Teenage Health Freak Encyclopedia: a description

Svenja Adolphs*, Kevin Harvey*, Louise Mullany* and Catherine Smith**

* The University of Nottingham, UK
** University of Birmingham, UK

The Teenage Health Freak Encyclopedia consists of the top 20 keywords in each of the five main topics represented in the Teenage Health Freak corpus. The following provides a description of the research methods that have been used to derive the keywords, the main topics and the data representation that form the basis of the entries in the Teenage Health Freak Encyclopedia [i].

The Teenage Health Freak Website

The basis of our analysis is a 2.2-million-word corpus of electronic requests for advice sent to the adolescent health website, Teenage Health Freak [ii]. Operated by UK-based GPs specialising in adolescent health, the Teenage Health Freak website has been running and continuously updated on a weekly basis since its launch in 2000. It is designed to be interactive, confidential and evidence-based, providing adolescents with accessible advice and information pertaining to a broad range of health issues. Adolescents are able to submit their health questions anonymously to the online GP persona, Doctor Ann.

The data submitted to the THF website thus afford an opportunity to analyse the variety of terms which adolescents use to conceptualise and articulate their health concerns when interacting through the anonymised medium of electronic discourse. The approach taken to extract patterns of language use is outlined in the next section.

Analysing specialist language samples using corpus linguistic techniques

Corpus linguistics is becoming increasingly popular in the analysis of health communication (Adolphs et al. 2004; Harvey et al. 2008). Corpus linguistics as a methodology relies heavily on large samples of homogenous data. Smaller, specialised corpora which include a high rate of spelling variation, are more difficult to analyse with the use of corpus linguistic methods. This is partly due to the level of recurrence required to make robust statements about the meaning and use of individual lexical items and phrases, and partly due to the fact that search mechanisms cannot easily predict spelling variations.

Data overview

The corpus used for this study consists of 113,480 submitted messages using the 'Ask Doctor Ann' facility on the Teenage Health Freak website between 2004 and 2009. These linguistic data provide a considerable snapshot of the health concerns routinely communicated by young people. The website possesses a privacy policy that informs potential senders of health problems to the website that in submitting such information they are not asked for their names, email addresses or other personal details. Thus all the data transmitted to the site is received in confidence and any information supplied which contains personal information is removed. The privacy policy also informs contributors that the information they provide may be used for research purposes and that, in using the website, they consent to the collection and use of the data which they supply.

Table 1 shows a break-down of the THF corpus by gender including average message lengths. As this table shows females send more messages to the website than males and the messages sent by females are on average much

longer. Overall, females contribute twice as many words to the corpus compared to males.

	All	Male	Female	Unspecified
Total Messages	113,480	41,830 (37%)	59,884 (53%)	11,766 (10%)
Total Words	2,217,919	667,277 (30%)	1,442,784 (65%)	107,858 (5%)
Median Message Length	10	8	13	6

Table 1: Summary of the corpus by gender and message length

As Table 1 shows, the corpus does not contain an equal number of messages from male and female contributors, and the same is true for the different age groups represented in the corpus. This means that it is not possible to make direct comparisons between the different demographics using raw frequency data alone. In order to be able to draw comparisons, the data has to be normalised by the amount of messages received from the demographic in question (Baker 2010: 19-21; McEnery et al 2006: 52-53). As we discuss later, our keyword analysis of the corpus has generated five broad topic categories of messages. We show below how the normalisation process is applied to one of those categories, i.e. that of 'smoking, drugs and alcohol'. The raw data for gender distribution in the messages in this category is shown in table 2. This shows that there are around 2,000 more messages sent by females than males about this top, with a further 881 messages unspecified for gender. However, as shown in Table 1, there are around 18 thousand more messages overall which are sent by females and less that 12 thousand messages overall do not specify a gender. Consequently, these figures cannot be accurately compared with each other.

Gender	No. of Messages
Male	2693
Female	4509
Unspecified	881

Table 2: Raw data for messages related to 'smoking, drugs and alcohol'

In order to take account of the underlying imbalance in messages sent by the different genders, we have normalised raw frequency counts by number of messages. In this case, we are using 1,000 messages as the unit of normalisation.

The normalisation process leads to new figures summarised in table 3. Relative to the number of messages sent by each gender, the numbers are much more even than the raw data suggests. This is particularly noticeable with the data for unspecified genders.

Gender	No. of Messages Normalised per 1,000
Male	64.38
Female	75.30
Unspecified	74.88

Table 3: Normalised data messages related to 'smoking, drugs and alcohol'

We will return to the use of normalised frequencies for individual lexical items identified as keywords in the corpus, as well as for gender and age information in our analysis below.

Data preparation

Although the messages that constitute the corpus data were 'born digital', the corpus had to be structured trough a number of processes to be fully usable for our analysis. The data was supplied in MS Access and was transferred into an MS Excel spreadsheet. A Python script was used to turn the MS Excel spread sheet into XML. Further clean-up was achieved by using a Python script to address any character encoding problems, remove empty messages and remove exact duplicate messages sent on the same date. In the data there were many occurrences of very similar messages being sent one after the other. In some messages spelling was corrected in the later messages, in others a few details were added, deleted or changed. It was decided that if it was highly likely that two similar messages were sent by the same person, then one of the messages should be removed from the corpus. In order to facilitate this, a script was used to identify possible duplicate messages. The script used the Lenvenshtein Distance Algorithm (see Rayson et al 2008), which calculates the similarity between two strings, to find messages that are similar. The calculation was performed both on the length of the longest message and also the length of the shortest message which allows the algorithm to find messages which have later been added to. Only messages from the same date were considered as possible duplicates.

Once the duplicates were identified and logged to the text file, another script reads the log file and presents each pair of potential duplicates to the user along with the timestamp of each message. This allows the user to consider each pair of messages and decide which, if any, should be deleted. The message that was deleted was either the one with more spelling errors or the shorter message.

The data presents many challenges with regards to the spelling used. Spelling was corrected as far as is possible by using the keyword procedure in Wordsmith Tools (Scott 2008) to identify consistently misspelled words. The reference corpus selected for this particular task was the written BNC. Once the keywords had been generated, the misspelled words were manually corrected. The resulting corrections were added to the corpus using the TEI <choice>, <corr> and <sic> tags so that the original text was not lost (see also Smith et al 2014).

Keywords

As a way into this substantial dataset and in order to provide a survey of the salient health themes in the corpus, the first stage of the analysis involved identifying keywords that appeared in the corpus of health messages. To this end, we generated a list of keywords which provide a thematic characterisation of the health email corpus. Keywords, according to McCarthy and Handford (2004: 174), are words which 'best define' a text or texts. They are words that occur with a significantly higher frequency in one data-set when compared with another. Keywords are an important indicator of both expression and content (Seale et al. 2006), and have been used by an increasing number of researchers as a reliable means of identifying key themes in health language corpora (Adolphs et al. 2004; Harvey et al. 2008). The advantage of using statistical keywords is that they remove the *a priori* biases of the analyst from the identification of themes of significance and interest (Baker 2004). Thus keywords present the analyst with evidence that a conventional thematic qualitative analysis might obscure from view (Seale et al 2006), and serve as an effective means of identifying salient themes that warrant further exploration in context (Baker 2006).

The Effects of Reference Corpora Choice on Keywords

Keywords are the starting point for much lexically oriented corpus based research (Archer 2009, Culpeper 2009, O'Keeffe et al 2007, Scott and Tribble 2006). In

this section we will consider the effects of the choice of reference corpus on the keyword list that is being generated. This is an important step in the research design both in terms of enabling replicability of results, and in terms of providing the basis on which the robustness of results can be assessed. This process informs the choice of reference corpora by highlighting the impact of mode, size, and genre of reference corpora on the resulting keyword list. As the basis of our analysis, we use the version of the Teenage Health Freak corpus that has been corrected for spelling errors. Keywords are extracted using WordsmithTools 5.0 (Scott 2008). The minimum frequency has been set at 5 and the maximum p value at 0.0001. This is equivalent to a log likelihood value equal to, or greater than, 15.13 (Rayson et al. 2004: 933).

The corpora selected for comparison are the British National Corpus (BNC) and the British component of the International Corpus of English (ICE-GB), both as full corpora and split into their spoken and written components, the Nottingham health communication corpus (NHCC), the Cambridge and Nottingham Corpus of Discourse in English (CANCODE), a component of the Cambridge and Nottingham E-Language Corpus (ELC) consisting of blogs and twitter feeds. The approximate word counts for these corpora are shown in Table 4 below together with their mode and domain.

Corpus	Word Count	Mode	Domain	Comparison to THF size (2,000,000)
BNC	100,000,000	90% written	General English	50 times larger
BNC written	90,000,000	written	General English	45 times larger
BNC spoken	10,000,000	spoken	General English	5 times larger
CANCODE	5,000,000	spoken	General English	2.5 times larger
ICE-GB	1,000,000	60% spoken	General English	1/2 the size
ICE-GB spoken	600,000	spoken	General English	1/3 the size
ELC	500,000	written (online)	General English (online)	1/4 the size
ICE-GB written	400,000	written	General English	1/5 the size
NHCC	250,000	spoken	English in health care settings	1/8 the size

Table 4: Information for selected corpora

The number of keywords generated against each reference corpus provides the starting point for our comparison.

Corpus	Number of key words generated
BNC	2,224
BNC written	2,260
BNC spoken	2,392
CANCODE	2,040
ICE-GB	1,167
ELC	797
NHCC	657

Table 5: Number of Keywords generated against selected corpora

As can be seen in Table 5, the British National Corpus and its sub-corpora produce similar numbers of keywords even though the spoken section of the BNC

is considerably smaller than the written section, and of course the full BNC. These findings align with those of Berber-Sardinha (2000) who suggests that the optimum size for a reference corpus is five times the size of the target corpus. He finds that a reference corpus that exceeds this multiplier does not generate a significantly larger set of keywords (Berber-Sardinha, 2000: 11). In our study, the number of keywords generated against CANCODE is almost as high as the results from the BNC even though this corpus is not above Berber-Sardinha's (2000) suggested threshold. The fact that the spoken BNC generated more keywords than the written BNC might indicate that spoken corpora will generate more keywords for the Teenage Health Freak data. This may also explain why CANCODE generates a high number of keywords. Although Berber-Sardinha (2000) does not consider the effects of comparing a corpus with another corpus that is considerably smaller, the pattern he finds suggests that we should expect less keywords to be generated when comparing the Teenage Health Freak corpus with corpora smaller than itself (Berber-Sardinha, 2000:10). This trend appears to be related to the size rather other features in the reference corpus, as there is a drop off in the number of keywords generated against the ICE-GB and the CANELC data which are both considerably smaller than the Teenage Health Freak corpus, but differ in both mode and genre.

In his paper examining the effect of different reference corpora on keywords, Scott (2009) outlines a method for comparing keywords generated against a range of corpora. He uses two measures: popularity and precision. Popularity refers to the presence of the keyword in a specified number of the different keyword lists under consideration (Scott, 2009:85). In his study of 22 different sized reference corpora popularity was determined as presence in at least 20 of the 22 lists (Scott, 2009:85). In his study of different genres of reference corpora, it was determined as a presence in at least eight of nine of the word lists (Scott, 2009:90). Precision is calculated as the number of popular keywords divided by the number of keywords in the list, times 100 (Scott, 2009:85). Scott's study uses two different target corpora: a 44,000 word section of a book, and a 615 word conversation. In his study of 22 different sized reference corpora, Scott (2009) reports precision figures for keywords generated using the 44,000 word target corpus as consistently over 50% rising to over 70% as the size of the reference corpus increased. With the smaller corpus, however, precision figures were consistently over 75%, reaching a high of 100% with two of the corpora towards the lower end of the size range.

When considering the effects of different genres on the keyword lists, Scott finds precision values between 20% and 70%. This shows that different genres have a much larger effect on keyword lists than the size of the reference corpora. He concludes that it is likely that different aspects of the 'aboutness' of the target corpus are being highlighted by the use of the different genres. He also points out that the precision measure is less appropriate for such a study as there are several variables at work at the same time.

In Scott's (2009) study all of the reference corpora are larger than the target corpus. This is not the case with the Teenage Health Freak corpus experiments. As can be seen in Figure 1, the smaller reference corpora with their correspondingly smaller lists of keywords have a considerable impact on the overall pattern of precision figures. In Figure 1, the popularity figure requires the keywords to be present in all seven corpora due to the disparity in the lengths of the keyword lists. If the popularity threshold is reduced, the precision calculation does not work for the smaller corpora as the popularity figure becomes larger than the total number of keywords on the list generated against the smallest corpus. The pattern seen with the different permutations of the BNC (the only corpora bigger than our target corpus) mirrors the findings of Scott (2009) in that

the larger the reference corpus the higher the precision value. When taken on their own they also have precision values between 75% and 85%.

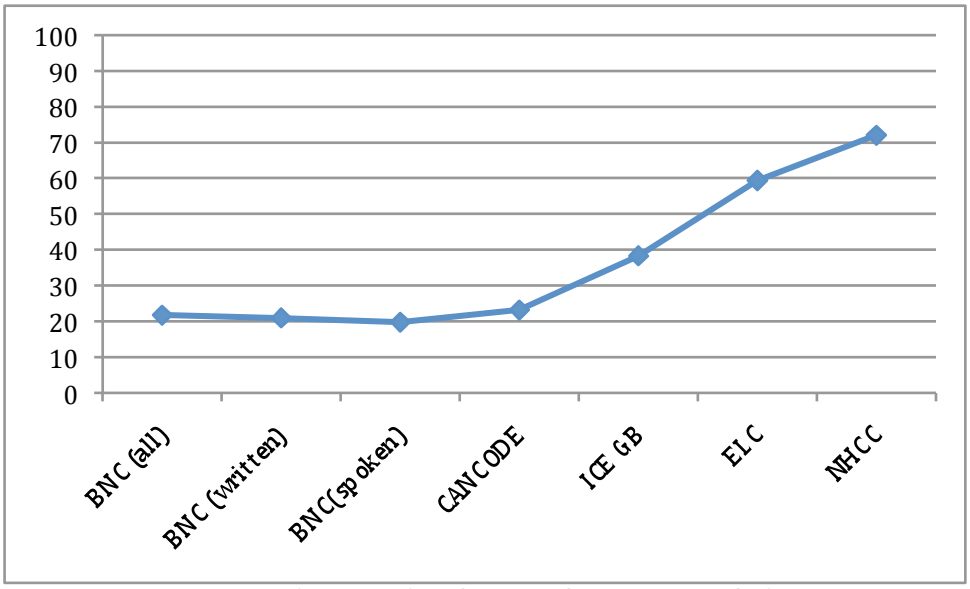

Figure 1: Precision figures for keywords generated for THF using all seven reference corpora

Due of the different number of keywords generated by the different corpora investigated in our research, and the fact that many studies combining corpus linguistics with other approaches focus on the keywords towards the top of the list, we repeated our calculation concentrating on only the top keywords from each list. The smallest corpus generates 657 keywords and popularity and precision were calculated in steps of 50 from 50 to 650 keywords. Figure 2 shows the results from this analysis where the keywords generated against each of the corpora are compared with those generated against the full BNC. When using an artificial cut-off for the keywords, the precision figure is the same for all corpora being compared at each point. For example, the first point for 'BNC(all) plus ELC' is calculated by taking the number of words shared between the two corpora in the top 50 words (popularity), dividing this by 50 (the number of keywords under consideration) and multiplying the result by 100.

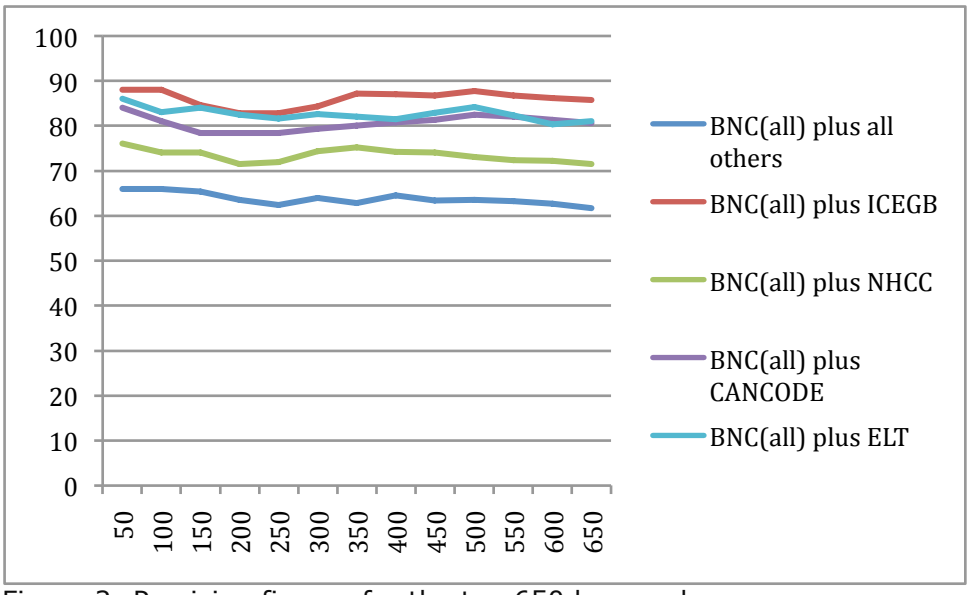

Figure 2: Precision figures for the top 650 keywords

Figure 2 shows that the corpus with the most keywords in common with the BNC is ICE-GB. Given that these are both designed as general reference corpora this is not surprising. ICE-GB however is only half the size of the Teenage Health Freak corpus. Thus it appears that although the size of the corpus has had an effect on the top 650 keywords, it is not as pronounced as the size difference between the BNC and ICE-GB might suggest based on Berber-Sardinha's research (Berber-Sardinha 2000). Between 80% and 90% of keywords generated in the top 650 against the BNC are also generated against ICE-GB. The E-Language Corpus also shares between 80% and 90% of the top 650 keywords with the BNC, with CANCODE located around the same level. These corpora can also both be considered to represent general English but differ in modes, representing internet language and spoken language respectively. Again they differ in size, with the CANCODE corpus being 2.5 times larger than the Teenage Health Freak corpus and the E-Language corpus being only a quarter of the size of the Teenage Health Freak corpus. This size difference does not seem to have had a significant impact on the results.

In contrast, the comparison between the BNC and the NHCC show much lower precision figures than the other pairs, although it should be noted that these figures still reach over 70%. The NHCC is by far the smallest corpus in this experiment, and the only corpus which represents specialized health communication genres rather than general English. It is possible that the size difference has impacted on the precision figures in this case, however, given that size does not seem to have affected the other pairings, it is worth considering other factors involved. The spoken mode did not have an impact in the comparison between the Teenage Health Freak corpus with CANCODE.

It can be assumed at this stage that mode of interaction is equally not a central factor in the comparison with the NHCC. By process of elimination, we suggest that it is the genre of the material which is causing the drop in precision figures, and thus introducing more new keywords when compared to the BNC. This conclusion is supported by a control test using the Cambridge and Nottingham Spoken Business English Corpus (CANBEC), a 1 million word corpus (the same size as ICE-GB) that contains only spoken data. This corpus shows similar precision figures to the NHCC especially at the top of the keyword list, as can be seen in Figure 3 below.

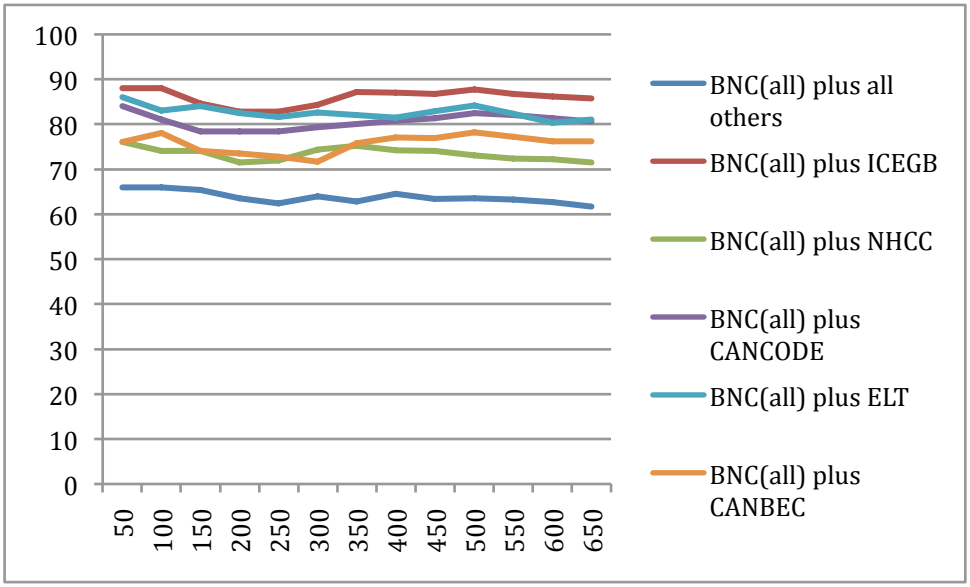

Figure 3: Precision figures for the top 650 keywords including CANBEC

In his paper, Scott (2009) does not draw any conclusions on the quality of the keywords generated by using different genres of reference corpora. He points out that unlike his study of different sized reference corpora, there are several different variables at play when considering differences in genres. However he does conclude that a significantly different reference corpus in terms of genre (in this case the plays of Shakespeare) generates more keywords than using the BNC. However, an examination of the key words shows that they still offer potentially useful keywords for the study of his source texts. Scott concludes that the use of different genres for the reference corpus can highlight different types of 'aboutness' in the target corpus (Scott 2009:90). As with Scott's study there are several different variables at play in the comparisons represented in Figure 3.

In order to minimise the interference from other variables, we suggest that selected reference corpora will be studied in pairs. In order eliminate the effects that are due to size, only corpora smaller than the target corpus will be considered. The pairs to be studied and the comparisons of mode and genre are show in Table 6.

Reference Corpora	Mode Comparison	Domain Comparison
ICE-GB Written vs ICE-GB Spoken	Written vs Spoken	
ELC vs ICE-GB Written	Internet vs Written	
ELC vs ICE-GB Spoken	Internet vs Spoken	
CANBEC vs ICE-GB Spoken		Business vs General
CANBEC vs NHCC		Business vs Health Care
NHCC vs ICE-GB Spoken		Health Care vs General
NHCC vs ELC	Spoken vs Internet	Health Care vs General

Table 6: Reference corpora pair comparisons

The precision figures for these pair comparisons can be seen in
Figure **4**. The precision figures for all but one comparison are consistently over 80%. Despite the high figures, this still means that varying the mode and domain of the reference corpora will generate around 15% of distinct keywords. This suggests that if broad coverage of keywords is important, it is worth using a variety of reference corpora to generate the keywords rather than relying on a single one. The comparison that stands out as being distinct is the one between the NHCC and the ELC. As Table 6 shows, this is the only comparison which changes both mode and domain thus supporting the idea that switching both the mode and domain of the reference corpus will generate a wider range of keywords.

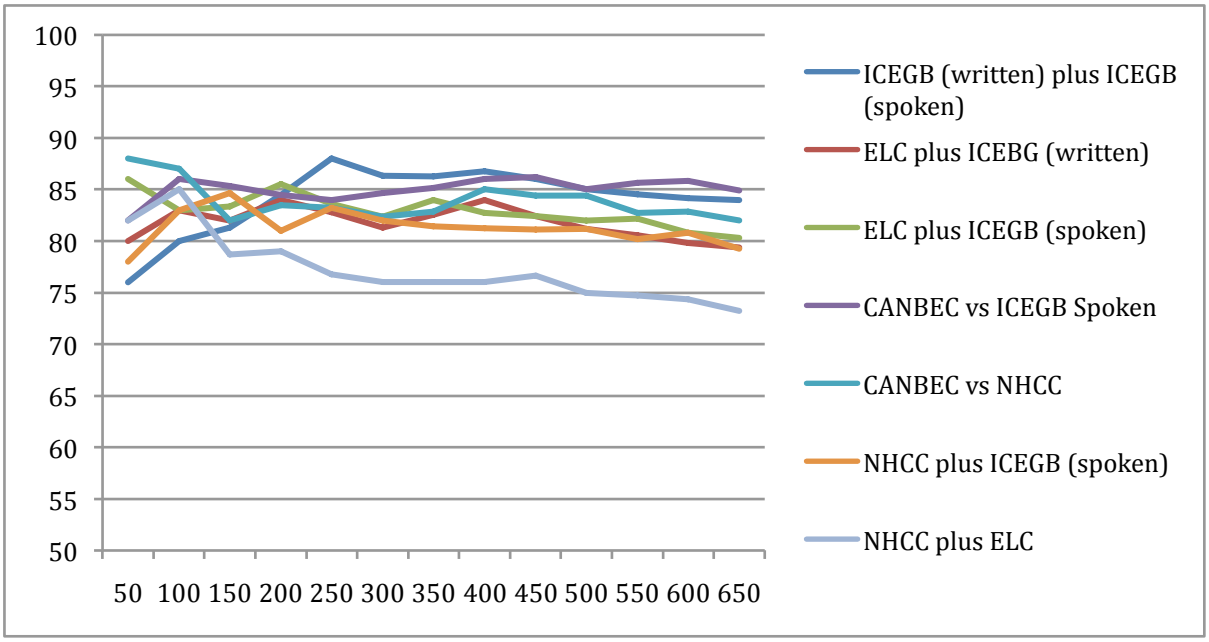

Figure 4: Precision figures for top 650 keywords in selected comparison pairs

When evaluating the results in Figure 3 and Figure 4 together, we can draw some conclusions with regards to the choice of reference corpora. Firstly, the high levels of shared keywords found in Figure 3 suggests that the BNC provides robust general keywords for our data. However, it also suggests that, should large general reference corpora not be available for some reason, smaller general English reference corpora also provide adequate keywords, especially if only the top 500 keywords are going to be considered. In both Figure 3 and Figure **4**, the NHCC stands out as providing quite distinct keywords when compared to the BNC and also the ELC. This corpus does stand out as being particularly small and is also the only corpus available which shares the health care genre with the target corpus. However, this is a very broad genre and the actual similarities in this comparison are marginal. Due to the difficulty of obtaining data from this sector, it is not possible to rerun these tests with a larger reference corpus collected in a health care setting. We conclude that using a different genre of reference corpora is likely to have more impact on the keywords generated than varying the mode. This supports Scott's suggestion that using reference corpora which vary by genre can highlight different aspects of the target text's 'aboutness' (Scott 2009:90).

Keywords and topics in the teenage health freak corpus

Following on from our study of different reference corpora above, we use the 100 million word British National Corpus (BNC) as the reference corpus in the following analysis. Keywords are thus defined as those that occur with a significantly greater frequency in the Teenage Health Freak corpus when compared with the BNC.

Keywords for the Teenage Health Freak corpus are generated with Wordsmith tools (Scott 2008) using the log likelihood statistic with a p value of 0.000001. Table 7 shows the top 40 content keywords generated from this analysis. As discussed, one of the functions of a keyword analysis is to highlight the 'aboutness' or themes in a text or corpus. Some themes emerge even when considering this relatively small set of keywords, e.g. words connected with sex, pregnancy and relationships (words 1, 7, 12, 28, 33, 34, 36).

1	SEX	11	REALLY	21	CAN	31	PERIOD
2	AM	12	BOYFRIEND	22	FEEL	32	IS
3	DO	13	HI	23	NORMAL	33	PILL
4	PENIS	14	HOW	24	WHY	34	CONDOM
5	WHAT	15	GET	25	DEAR	35	FRIENDS
6	HAVE	16	WORRIED	26	DR	36	GIRLFRIEND
7	PREGNANT	17	DON'T	27	SCARED	37	GIRL
8	PLEASE	18	GAY	28	MASTURBATE	38	SMOKING
9	ANN	19	WANT	29	FRIEND	39	BOY
10	HELP	20	VAGINA	30	MUM	40	THINK

Table 7: Top 40 THF keywords generated against the BNC

In total, over two thousand keywords were generated using this procedure and when the full list is taken into account, five main themes emerged from the data. The analysis of clustering keywords for the purpose of assigning them to broader themes was carried out manually. The emerging topics are listed in Table 8. 'Types' here refers to the number of different words (or word forms) classified in the topic while 'tokens' refers to the total number of instances of all the types present in the data. By far the most dominant topic in the data is 'sex, pregnancy and relationships'. Around one third of all messages submitted to the website contain at least one keyword classified under this topic. In contrast, the topic of 'weight and eating' accounts for only 5 percent of the messages but still totals over 6,000 messages. The topic of 'weight and eating' contains on average the longest messages of all the topics, suggesting that the messages in this topic (along with that of 'body changes') are more detailed and involved than some of the larger topics.

Topic	**Types**	**Tokens**	**Median Message Length**
Sex/Pregnancy/Relationships	166	62,804	13
Sexual Body Parts	80	24,492	13
Body Changes	37	14,754	18
Weight and Eating	40	12,703	19
Smoking/Drugs/Alcohol	46	11,796	10

Table 8: Topical keyword summary

Further keyword lists were generated for each of the main themes. The themes became sub-corpora of the overall Teenage Health Freak corpus, with further divisions relating to age and gender of the advice seekers in each of the main themes.

As a next step in the analysis, individual keywords are considered in more detail, both in terms of their patterns of use, distribution across different age groups and gender, and their prominence over time (see Adolphs et al 2011). Using as an example the word 'pill', we illustrate below the information that was extracted from the corpus.

Having identified 'pill' as a keyword, a KWIC search starts to provide some further context about the way in which this word is used. Below is a sample of 20 randomly chosen concordance lines.

hi i'm on the	pill but recently missed quite
2 go on	pill as i don't want
I'm not on the	pill. I was wondering if
you go on the	pill to reduce your
and received the mini	pill going by the name
on birth control (the	pill) for a while

i did take the	pill. It made me fell
took the morning after	pill the following day. i
about going on	pill because me and my
anti-botits and then my	pill got messed up and
forgot to take my	pill. i have missed 3
you take the contraceptive	pill more than once since
for an emergency contraceptive	pill but i am worried
get the contraceptive	pill and does a parent
go on the combined	pill
i have taken	pill called [name] 3days three
I am on the	pill
as i take the	pill does it start
took the morning	pill 4hours ago, ,
forgot to take a	pill and am worried i'm

Patterns emerging from the concordance output include concerns about the appropriate course of action when missing a pill, the use of the emergency contraceptive pill, and the different types of the pill. This output also illustrates the issue of spelling variation discussed above and provides a typical example of the nature of the data in the corpus.

Additional information about the use of the word 'pill' can be extracted by combining the different sub-corpora to reflect the distribution over time (Table 9), by age (Table 10) and by gender (Table 11).

Years	Messages		Words		Median Message Length	
	Raw	Normalised per 1000 messages	Raw	Normalised per 1000 messages	Containing word	All
2004	381	15.47	491	97.65	35	10
2005	401	16.16	515	102.45	33	10
2006	340	15.91	442	107.56	33	10
2007	260	12.63	321	81.65	34	9
2008	114	8.43	147	59.56	37	9
2009	75	8.72	106	65.58	52	9

Table 9: 'Pill' Distribution by year

Ages	Messages		Words		Median Message Length	
	Raw	Normalised per 1000 messages	Raw	Normalised per 1000 messages	Containing word	All
10 or younger	5	0.95	5	7.08	10	9
11	8	1.66	8	9.83	9	10
12	18	1.47	19	9.28	19	9
13	72	3.37	78	20.44	21	9
14	158	7.41	183	43.00	28	10
15	338	24.82	425	124.80	31	13
16	369	41.96	505	195.20	40	16
17	493	37.86	675	215.62	43	11
unspecified	110	8.38	124	87.20	12	6

Table 10: 'Pill' Distribution by age

Gender	Messages		Words		Median Message Length	
	Raw	Normalised per 1000 messages	Raw	Normalised per 1000 messages	Containing word	All
Male	77	1.84	100	14.98	43	8
Female	1419	23.69	1843	127.73	36	13
Unspecified	75	6.37	79	73.24	9	6

Table 11: 'Pill' Distribution by gender

This data shows that the overall percentage of messages containing the word 'pill' reaches a peak in 2005 and then decreases until 2009. In terms of age groups sending messages containing the word 'pill', there is a steady increase from the '10 or younger' category to the 16 year old category when requests for advice mentioning the word 'pill' peak. As can be expected there are significantly more females than males using the word 'pill' in their messages as the gender comparison in Table 11 shows.

Conclusion

We have illustrated the processes and procedures that have been applied in order to render a 2 million word corpus of on-line communication usable for linguistic and demographic analysis. The approach we have described is transferable to other domains of online discourse and to the exploration of other specialised corpora, where issues of data preparation, the use of reference corpora in keyword analyses, and the issue of spelling variation are a key concern. The analysis of this kind of data allows applied linguists to advance their own methodological frameworks by developing new ways of extracting linguistic patterns from corpora that are challenging in terms of their small size and spelling variation, while at the same time offering a new basis for interdisciplinary endeavours with health sciences and engagement with non-academic audiences.

ENDNOTES

[i] Research reported in this article has been funded by the *Economic and Social Research Council*, Grant Number RES-000-22-3448.

[ii] Teenage Health Freak website URL: http://www.teenagehealthfreak.org

REFERENCES

Adolphs, S., B. Brown, R. A. Carter, P. Crawford and O. Sahota. 2004. 'Applied clinical linguistics: corpus linguistics in health care settings.' *Journal of Applied Linguistics* 1/1: 9-28.

Adolphs, S., L. Mullany, K. Harvey and C. Smith. 2011. 'Am I Normal? What adolescents want to know about health.' *Health Education Booklet*. Nottingham: University of Nottingham.

Archer D. 2009. *What's in a Wordlist: Investigating Word Frequencies and Keyword Extraction*. Burlington, VT: Ashgate.

Baker, P. 2004. 'Querying keywords: questions of difference, frequency and sense in keywords analysis.' *Journal of English Linguistics* 32/4: 346-359.

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Berber-Sardinha, A.P. 2000. 'Comparing corpora with WordSmith Tools: How large must the reference corpus be?' *WCC '00 Proceedings of the workshop on Comparing corpora*. Morristown, NJ, ACL: 7-13.

Harvey, K., D. Churchill, P. Crawford, B. Brown, L. Mullany, A. Macfarlane and A. McPherson. 2008. 'Health communication and adolescents: what do their emails tell us?' *Family Practice* 25: 1-8.

McCarthy, M. J. and M. Handford. 2004. ''Invisible to us': A preliminary corpus-based study of spoken business English' in Connor, U. and T. Upton (eds.): *Discourse in the professions: Perspectives from corpus linguistics*. Amsterdam: John Benjamins: 167–201.

McEnery, T., R. Xiao and Y. Tonio. 2006. *Corpus-based language studies: An advanced resource book.* New York: Routledge.

O'Keeffe, A., M. J. McCarthy and R. A. Carter, 2007. *From Corpus to Classroom: language use and language teaching*. Cambridge: Cambridge University Press.

Rayson, P., D. Berridge and B. Francis. 2004. 'Extending the Cochran rule for the comparison of word frequencies between corpora' in Purnelle G., C. Fairon and A.

Rayson, P., D. Archer, A. Baron, and N. Smith. 2008. 'Travelling through time with corpus an- notation software' in Lewandowska-Tomaszczyk, B. (ed.): *Corpus Linguistics, Computer Tools, and Applications – State of the Art*. PALC 2007, *Studies In Language*, Frankfurt am Main. Peter Lang: 29-46.

Scott, M. and C. Tribble. 2006. *Textual Patterns: Key words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

Scott, M. 2009. 'In Search of a Bad Reference Corpus' in Archer, D. (ed.): *What's in a Word List? Investigating Word Frequency and Keyword Extraction*. Ashgate Publishing: 79-91.

Scott, M. 2008. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

Seale, C., G. Gobo, J. Gubrium and D. Silverman, D. (eds.) 2006. *Qualitative Research Practice*. Sage Publications.

Smith, C., Adolphs, S., Harvey, K. and Mullany, L. (2014) Spelling Errors and Keywords in Born-Digital Data: A Case Study using the Teenage Health Freak Corpus. *Corpora 9.2.*