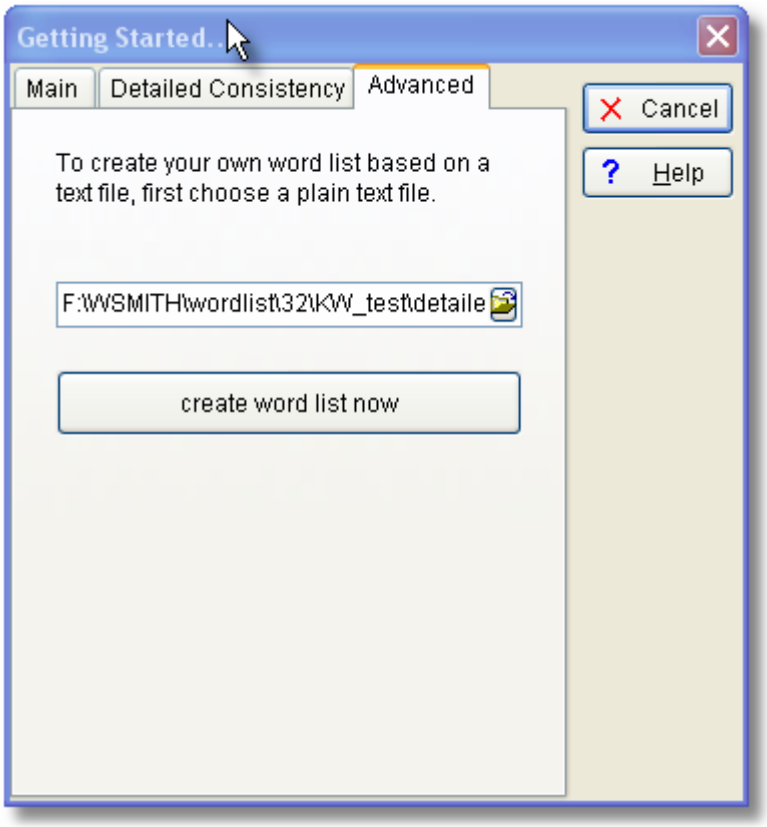
The KeyWords tool settings were as follows: minimum frequency = 3; maximum keywords = 5000; negative keywords to be excluded; p value = 0.000001; procedure = log likelihood (Dunning 1993). Keywords were computed for the two source texts using each of the 22 reference corpora. Keywords and frequency information for each were saved as text:

Figure 2. Saving keywords as text



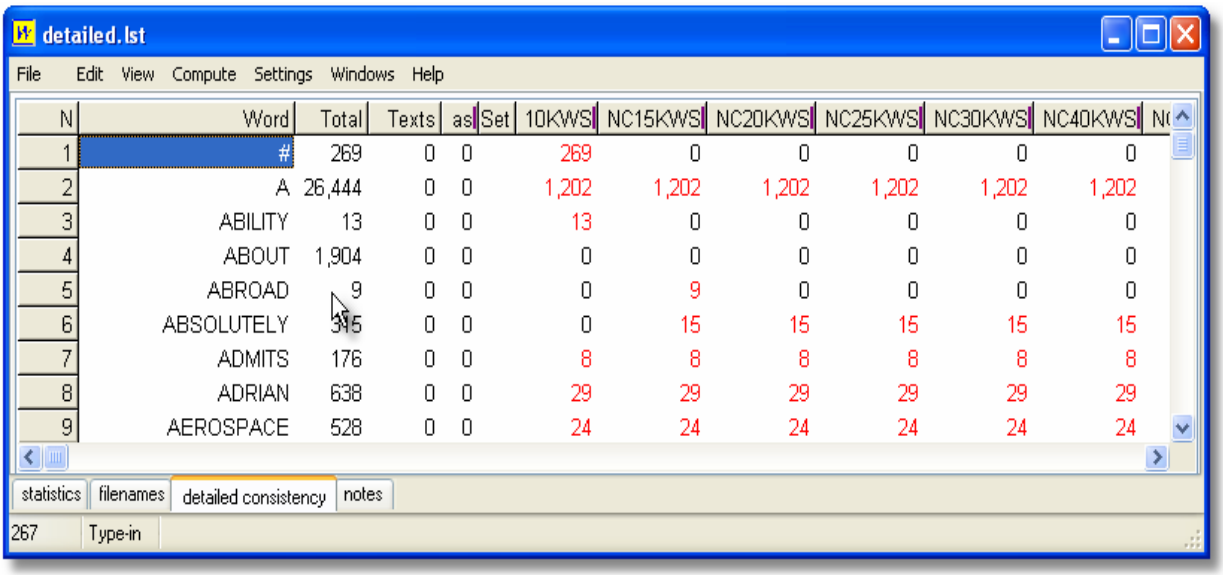
These were then imported into the WordList tool

Figure 3. Importing a word-list from plain text



In this way, it was possible to treat each set of keywords as a word-list and examine which of the items were found in the different sets based on the 22 RCs, using WordList's detailed consistency procedure:

Figure 4. Detailed Consistency view of the 22 keyword sets



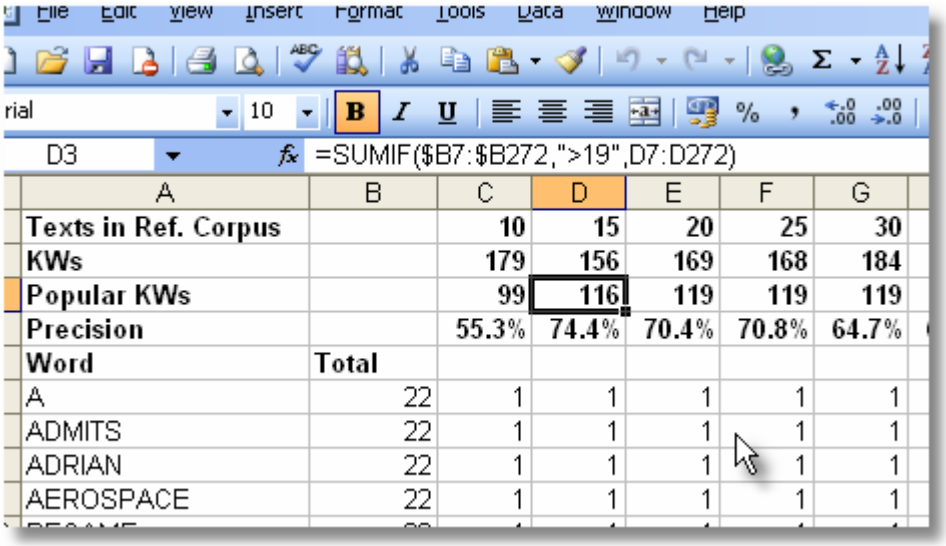
N	Word	Total	Texts	as	Set	10KWS	NC15KWS	NC20KWS	NC25KWS	NC30KWS	NC40KWS
1	#	269	0	0		269	0	0	0	0	0
2	A	26,444	0	0		1,202	1,202	1,202	1,202	1,202	1,202
3	ABILITY	13	0	0		13	0	0	0	0	0
4	ABOUT	1,904	0	0		0	0	0	0	0	0
5	ABROAD	9	0	0		0	9	0	0	0	0
6	ABSOLUTELY	315	0	0		0	15	15	15	15	15
7	ADMITS	176	0	0		8	8	8	8	8	8
8	ADRIAN	638	0	0		29	29	29	29	29	29
9	AEROSPACE	528	0	0		24	24	24	24	24	24

This detailed consistency procedure allows one to sort the keywords in a number of ways (in the figure above they are sorted alphabetically) e.g. according to whether they are found to be key in the various RC. Item frequencies are constant where greater than 0 in the figure above, e.g. *absolutely* was found to be key, and had a source text frequency of 15, in all comparisons except that with the smallest RC, where it was not picked up as key.

Overall, there were 267 keyword types in the 22 lists from text A6L, and 18 keyword types from text KNG.

These results were exported into MS Word™ so that the numbers could be simplified, in such a way that zeroes were replaced with space and numbers standardised as ones, and these results brought into MS Excel™.

Figure 5. Excel spreadsheet of results



	A	B	C	D	E	F	G
	Texts in Ref. Corpus		**10**	**15**	**20**	**25**	**30**
	KWs		**179**	**156**	**169**	**168**	**184**
	Popular KWs		**99**	**116**	**119**	**119**	**119**
	Precision		**55.3%**	**74.4%**	**70.4%**	**70.8%**	**64.7%**
	Word	**Total**					
	A	22	1	1	1	1	1
	ADMITS	22	1	1	1	1	1
	ADRIAN	22	1	1	1	1	1
	AEROSPACE	22	1	1	1	1	1

This figure shows a fragment of the Excel data: for the smallest RC (based on 10 texts only), 179 keywords were identified in relation to text A6L, but 156 when using the 15-text RC, rising to 184 with the 30-text RC.

Two further variables were computed in Excel: popularity and precision. Popularity was defined (as can be seen in the figure above) as presence of each keyword in at least 20 of the 22 sets. Thus of the 179 keywords found using the smallest RC, 99 were common to at least 20 of the 22 sets. This was based on the rationale that the keywords identified using most of the RC sets are more likely to be useful than those identified in a minority, and is the only indicator of quality used in the study. (Other indicators of keyword quality might include informant testing with a variety of informants ranging from the naïve – '12 good men and true' – to the linguistically sophisticated, or for example the original authors.)

Precision was computed following Oakes ('the proportion of retrieved items that are in fact relevant (the number of relevant items obtained divided by the total number of retrieved items)' 1998:176). In this case the calculation involved dividing the total number of keywords (179 for the smallest RC) by the number of popular keywords (99) which gives a precision value of 55% for the smallest RC.

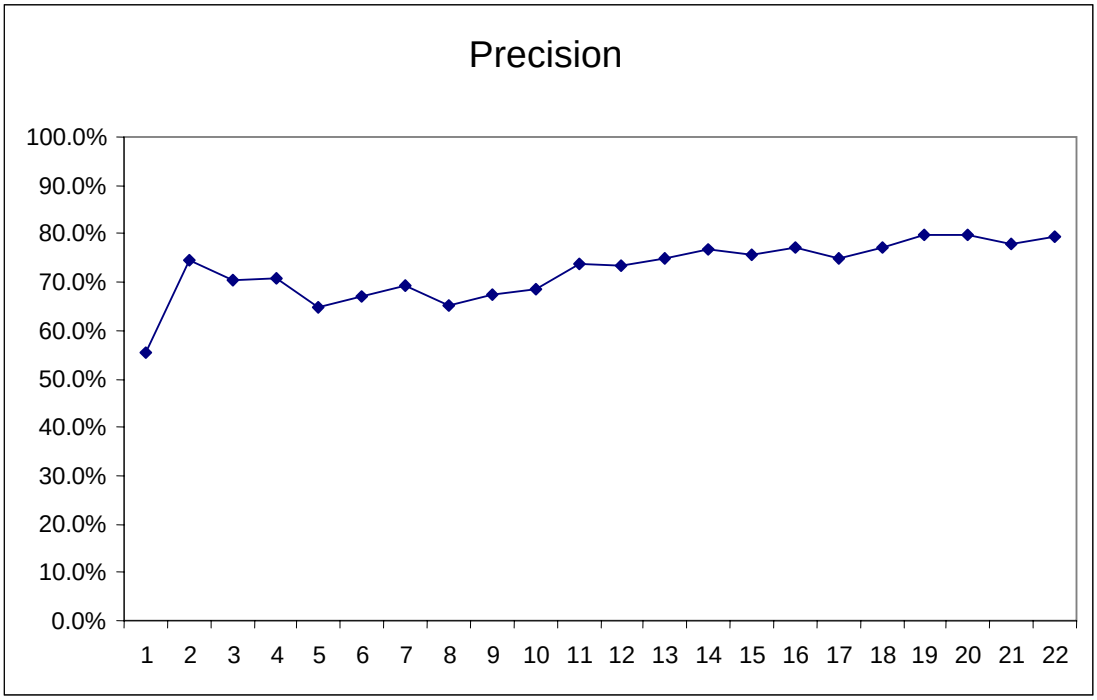
The study uses comparative precision as its main measure.

Results

Research question 1: What distribution is obtained if a set of keyword calculations is made using RCs of different sizes, using randomly selected BNC texts regardless of genre? For example as the size of the RC increases does the quality of the keyword results increase? If so, is there any noticeable threshold below which the quality is unacceptable?

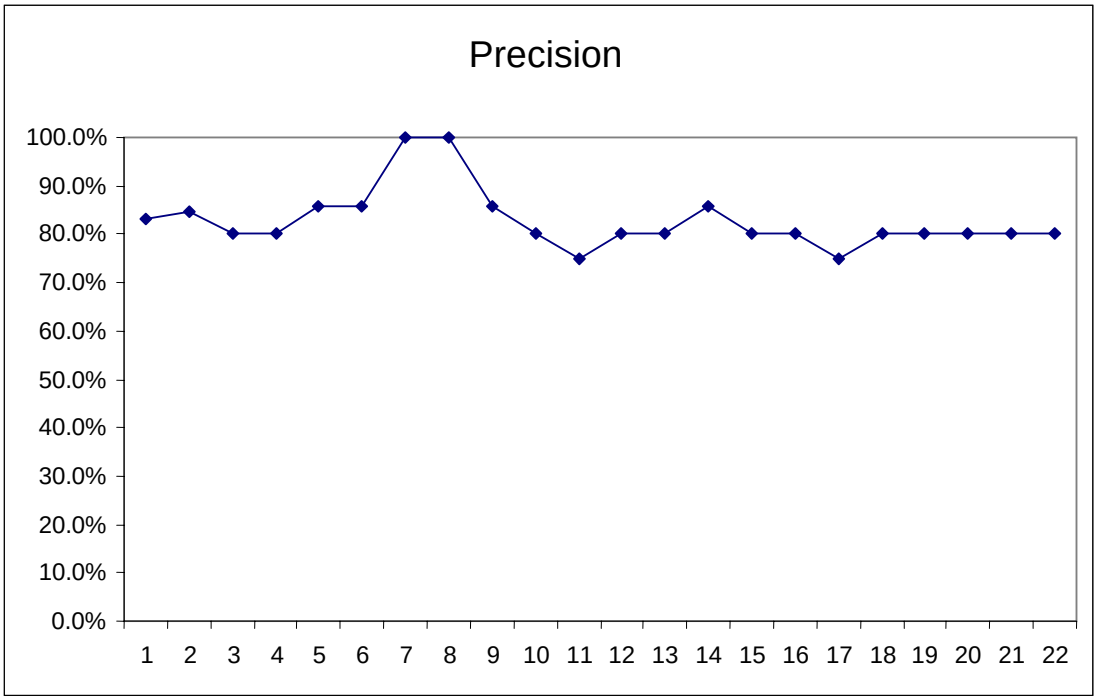
A first set of results for this research question, based on the two texts is shown in the two following figures.

Figure 6. Precision values for text A6L



This figure shows increasing precision values as the size of the RC increases. The 22 different RC sizes are on the horizontal axis and exact RC sizes for each can be found in Table 1. above. The plot suggests that, after a rocky start, where precision is fairly inconsistent but still high at over 50%, the precision gently increases to a maximum value corresponding at or near the biggest RC. The text in question is a lengthy (44 thousand words) section of a book profiling well-known business leaders.

Figure 7. Precision values for text KNG



For this very short (615-word) doctor-patient interview we get rather different results. There were only 18 keywords identified over the 22 lists. Again the precision values are high, all over 75%, but these are clearly higher values than with the much longer text, which generated many more keywords. Here we do not get increasing precision values as the size of the RC increases; instead there seems to be a ceiling effect with fairly small RCs based on 50 or 60 texts.

It seems that an appropriate answer to the first research question is that there is no clear and obvious threshold below which poor keyword results can be expected. Precision values when using a mixed bag of RC texts, even if the set is small, are high; there is no obvious cut-off point; very much the same keywords are generated whatever the RC used. We have not yet found a really bad reference corpus.

Research question 2: What sort of keyword results obtain if a deliberately strange RC is used, one which has little or no relation to the text in question apart from being based on the same language? Is the quality of results (un)acceptable?

For this research question, an RC was used that was based on all of Shakespeare's plays. The genre is drama, the period late sixteenth and early seventeenth Century. Will this absurd RC give rise to usefully poor results?

Using the leaders of commerce text from the 1990s with the Shakespeare RC, there were altogether 606 keywords. Although the source text is lengthy at 44,000 words, 606 distinct keyword types seems a large number. With the BNC RC, there were just over a quarter of that number, 161 keywords. 143 of these were common to both sets. If we assume that the BNC RC is the better RC, at first sight it might seem that using an inappropriate RC may generate a lot of unwanted keywords.

A few keywords picked up by the BNC RC were not in the Shakespeare-generated set: common pronouns or conjunctions (I, we, them,you, when) , high frequency verbs and nouns (finds, go, have, make, own, sir, take, taught, thing) a couple of numerical and time-related words (thousand, never) and two which were clearly unknown to Shakespeare (Sikorsky, jojoba). A further 463 keywords were generated only when using the Shakespeare RC. Those beginning with *o* are presented as a sample:

Table 2. keywords picked up only using Shakespeare RC

objective(s), obviously, of, offered, oil, on, one, only, opened, operate, operations, opportunity (ies), option, organisations, original, other, outside, over, overseas, owned

This table suggests that although many *more* keywords were picked up using our deliberately inappropriate RC, the keywords themselves are not absurd. Most of these words beginning with *o* have to do with business operations and are indicative of the aboutness (Phillips, 1989) of the text. In the case of the doctor-patient consultation, the numbers are much more manageable.

Table 3. Doctor-Patient keywords with 2 RCs

Shakespeare RC	**BNC RC**
YES	YES
THAT'S	THAT'S
RIGHT	DOCTOR
DOCTOR	RIGHT
OH	CRAMP
CAN'T	TEASPOONFUL
CRAMP	QUININE
JUST	I'LL
AHA	TABLETS
ER	NO
MR	HEARD
QUININE	EASES
TABLETS	YOU
TEASPOONFUL	AHA
I'M	
EIGHTY	
REALLY	
OPERATION	
GETTING	
INFIRMARY	

OCTOBER
PHONE
GET
EASES
ANY

Again we get more keywords when using Shakespeare as the RC. It is however difficult to claim that as a set those on the left side are in any way worse than those on the right. We are using a necessarily subjective criterion, but research question 2. (is the quality of results unacceptable?) can now be answered provisionally: no, we still have no really bad RC.

Research question 3: What quality of keyword results obtains if genre is included as a variable, so that BNC texts are compared by genre with the source texts?

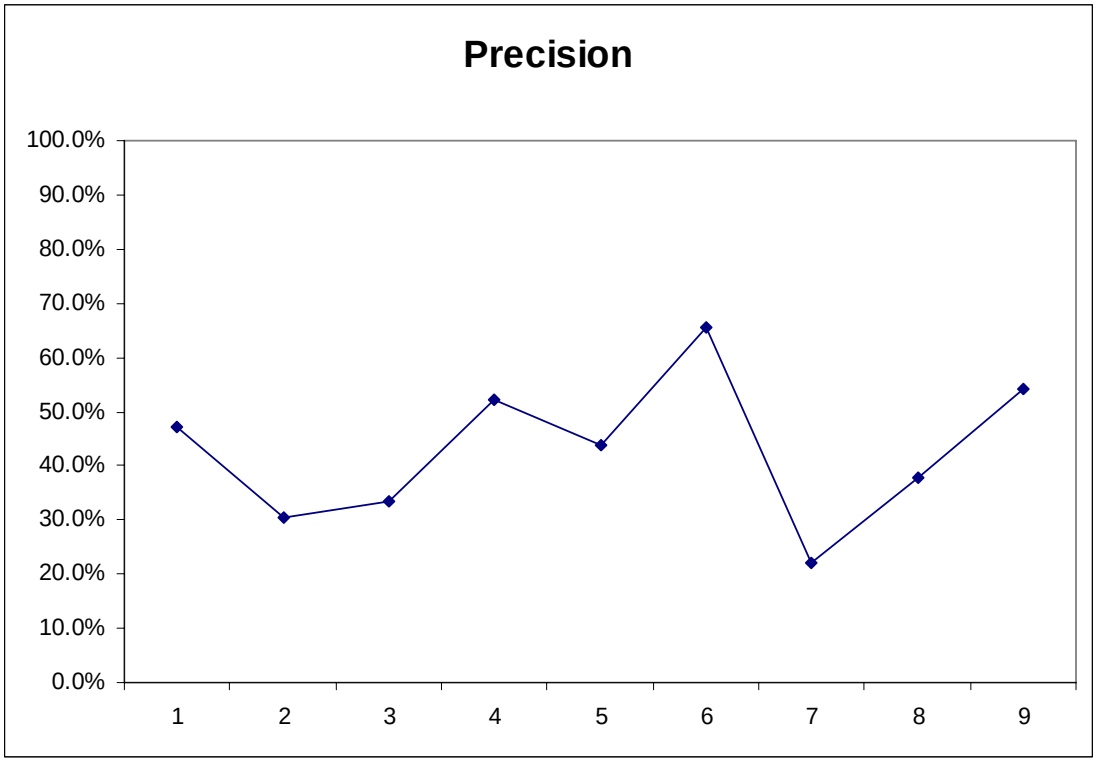
For this part of the study, sets of keywords based on text A6L were again computed using the BNC as the source for the RCs. A series of nine RCs were constructed, using the following sub-sections of the BNC itself, therefore no longer mixed genres. The classification used David Lee's categories embedded in the BNC World Edition headers.

Table 4. Genred RCs

1. commerce texts[2] (112 of them) (A6L is a business text)
2. academic medicine texts (24)
3. prose fiction texts (432)
4. non-academic Humanities texts (111)
5. non-academic politics, law, education texts (93)
6. spoken broadcast discussions (53)
7. conversations (153)
8. oral history interviews (119)
9. spoken meetings (132)

Procedures for computing popularity and precision were as in the first part of the study, except that the criterion for popularity was now based on presence in eight or nine of the keyword sets. The following figure shows precision figures over the nine RCs.

Figure 8. Precision values for text A6L with genred BNC RCs



The horizontal numbers correspond to the numbers in the list in the table above, so that number 7, where there seems to be a big dip, represents 153 conversations. These conversations generated a lot of keywords but a low precision as measured by agreement between these various RCs.

Precision values here are noticeably lower than in the other similar graphs above, between 20% and 70%.The line rises and falls much more steeply.

To interpret this finding, let us remind ourselves that the measure of precision here may be much less appropriate than it was earlier. In the first part, there was a fairly straightforward increase from RC1 to RC22, along a single dimension, the number of BNC texts in each RC. Here, on the other hand, we have several variables in operation at once, the number of texts varying quite unsystematically, and the type of genre also in no particular order except that the spoken ones are the last 4. To expect there to be agreement between these RCs is to assume that they are alike in some way – we could reasonably assume this in the first part but not here. There are two main ways in which we may assume they are not alike: a. because they come from different media and genres, b. because they are about different topics.

It is therefore possible that any of the sets generated might be useful – or useless – for a given purpose.

Research question 3. must therefore remain unanswered for the time being. We do not from this know what the quality of the keywords is, though it does seem that the keywords generated do differ if genre-different RCs are used, much as the keywords differed when the genre Elizabethan drama was used, in comparison with the mixed bag BNC RC. If this is so, different aspects of the source text's aboutness are being picked up.

Conclusions

These three mini-studies have important limitations.

The texts in the study are all incomplete extracts from larger texts apart from the doctor-patient interview and the Shakespeare plays. This is probably not a major limitation in itself, since there is no reason to suppose that the keyword method depends exclusively on the identification of

clear text boundaries; indeed, it is likely that the method can be used successfully with segments of texts, and it certainly has been used to compare groups of texts with an RC.

We have only examined a couple of texts in comparison with our 32 different RCs. keywords have been found by comparing one text at a time with an RC, not by comparing sub-corpora with larger RC corpora. No other method has been used to evaluate the quality of keyword sets apart from agreement between the different RCs and subjective appreciation. Informant studies could also be carried out.

In conclusion it seems that the first part suggested that, using a mixed bag RC, the larger the RC the better – but not in the case of the small doctor-patient consultation: a moderate sized RC may suffice. This suggests that the keyword procedure is fairly robust. The second part suggests that keywords identified even by an obviously absurd RC can be plausible indicators of aboutness, which reinforces the conclusion that keyword analysis is robust. The third part suggested that genre-specific RCs identify rather different keywords, which itself led to the conclusion that the aboutness of a text may not be one thing but numerous different ones.

The Snark is still out there. Somewhere.

References

Aston, G. and Burnard, L., *The BNC Handbook* (Edinburgh: Edinburgh University Press, 1998).

Berber Sardinha, A.P., *Lingüística de Corpus* (Barueri, SP, Brazil: Manole, 2004).

Dunning, T., 'Accurate Methods for the Statistics of Surprise and Coincidence' *Computational Linguistics*, 19:1 (1993), 61–74.

Oakes, M., *Statistics for Corpus Linguistics* (Edinburgh: Edinburgh University Press, 1998).

Phillips, M., 'Lexical Structure of Text', *Discourse Analysis Monographs,* 12, (Birmingham: University of Birmingham,1989).

Scott, M. and Tribble C., *Working with Texts: Keyword and Corpus Analysis in Language Education* (Amsterdam: Benjamins, forthcoming)

Appendices

Fragment of text A6L

Sir Adrian Cadbury
Born: Birmingham, 1929.
Educated: Eton; King's College, Cambridge.
Sir Adrian Cadbury is chairman of Cadbury Schweppes plc. He joined Cadbury Brothers Ltd in 1952 and became chairman in 1956. After the merger between Cadbury and Schweppes he succeeded Lord Watkinson as Chairman of the combined company at the end of 1974. He was knighted for his services to industry in 1977.
He is a director of the Bank of England and of IBM UK Holdings Ltd, and chairman of Pro Ned, an organisation that encourages the appointment of non-executive directors to company boards. He also heads the CBI Business Education Task Force.
He is chancellor of the University of Aston in Birmingham, a trustee of the Bournville Village Trust and president of the Birmingham Chamber of Industry and Commerce. He was made a freeman of the City of Birmingham in 1982.
Sir Adrian Cadbury is not one of those who subscribes to the popular theory that a truly professional manager can take over the helm of any type of business with only a superficial knowledge of the nuts and bolts.
'I'm very sceptical of the ability to shift from managing a bank to managing a steel mill, for example. I have grave doubts about that. I think it is essential to understand the key factors for success or failure in your type of business and I'm not convinced you can do that without actually understanding the process in some detail.'

Text KNG

Hello Doctor.
Good morning Mrs .
Yes.
Well young lady, what can we do
I'm just up to see about this operation.
What's what operation?
You know Royal Infirmary
Aha.
 th that was temporary.
Yes.
And he's the consultant told me, it would take two to three days.
 I think it's years he means.
Have you not heard any more about it?
Two cancellations Doctor, two.
And that's all you've heard?
That's all I've heard.
Cos we've never heard any more.
No.
I just thought I'd come up and speak to you about that.
Yes.
And it's a thingy that I can't forget about.
I can't make any appointments for going
That's right.
anywhere.
That's right.
You know.
Right well I'll get on to them this morning.
Will you?
That was Mr ?
Yes.
Mr was
That's right.
the man.
Yes.
Right.
Yes.
I mean, that's a long time isn't it?
Oh yes.
Yes.
Phone Mr Royal Infirmary Mrs operation.
 As soon as possible.
Now could I have some cramp er tablets Doctor?
Yes.
For my hands any my feet.
And this is where I used to get pain.
Now I can be constipatied constipated and I can be the other way.
Aha.
And Dr that was the only w doctor ever I knew up here.
Yes.
And he always gave me this bottle
That's the syrupy stuff?
Yes, and he told me to take a spoonful
That's
at night a teaspoonful
A teaspoonful before you go to your bed .
Yes.
That's right.
So could I have that?

Yes.
Please.
I I'm not really a doctor person really, but this is really troubling me up here you know .
Oh yes.
Oh aye.
You should heard long before this.
Oh it's a terrible thing Doctor .
That's that's a terrible old thing .
And I'm eighty eighty two, I'll be eighty three in December.
Aha.
And you're not getting any younger.
I'm not getting any younger, but mind you I'd like to get it done.
Yes.
Because I can't take any freedom.
That's right.
That's right.
Now.
Mark this in here.
We were getting the sun weren't we?
Aye, today, today.
Now that was yours Your er
Cramp.
Your erm quinine.
Was it you quinine?
Tab tablets.
Was it
Oh yes.
the quinine tablets?
The old fashioned ones.
For the cramp.
was for this c for the cramp?
Yes yes it is still.
Sometimes I got to get up in the night and walk about and
Mhm.
my hand's cold.
But that oil, it seemed to help me a lot .
Oh yes, it just eases
Just a teaspoonful.
That's right, just eases things through. ?
No I can't do with anything.
No no no no.
Doctor used to say, Never you take a laxative.
No no.
No that's the worst thing you could do.
Yes.
Have you had your holidays Doctor?
No no.
No?
October.
October.
There we are and I'll I'll get on to the Royal this morning.
Thanks ever so much Doctor.
And I'll be
And we'll try and get worked out to you this week.
greatly obliged to you.
Right Mrs ,
Yes.
I'll just go straight through just now.
Right and thank you Doctor .
Now have we got your phone number?

Er.
Yes.
Wait a bit, Yes that's right.
That's right , .
Okay.
we know where to find you.
Thanks Doctor, thanks very much.
Right that's enough.
Bye bye.
Right cheerio now.

Notes

[1] See Scott & Tribble (in press) chapters 4 and 5.
[2] No claim is made here that the BNC texts are themselves complete; this is recognised by Aston and Burnard (1988: 28).