Spelling errors and keywords in born-digital data: a case study using the Teenage Health Freak Corpus

Smith, Catherine; Adolphs, Svenja; Harvey, Kevin; Mullany, Louise

DOI:
10.3366/cor.2014.0055

Citation for published version (Harvard):
Smith, C, Adolphs, S, Harvey, K & Mullany, L 2014, 'Spelling errors and keywords in born-digital data: a case study using the Teenage Health Freak Corpus', *Corpora*, vol. 9, no. 2, pp. 137-154.
https://doi.org/10.3366/cor.2014.0055

Spelling errors and keywords in born-digital data: a case study using the Teenage Health Freak Corpus

Catherine Smith,[1] Svenja Adolphs,[2]
Kevin Harvey[2] and Louise Mullany[2]

Abstract

The abundance of language data that is now available in digital form, and the rise of distinct language varieties that are used for digital communication, means that issues of non-standard spellings and spelling errors are, in future, likely to become more prominent for compilers of corpora. This paper examines the effect of spelling variation on keywords in a born-digital corpus in order to explore the extent and impact of this variation for future corpus studies. The corpus used in this study consists of e-mails about health concerns that were sent to a health website by adolescents. Keywords are generated using the original version of the corpus and a version with spelling errors corrected, and the British National Corpus (BNC) acts as the reference corpus. The ranks of the keywords are shown to be very similar and, therefore, suggest that, depending on the research goals, keywords could be generated reliably without any need for spelling correction.

Keywords: computer-mediated communication, keyword analysis, spelling variation

1. Introduction

Corpora are more frequently being created from texts taken from the Internet and, therefore, the issue of spelling errors and non-standard spellings

[1] ITSEE, Room 228, ERI Building, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom.
[2] School of English, Trent Building, University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom.
Correspondence to: Catherine Smith, *e-mail*: c.j.smith@bham.ac.uk

This work was supported by Economic and Social Research Council [grant reference RES-000-22-3448].

Corpora 2014 Vol. 9 (2): 137–154
DOI: 10.3366/cor.2014.0055

www.euppublishing.com/loi/cor

are increasing for corpus compilers. Unlike corpora built with published, well-edited materials, the increasing spread of user-generated content present in the Web 2.0 world presents new challenges to the corpus creator in terms of ‘typos’ since this material is not edited to the standard of traditional print media or even large company websites. In addition, in the same way that non-standard spelling poses problems for corpus researchers working with historical corpora (see Baron *et al.*, 2009) the use of non-standard spellings in e-language is reintroducing the problem to modern corpora. The use of chat and text abbreviations is widespread in the Internet community and innovations are always being introduced (Baron, 2008; and Crystal, 2006, 2011). In this paper, we aim to investigate the impact that spelling errors and non-standard spellings may have on keywords generated with a born-digital corpus.

The corpus used in this case study is comprised of health questions sent to Doctor Ann through the Ask Doctor Ann facility on the Teenage Health Freak website.[3] The messages date from January, 2004, to December, 2009 – a period of six years. With the exception of the removal of very similar messages sent within a short time frame, the corpus is unedited and contains all messages that were sent to the website during the time period in question; only a small number of these are published on the website with answers from the medical staff. The messages themselves are typed directly into a web form. Apart from the automatic removal of personal details (such as e-mail addresses) the corpus contains exactly what was typed into the form by the users. In total, the corpus contains 113,480 messages and 2,217,919 words. This corpus is unique in many respects and is of particular interest to the medical community, since the questions posed are unsolicited and the users of the site are able to ask about whatever aspect of their health is of concern to them. The fact that the corpus is generated by adolescents is also of interest to the linguistic community. Adolescents are typically seen as language innovators (Stenström *et al.*, 2002) and this is no exception in the on-line community (Crystal, 2006: 94).

Keywords are the starting point for a great many lexically orientated corpus studies (Baker, 2006; Harvey *et al.*, 2008; and Scott and Tribble, 2006: 55–72). They provide a good lead into the study of a large corpus, highlighting words and topics which are unusually frequent in the data and, therefore, may warrant further investigation. Such studies are classed by Rayson (2008) as type III corpus studies. These studies are categorised by ‘the use of corpus-based comparative frequency evidence to drive the selection of words for further study’ (Rayson, 2008: 523). Scott (1997: 236) defines a keyword as ‘a word which occurs with unusual frequency in a given text … by comparison with a reference corpus of some kind’. Statistical procedures are used to determine whether the comparative frequencies of a word in the target and reference corpora are significant enough to classify the word as key in the target corpus. Several measures have been used

[3] See: http://www.teenagehealthfreak.org

to determine which words should be considered to be key, including the chi-squared test (Hofland and Johansson, 1982), the Mann–Whitney test (Kilgarriff, 2001) and the T-score (Paquot and Bestgen, 2009). The most widely used statistical procedure, however, is log-likelihood.

The log-likelihood statistic is calculated using two figures: the observed frequency (the counts for the word in question in the data) and the expected frequency (the frequency that would be expected if the distribution was due to chance alone). Fundamental to the log-likelihood statistic, as with all other measures mentioned above, are the word counts for each of the words in both the target and reference corpora. If one of the corpora contains problems with non-standard spelling, or contains a large volume of spelling errors, the counts for the words will be affected and, consequently, also the log-likelihood score. If, for example, the word *what* occurs in the corpus with the standard spelling and also with the two chat-style abbreviations, *wat* and *wot*, then the overall counts for *what* in the target corpus will be reduced since it is being represented by several different orthographical forms. This problem becomes even greater for the statistical calculation if, as is the case here, only the target corpus contains such inconsistencies. In this case, the non-standard and incorrect spellings could dominate the keyword list since they are not likely to be present in the reference corpus at all. This kind of challenge is encountered in a variety of different corpora: historical corpora, for example, in which spelling conventions were yet to be established (Baron *et al.*, 2009), also regional corpora, where dialects are being represented (Kay, 2006), as well as born digital corpora, which use non-standard spelling, of the kind employed in this research.

In this paper, we first analyse the volume and type of spelling errors and non-standard spellings found in the Teenage Health Freak Corpus to establish the scale and nature of the problem faced. The variant spelling in the corpus is then corrected as far as possible to create a second normalised corpus. These two corpora are then compared with respect to the keywords they produce against the same reference corpus in order to establish any differences in keyword rank due to spelling variation. Here, the term spelling error is used as an umbrella term for spelling error, deliberate non-standard spelling and other abbreviations or acronyms that represent words or phrases. This is done as a matter of convenience and is not intended as any judgment on the language use itself.

2. Corpus approaches to spelling variants and errors

Interest in spelling variation in corpus linguistics has, typically, been focussed on historical corpora and learner corpora. In historical corpora the problem is caused by a lack of standardised spelling. Research by Baron *et al.* (2009: 53) suggests that variant tokens make up between 35 percent and 40 percent of all tokens in English corpora from the period 1400 to 1550; this gradually reduces to below 10 percent by 1650. The problems that this level of variation cause for part-of-speech tagging and subsequent semantic tagging using the online tool Wmatrix (Rayson, 2009)[4] inspired the creation of a variant detection tool or VARD (Archer *et al*., 2003; and Rayson *et al*., 2007: 1). The latest version of VARD achieved impressive results with historical corpora with levels of precision at over 90 percent and recall reaching 65 percent with sufficient training (Baron and Rayson, 2009: 14). Recently, VARD has been developed from a tool that was specifically designed to deal with historic English corpora to a tool that can, potentially, be used to deal with any kind of spelling variation or error in corpora written in any language (Baron and Rayson, 2009: 4–9, 13). VARD can be customised in several ways – for example, by changing the dictionary of accepted spellings or adding new letter-replacement rules, and can also be trained with a specific dataset (Baron and Rayson, 2009: 9). The performance of VARD on a corpus of children's writing was much lower than the figures achieved with historical corpora, with levels of precision of around 80 percent, but a recall no higher than 20 percent. In this experiment, the only customisation performed was to train VARD with samples of the corpus; much better recall figures could be expected if all of the customisation features of VARD had been employed (Baron and Rayson, 2009: 19).

In learner corpora, spelling errors are only one of several types of error that are of particular interest to the compilers and users of the data. Most of the work in identifying the errors is done manually, but a few tools have also been developed to assist with the process of finding and correcting errors. A tool has been developed for computer assisted error annotation, Université Catholique de Louvain Error Editor (UCLEE), but it functions to speed up the manual process of mark-up rather than assist with the identification of errors themselves (Dagneaux *et al*., 1998: 167–8). Rayson and Baron (2011) have also tested VARD's performance on learner corpora to see if it could be used to aid with error annotation. With training, as with the test on children's writing, they were able to achieve a very high precision figure of 90.8 percent but the recall level was again much lower at 23.4 percent. This is attributed, in this case, to the large number of real-world errors found in learner corpora which are not yet handled by VARD (Rayson and Baron, 2011: 122).

Studies of modern born-digital data have tended not to address the issue of normalising spelling variation: the studies have typically focussed on the innovations of language and orthography used (Hoffman, 2007; Ooi *et al*., 2007; and Tagg, 2009). In a study of Singaporean English blogs, Ooi *et al*. find the same kind of reduction in performance when using Wmatrix's semantic tagger as was found with historical corpora. In this study, Wmatrix is used to analyse two corpora, one containing blogs written by teenagers and the other by undergraduates. In total 3,712 types from the undergraduate corpus were left unclassified by the semantic tagger and a much higher

[4] Wmatrix is an online tool for the analysis and comparison of corpora. In particular, it facilitates part-of-speech and semantic annotation of corpora which can then be analysed using the same statistical procedure as is used for keywords.

11,137 types from the teenagers corpus. Rather than considering any kind of pre-processing to normalise the corpus, Ooi *et al.* (2007) suggest that corpus tools like Wmatrix need to evolve to be able to deal with this new emerging form of English. Tagg (2009) is also interested in the creative use of language but in this case with a corpus of text messages. Whilst Tagg chose not to normalise the language for this particular study, she does note that some experimentation with retraining VARD on this data suggests that its use could be possible with the corpus. In a later conference presentation, the latest version of VARD was used on the corpus with success (see Tagg *et al.*, 2010).

3. Spelling errors in the Teenage Health Freak Corpus

In order to gain an insight into the scale and type of spelling errors present in our corpus of health questions from teenagers, fifty messages were sampled at random from each year in the corpus (a total of 300 messages). These messages were checked manually for incorrect and unconventionally spelt words and the nature of each spelling error was analysed. The volume of spelling errors is summarised in Table 1. As the table shows, it is estimated that 7.6 percent of the words in the Teenage Health Freak Corpus could be incorrectly or unconventionally spelt. If the samples are representative of the whole corpus, this would amount to around 168,600 words. It is not common to report data regarding spelling variation in electronic communication corpora, but Tagg *et al.* (2010) report figures from a text message corpus which are similar to those found in the Teenage Health Freak data. As part of a test of automated spelling normalisation with VARD, a test sample is manually corrected. The sample consisted of 2,430 messages containing a total of 41,342 words. Of these words, 3,166 required standardisation meaning that 7.7 percent of the words in the sample were incorrectly or unconventionally spelt (Tagg *et al.*, 2010). In addition, both corpora revealed that while some messages contained many unconventionally spelt words, other messages showed no variation at all.

The errors can be classified into five broad categories: chat-style abbreviations; phonetic errors; typographical errors; deliberate errors for emphasis; and finally errors that did not fit into any of the other categories. Chat-style language includes abbreviations and acronyms which might be expected in text messages or instant messaging and in our samples these include the following transformations: *u > you*, *4 > for*, *cuz > because*, and *sum > some*. Typographical errors are errors which are most likely to be caused by mistyping and include the following examples: *iam > i am*, *resulst > results*, and *alchohl > alcohol*. Phonetic errors are words which can be pronounced reasonably in the same way as the original word and are less likely to have been caused by mistyping. In our sample, this includes: *probarbly > probably*, *egsisting > existing*, and *marige > marriage*. The

Year	No. of words	No. of errors	Percentage errors
2004	1,209	76	6.3
2005	1,403	116	8.3
2006	758	89	11.7
2007	898	55	6.1
2008	1,000	70	7.0
2009	871	60	6.9
All years	6,139	466	7.6

Table 1: Volume of spelling errors in 300 sample messages from the THF corpus

Error class	Total no. of occurrences
Typographical	257 (ignoring apostrophes 134)
Chat-style	125
Phonetic	83
Emphasis	3
None	1

Table 2: Classification of spelling errors in 300 sample messages from the THF corpus

next category of emphasis is less of an error and more of a manipulation of the language. However since this use of language poses the same problems for keyword analysis as the other spelling errors it is considered here. The emphasis category includes deliberate errors made for emphatic purposes. These are typically additions of letters as in these examples from our sample messages: *soooo* > *so*, and *yoooo* > *yo*. The final category includes everything that does not fit into any of the other categories; in our samples this includes only one example: *pencise* > *penis*.

Table 2 shows the number of errors in each category present in our sample messages. As the table shows, the vast majority of errors are accounted for by typographical errors and chat-style errors. The nature of chat-style errors should make correcting them relatively easy since there is a great deal of internal consistency with this type of language use (see Tagg, 2009: 136–8 for a summary of the use of such language in a text message corpus). An interesting observation on spelling in general, but which is particularly true of the use of chat-style abbreviations, is that there is huge difference between messages: some users avoid all chat-style language and others make full use of it. This may reflect the familiarity of the user with instant messaging, forum writing and perhaps text messaging, but also reflects the choice of register that is regarded as appropriate for addressing medical questions to Dr Ann with some selecting very formal registers and other much more informal ones. Compared to chat-style errors, the typographical errors are not as consistent. However, along with the phonetic errors the vast majority consist of a single omission, addition, deletion, substitution or transposition. This means that distance-based spelling detection algorithms may do a reasonable job at identifying these errors (Jurafsky and Martin, 2000: 144–6).

4. Spelling correction procedure

In order to try to correct the spellings in the corpus, an evaluation of VARD was conducted as this was the program that was used to regularise the spelling in Baron *et al.* (2009). At the time the project started the latest version of VARD was VARD 2.2. This version was still specifically aimed at the regularisation of historical corpora although it had also been used for other language varieties. Communication with VARD's creator, Alistair Baron, suggested that VARD 2.3 would be better suited to our data as many changes were being made in the automated processing to widen its application. During the course of the project VARD 2.3 was released but, unfortunately, the release date was too far into the project for it to be used to correct the spelling in the research reported here.

Rather than using a specifically designed tool, the spelling was corrected with the help of WordSmith Tools' keyword procedure (Scott, 2008). The main advantage of this procedure is that it is well-known amongst corpus linguists and is available in most of the standard software tools and would, therefore, be easily replicable in other corpora. The written component of the BNC was selected as the reference corpus as this minimises the number of non-standard tokens present in the reference corpus. As described above, the keyword procedure works to highlight words which are unusually frequent in the target corpus in comparison to their frequency in the reference corpus. Therefore, any words which are incorrectly spelt but occur in the same variant form frequently in our corpus were present in the keyword list. The resulting keyword list was reviewed manually and if the vast majority of variant forms of a word could be corrected to the same word these were corrected. This analysis was supported by the use of concordance lines to allow the intended word to be determined from the context. In cases where the error could not be corrected to a single word, the word was not corrected and the error remained. Once the whole list had been processed in this way, a script was used to mark the errors and provide their corrections. As the corpus was in XML, the spelling corrections were indicated using the

Text Encoding Initiative (TEI) tags, *choice*, *sic* and *corr*, so that the original spelling was also preserved. The resulting XML for each corrected word looks like this:

```
<choice>
        <sic>plz</sic>
        <corr>please</corr>
</choice>
```

Using this method, 2,732 types were corrected and this amounts to 88,542 tokens, or just over 50 percent of the predicted error total. The 2,732 incorrect types were corrected to just 900 types, suggesting there could be a significant impact on word frequencies. As missing apostrophes have an effect on word tokenisation and, therefore, on frequencies and keywords, they were also counted as spelling errors in this study. Forty-six of the corrected types only involved missing apostrophes, making the total number of types corrected that did not involve apostrophes 2,686. If the types that are only considered incorrect due to a missing apostrophe are ignored in the spelling correction process, the number of corrected tokens falls by 35,077 to only 53,465. In view of the large number of changes involving only apostrophes, the effects on keywords will be measured with and without the apostrophe corrections made.

5. The effects of spelling errors on key words

5.1 Methodology

The procedure for comparing the keyword lists was based on that used by Baron *et al.* (2009) for comparing the effects of non-standardised spelling in early modern English on the results of keyword analysis. In their paper, Baron *et al.* use two statistical measures, Spearman's rank correlation coefficient and Kendall's tau rank correlation coefficient, to compare two keyword lists. These statistical measures both focus on differences in ranks, but they are sensitive to different types of movement within the ranked lists that are being compared. As Spearman's rho uses the squared difference in the ranks between the two lists (Conover, 1999: 287) this measure is more sensitive to movements over a greater distance within the ranked lists and is relatively insensitive to large numbers of small movements within the ranks. In contrast, Kenall's tau is based on concordant and discordant pairs (Sprent and Smeeton, 2007: 318) and is, therefore, more sensitive to the volume of changes rather than the actual difference between the two ranks and, thus, the distance of the movement. In most cases, Spearman has been found to give a slightly higher figure than Kendall (Sprent and Smeeton, 2007: 323).

	Original spelling	Corrected (ignoring apostrophes)	Corrected spelling
Total keywords generated against BNC	3,608	1,934	1,900

Table 3: Total number of keywords generated

It should be noted here that Scott in his Frequently Asked Questions for WordSmith 5.0 advises that it is unsafe to rely on rank order in keyword lists (Scott, 2011). As log-likelihood is a statistical measure of significance, it is true that words are either key at the specified *p* value or are not key. However, in practice, researchers must often resort to relying on the top N keywords and, in this regard, changes in rank caused by spelling errors or a change in reference corpus could have an impact on the focus of the research. Changes in rank order also give a broad overview of the impact of such changes on the keyword lists. The procedure used is outlined below.

- Generate keyword lists to compare using WordSmith Tools and specified parameters for corrected and uncorrected spelling;
- Remove any words not present in both of the resulting keyword lists;
- Rank remaining entries in both lists from 1 to *n* (*n* will be the same for both lists); and,
- Use the ranks as the input to correlation graphs and statistical procedures.

5.2 Effect on keyword lists

The BNC was chosen as the reference corpus to test the effect of spelling variants on the keyword lists. The keywords used in this section were all generated using WordSmith Tools (Scott, 2008) using the log-likelihood statistical measure. The minimum frequency threshold was five and the *p* value used was 0.000001. Keywords were generated against the reference corpus for the Teenage Health Freak Corpus based on the original spelling, the fully corrected spelling and the corrected spelling ignoring missing apostrophes. The first thing to consider is the number of keywords generated for each version of the corpus. These are shown in Table 3.

By correcting the spelling in the corpus, the number of keywords generated is reduced by over 1,500. An examination of the words that are key only when the spelling errors have been corrected show that they contain medical or medical-related terms. In all of these cases, however, they occur

Corrected spelling		Original spelling	
Spelling	Rank in list	Spelling	Rank in list
transsexual	952	*transexual*	1,038
achy	1,492	*achey*	3,415
tonsillitis	1,498	*tonsillitis*	3,389
bingeing	1,623	*binging*	2,466
syphilis	1,658	*syphilis*	1,673
dizziness	1,609	*dizziness*	2,885
oestrogen	1,731	*estrogen*	2,048
tetanus	1,778	*tetnus*	2,523
disease	1,850	*disese/diseas*	2,885/3,415
lymph	1,898	*lymphnodes*	2,523

Table 4: Medical terms from the keyword list based on the corrected corpus and their equivalents in the original corpus

Corrected spelling		Original spelling	
Spelling	Rank in list	Spelling	Rank in list
deodorant	788	*deodrant/deoderant/deodorant*	1,313/1,784/3,526
accidentally	903	*accidently/accidentaly*	887/2,885
regularly	980	*regulary/regulaly*	672/2,885
noticeable	991	*noticable/noticible*	795/2,279

Table 5: Words that in the top 1,000 keywords from the corrected corpus and their equivalents in the original corpus

in some alternatively spelt form in the keyword-list based on the uncorrected spelling so they are not lost entirely even when the spelling is not corrected. Also, as Table 4 shows, these words occur a long way down the keyword lists for both the original and the corrected spelling.

The other words that are only present in the corrected spelling list also appear a long way down the keyword list. Only four examples occur in the top 1,000 keywords and these also occur in various spellings in the keyword list generated from the original spelling. These words are shown in Table 5. In all of these cases, correcting the spelling from what are typically

Spelling variant	Frequency	G2	Rank
embarrased	125	947.95	368
embarassed	108	801.61	413
embaressed	63	483.62	578
embarresed	60	460.59	589
embrassed	21	161.21	1,104
embarrassed	125	156.97	1,134
embarased	14	107.47	1,416
embarrsed	11	84.44	1,673
embaresed	9	69.09	1,907
embarrised	9	69.09	1,907
embarested	5	38.38	2,885
embarised	5	38.38	2,885
embarresd	5	38.38	2,885
embarsed	5	38.38	2,885
imbarrased	5	38.38	2,885

Table 6: Variant spellings of *embarrassed* in the corpus

multiple incorrect forms when added together means that the frequency becomes high enough for the correctly spelt version to be considered key. Only *deodorant* occurs with the correct spelling in both keyword lists, but the correctly spelt version occurs considerably lower in the original spelling list than the incorrect spellings do.

The keywords which occur only in the list generated with original spelling are predominantly examples of non-standard spelling and, with the exception of some of the standard chat-style abbreviations, they tend to occur towards the bottom end of the keyword list. There would, however, be many more keywords to analyse had the spelling not been corrected. In some cases, the correction of the spelling also makes a significant difference to the frequency count for the word and, therefore, its rank in the keyword list. An example of such a word is *embarrassed* which has multiple spellings, as can be seen in Table 6.

In the version of the corpus with the corrected spelling, the word frequency for *embarrassed* is 596 making it 171st in the keyword list rank with a G2 value of 2,206.33. This is higher than the highest entry in the

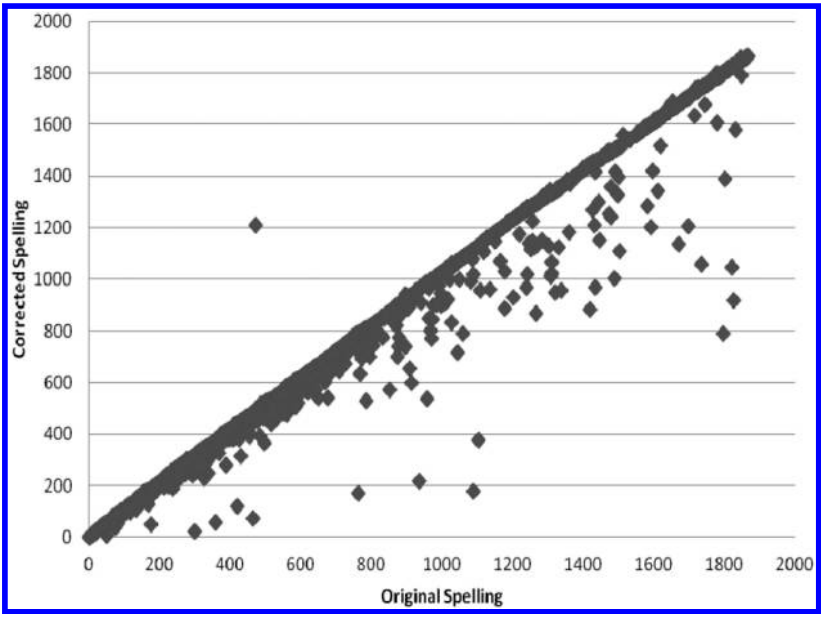

Figure 1: Correlation graph for fully corrected spelling against original spelling

original spelling keyword list where *embarrased* is ranked 368th with a G2 value of 947.95 and considerably higher than the correctly spelled version which is ranked 1,134 with a G2 value of only 156.97. It seems then, that even if correctly spelt keywords are not lost altogether when the spelling errors remain in the corpus, the spelling variation could make a difference to the ranks of the keywords. This was the hypothesis which formed the basis of Baron *et al.*'s (2009) investigation into spelling variants in early modern English and will also form the basis of our investigation.

The Spearman rank correlation coefficient and the Kendall correlation coefficient were calculated between two pairs of keyword lists all generated using the BNC as the reference corpus. The pairs were the original spelling and the fully corrected spelling, and the original spelling against the corrected spelling ignoring missing apostrophes.[5] An indication of the type of correlation present in the data can be seen in Figure 1 and Figure 2. These graphs suggest a very strong positive correlation in the

[5] Butler (1985: 147) states that Spearman's *r* should not be used if there are 'a large number of tied ranks'. In place of Spearman's *r* the ranks should be used as input to the Pearson product-moment correlation coefficient (the result of this calculation is, however, still known as Spearman's *r*). In both cases, the number of tied ranks here is less than 25 percent of the total pairs in the study and with no indication of what a 'large number' might be it was decided to calculate the statistic using both methods for completeness. In our data, there proved to be no difference in the calculations until the seventh or eighth decimal place so it was decided to report the results from the straight Spearman's *r* calculation for simplicity.

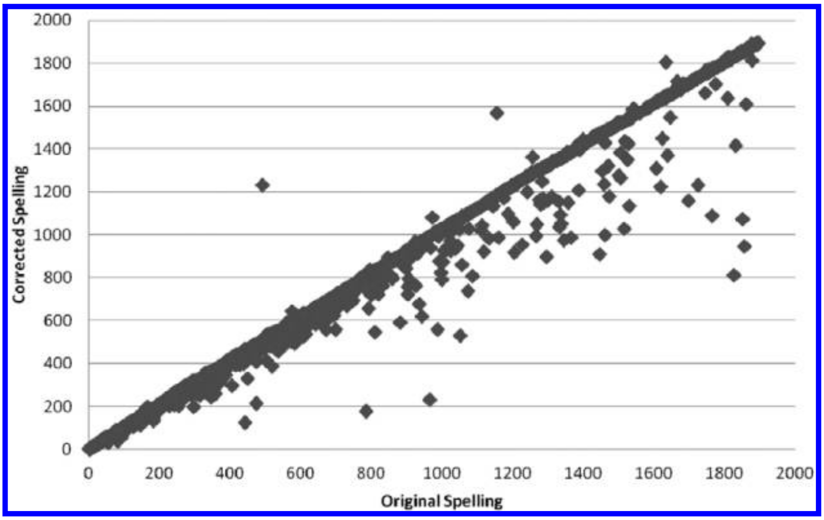

Figure 2: Correlation graph for corrected spelling ignoring missing apostrophes against original spelling

keyword ranks. Anything above the perfect line of correlation represents words that are so consistently spelt incorrectly in the Teenage Health Freak Corpus that they appear very high in the keyword list of uncorrected words (since the incorrectly spelt form does not occur in the reference corpus). When the spelling is corrected, their rank in the keyword list drops as a genuine comparison is made with the correctly spelt word in the reference corpus. The strongest example of this in the corpus is the word *conscious* (point 473, 1,209) which is misspelt as *concious* so frequently in the corpus that it ranks very highly in the keyword list until it is corrected and, therefore, compared with the actual frequency of the word in the reference corpus. Anything below the perfect line of correlation represents words where, when all the spelling variants are corrected, they collectively make a big enough difference to the frequency for them to move higher up the keyword list for the corrected text, as with the example of *embarrassed*. It can also be seen from the graphs that the later items towards the bottom of the ranked list are more affected than those towards the top.

Both Spearman's rho and Kendall's tau return a number between –1 (a perfect negative correlation) and +1 (a perfect positive correlation) with 0 indicating no correlation at all. The results for these comparisons can be seen in Table 7. These numbers are both very close to +1 suggesting a very positive correlation between the ranks of the keyword lists. Not correcting missing apostrophes leads to a slightly better correlation than the fully corrected text, but both are very high. The critical tables for Spearman's *r* stop at 30 degrees of freedom and since we have many more than thirty samples, the result of Spearman's *r* can be converted to a *t*-value and that checked with N-2

	Fully corrected spelling against original spelling	Corrected spelling ignoring apostrophes against original spelling
Kendall's Tau	0.961	0.963
Spearman's *r*	0.989	0.991
Spearman's *r* converted to t-value	293	315

Table 7: Results from Spearman's rank correlation coefficient and its conversion to t-value

degrees of freedom against the critical values for *t*. In both cases, our degrees of freedom (1849 and 1893) are off the scale for *p* values of *t* value but both scores are much greater than the values needed at the highest degrees of freedom available and, therefore, we can conclude that there is a significantly high correlation to suggest that the spelling variants in the corpus make little difference to the ranks of the keyword lists. These figures are certainly higher than those reported by Baron *et al.* whose overall Spearman's rho score for the Innsbruck Letter corpus in manually standardised and original form was 0.705 and Kendall's tau, 0.530. This is to be expected given the much higher rates of variation present in the Innsbruck Letter corpus. When comparing automatically standardised (using VARD) and original texts over several decades both correlation scores were much higher and above 0.9 by the 1600s. Figures are only reported in graphical form in this part of the study, so it is not possible to give the highest correlation reached in that study; however, Spearman is very close to +1 and Kendall, over 0.9 (Baron *et al.*, 2009: 58).

6. Conclusion

This case study demonstrates that, even with born-digital data, where the instances of spelling errors and non-standard spelling are likely to be higher than in other written corpora (i.e., newspaper or other print media corpora), spelling variation does not necessarily have a significant impact on the keywords that are generated. Although the volume of keywords generated was much greater with the uncorrected version of the corpus, the differences in the ranks of shared words were very small. The words found to be key in the uncorrected keyword list and not in the corrected keyword list tended to appear towards the bottom. They would, therefore, present less of a problem for research that focusses on words towards the top of the list. In general, we can conclude that while correcting the spelling made a large difference to the ranks of some words (for example, *embarrassed*, noted above) the overall effect on the keyword ranks is small. So, together with the fact that even the correctly spelt technical terminology only present in the corrected keyword list still appears in some form on the uncorrected list, this suggests that for born-digital data the correction of spelling is not necessarily essential before corpus analysis can take place. Of course, for other types of corpus analysis on born digital data, such part-of-speech and semantic tagging that prompted the creation of a tool like VARD, where spelling variation is a much more significant problem, might suffer the same loss of accuracy as has been found in historical corpora.

More work is needed to establish whether the figure of 7 percent for non-standard spelling in the Teenage Health Freak Corpus and in Tagg's text message corpus holds true for other CMC corpora. This will be a key factor in assessing how generalisable the results of this study are. While the range of messages in the Teenage Health Freak Corpus is quite wide, there is a focus on health concerns which is unlikely to be found in more general CMC corpora, and this topical focus may also contribute to the results seen in this study. The same research should be repeated on more general CMC corpora to support the suggestion that spelling correction is not of great import in keyword studies of such data. The effort involved in such an investigation is somewhat lessened, now, in the era of Internet corpora and 'big data'. In addition, it is possible that a more accurate method of error detection and correction might show a much greater variation in key word rank, since, in this study, the level of correction was around 50 percent. Now that the VARD has been adapted to deal well with spelling variation of non-standard modern English, specifically CMC, it is possible that repeating this study on a version of the corpus that has been automatically corrected using VARD would shed more light on the impact of spelling on keyword rank. However, overall we suggest that this study demonstrates that while non-standard spelling should be a concern to corpus researchers working with born digital data, they can be confident in carrying out initial keyword based investigations on the uncorrected data.

References

Archer, D., T. McEnery, P. Rayson and A. Hardie. 2003. 'Developing an automated semantic analysis system for Early Modern English' in D. Archer, P. Rayson, A. Wilson and T. McEnery (eds) Proceedings of the Corpus Linguistics 2003 Conference. University Centre for Computer Corpus Research on Language: Lancaster University, pp. 22–31. Accessed November 2012, at: http://ucrel.lancs.ac.uk/publications/CL2003/papers/archer.pdf

Baker, P. 2006. 'The question is, how cruel is it?' Keywords, Foxhunting and the House of Commons. Word Frequency and Keyword Extraction, AHRC ICT Methods Network Expert Seminar on Linguistics. Lancaster University. Available online, at: http://www.arts-humanities.net/system/files/es1_07baker.pdf (accessed October 2011).

Baron, N.S. 2008. Always On: Language in an Online and Mobile World. Oxford: Oxford University Press.

Baron, A. and P. Rayson. 2009. 'Automatic standardization of texts containing spelling variation, how much training data do you need?', M. Mahlberg, V. González-Díaz and C. Smith (eds) Proceedings of the Corpus Linguistics Conference 2009. Accessed October 2011, at: http://ucrel.lancs.ac.uk/publications/cl2009/314_FullPaper.pdf

Baron, A., P. Rayson and D. Archer. 2009. 'Word frequency and keyword statistics in historical corpus linguistics', Anglistik: International Journal of English Studies 20 (1), pp. 41–67.

Butler, C. 1985. Statistics in Linguistics. Oxford: Blackwell.

Conover, W.J. 1999. Practical Nonparametric Statistics. (Third edition.) New York: John Wiley and Sons.

Crystal, D. 2006. Language and the Internet. (Second edition.) Cambridge: Cambridge University Press.

Crystal, D. 2011. Internet Linguistics: A Student Guide. Oxford: Routledge.

Dagneaux, E., S. Denness and S. Granger. 1998. 'Computer-aided error analysis', System 26 (2), pp. 163–74.

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 (4), pp. 304–11.

Hoffman, S. 2007. 'Processing Internet-derived text: creating a corpus of Usenet messages', Literary and Linguistic Computing 22 (2), pp. 163–74.

Hofland, K. and S. Johansson. 1982. Word Frequencies in British and American English. Bergen: The Norwegian Computing Centre for the Humanities.

Jurafsky, D. and J.H. Martin. 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. (International edition.) New Jersey: Prentice Hall.

Kay, C. 2006. Issues for Historical and Regional Corpora: First Catch Your Word. Word Frequency and Keyword Extraction, AHRC ICT Methods Network Expert Seminar on Linguistics. Lancaster University. Accessed October 2011, at: http://www.arts-humanities.net/system/files/es1_04kay.pdf

Kilgarriff, A. 2001. 'Comparing corpora', International Journal of Corpus Linguistics 6 (1), pp. 1–37.

Ooi, V.B.Y., P.K.W. Tan and A.K.L. Chiang. 2007. 'Analyzing personal weblogs in Singapore English: the Wmatrix approach', Studies in Variation, Contacts and Change in English 2. Accessed October 2011, at: http://www.helsinki.fi/varieng/journal/volumes/02/ooi_et_al/

Paquot, M. and Y. Bestgen 2009. 'Distinctive words in academic writing: a comparison of three statistical tests for keyword extraction' in A. Jucker, D. Schreier and M. Hundt (eds) Corpora: Pragmatics and Discourse, pp. 247–69. Amsterdam, Rodopi. Accessed October 2011, at: http://sites.uclouvain.be/cecl/archives/PAQUOT_BESTGEN_2009_Distinctive_words_in_academic_writing_ICAME2008.pdf

Rayson, P. 2008. 'From key words to key semantic domains', International Journal of Corpus Linguistics 13 (4), pp. 519–49.

Rayson, P. 2009. 'Wmatrix: a web-based corpus processing environment', Computing Department, Lancaster University. Accessed October 2011, at: http://ucrel.lancs.ac.uk/wmatrix/

Rayson, P. and A. Baron. 2011. 'Automatic error tagging of spelling mistakes in learner corpora' in F. Meunier, S. De Cock, G. Gilquin and M. Paquot (eds) A Taste for Corpora: In Honour of Sylviane Granger, pp. 109–26. Amsterdam: John Benjamins.

Rayson, P., D. Archer, A. Baron and N. Smith. 2007. 'Tagging historical corpora – the problem of spelling variation' in Proceedings of Digital Historical Corpora, Dagstuhl-Seminar 06491. Accessed October 2011, at: http://www.comp.lancs.ac.uk/~paul/publications/rabs_extAbs_dagstuhl06.pdf

Scott, M. 1997. 'PC analysis of key words – and key key words', System 25 (1), pp. 1–13.

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

Scott, M. 2011. WordSmith 5.0 Answers to FAQs. Accessed October 2011, at: http://www.lexically.net/wordsmith/version5/faqs/answers.htm#different_keynesses

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

Sprent, P. and N.C. Smeeton. 2007. Applied Nonparametric Statistical Methods. (Fourth edition.) London: Taylor and Francis.

Stenström, A., G. Andersen and I.K. Hasund. 2002. Trend in Teenage Talk: Corpus Compilation, Analysis and Findings. Amsterdam and Philadelphia: John Benjamins.

Tagg, C. 2009. A Corpus Linguistics Study of SMS Text Messaging. Unpublished PhD thesis. Birmingham: University of Birmingham.

Tagg, C., A. Baron and P. Rayson. 2010. ‘I didn’t spel that wrong did i. oops’: analysis and standardisation of SMS spelling variation. ICAME 2010. Accessed November 2012, at: http://comp.eprints.lancs.ac.uk/2310/1/CorTxt_and_VARD_-_ICAME_presentation-Final.pdf