# Querying Keywords : Questions of Difference, Frequency, and Sense in Keywords Analysis

Paul Baker

*Journal of English Linguistics* 2004 32: 346
DOI: 10.1177/0075424204269894

# Querying Keywords

## Questions of Difference, Frequency, and Sense in Keywords Analysis

PAUL BAKER

*Lancaster University, United Kingdom*

The corpus analysis program WordSmith Tools allows two corpora to be compared against each other in order to acquire a list of keywords: words that appear statistically more often in one text than the other. This article examines some potential problems of analyzing keywords and ways to overcome them. A keywords analysis may force the researcher to overplay differences rather than similarities, and it focuses on difference at the lexical rather than the semantic, grammatical, or pragmatic levels. An examination of dispersion patterns, concordances, and key clusters are useful supplementary forms of analysis. The use of annotated data to derive key categories or grouping like words into similar groups is also helpful.

***Keywords:*** *keywords; corpus; methodology; frequency; discourse*

In recent years, corpora have begun to play an important role in discourse analysis (e.g., Teubert 2000; Krishnamurthy 1996; Piper 2000; Fairclough 2000; Flowerdew 1997). Corpus-based analysis allows researchers to identify more or less objectively widespread patterns of naturally occurring language and rare instances, both of which may be overlooked in a small-scale analysis. Corpus linguists have access to a range of procedures that can be implemented in the analysis of text (e.g., collocations, frequency lists, dispersion plots, concordances). One statistical procedure that has proven to be popular involves the creation of *keyword lists*. The earliest writers who referred to keywords intuitively focused on words that they believed embodied important concepts that reflected societal or cultural concerns (e.g., Firth 1957; Williams 1983). However, taking a corpus linguistics approach, Scott (1999) derives keywords via a specific statistical process. A word is key if it occurs in a text at least as many times as a user has specified as a minimum frequency, and its frequency in the text when compared with its frequency in a reference corpus is such that its statistical probability as computed by an appropriate procedure (e.g., Scott allows users to specify either Dunning's log-likelihood score

**AUTHOR'S NOTE:** The author would like to thank Mike Scott and two anonymous reviewers, all of whom provided comments to earlier drafts of this article and improved it considerably.

Journal of English Linguistics, Vol. 32 / No. 4, December 2004 346-359
DOI: 10.1177/0075424204269894

[1993], or the chi-squared test) is smaller or equal to a *p* value specified by a user. Scott's definition of keywords is therefore not based on concepts that are subjectively viewed as important to culture but allows for any word potentially to be key if it occurs frequently enough when compared to a reference corpus. Scott notes that three types of keywords are often found: proper nouns; keywords that human beings would recognize as key and are indicators of the "aboutness" of a particular text; and finally, high-frequency words such as *because*, *shall*, or *already*, which may be indicators of style, rather than aboutness.

Scott's WordSmith suite of tools allows a frequency list taken from one file (or corpus) to be compared against the frequency list of another corpus (either a larger "reference" corpus or one that is of a similar size). When two texts of equal size are compared, two corresponding keyword lists are produced, usually of a similar length. When a smaller text is compared with a larger text, only the words that are key in the smaller text appear, alongside a smaller number of negative keywords: words that have appeared in the smaller text *less* often than would be expected from their appearance in the reference corpus. A keyword list is usually presented in order of keyness: the most statistically significant or "strongest" keywords appearing first. An examination of the keywords that occur when two corpora are compared together should reveal the most significant lexical differences between them, in terms of aboutness and style.

Researchers have used keyword lists in order to gain descriptive accounts of particular genres. For example, Tribble (2000) derives a keyword list from comparing a corpus of romantic fiction with a general corpus and finds evidence to suggest features of spoken language in romantic fiction such as more first and second person pronouns and proper nouns, and fewer complex noun phrases (the words *the* and *of* were negative keywords).

Keywords can also be useful in helping to spot traces of discourse within language. While the term *discourse* has multiple meanings, I use it here to refer to a "system of statements which constructs an object" (Parker 1992, 5). Discourse is further categorized by Burr (1995, 48) as "a set of meanings, metaphors, representations, images, stories, statements and so on that in some way together produce a particular version of events." Parker and Burman (1993, 156) point out that discourses emerge as much through our work of reading as from the text. Keywords will therefore not reveal discourses but will direct the researcher to important concepts in a text (in relation to other texts) that may help to highlight the existence of types of (embedded) discourse or ideology. Examining how such keywords occur in context and which grammatical categories they appear in, and looking at their common patterns of co-occurrence should therefore be revealing. For example, Johnson, Culpeper, and Suhr (2003) investigate keywords in a preselected set of British newspaper articles across a five-year period. All of the articles contain reference to the concept of *political correctness* in some way. They find that the stron-

gest keywords differed over time as focus around political correctness shifted from a range of minority identities and the media in 1994 to racism in 1999. Fairclough (2000) compares a corpus of "New Labour" (i.e., from the Blair period of government) documents, speeches, and newspaper articles with a corpus of older Labour texts and subsequently carries out analyses to show how Labour's ideological stance had changed over time to stress business interests and competition. New Labour keywords included *partnership*, *new*, *deliver*, *deal*, *business*, and *promote*.

Keywords are therefore an extremely rapid and useful way of directing researchers to elements in texts that are unusually frequent (or infrequent), helping to remove researcher bias and paving the way for more complex analyses of linguistic phenomena. However, it is essential to realize that a keyword list provides the researcher only with language patterns, which must be interpreted in order to answer specific research questions. This issue focuses on some of the matters of interpretation that were brought to light when a keyword analysis was carried out in order to determine the differences between two large bodies of text of equal size. It is not the intention of this article to denigrate keyword analysis, but rather, to make researchers aware of possible areas of over- or underinterpretation and suggest ways of ameliorating these issues.

## Keywords

The data used in the analysis consisted of one million words of gay male erotic narratives and one million words of lesbian erotic narratives collected from the Web site www.nifty.org.[1] These sets of narratives, each containing texts from many authors, were chosen in order to compare discourses of gender in these two sets of texts. Because gay narratives mainly involve gay men and lesbian narratives involve lesbian women, it is relatively easy to compare the different ways that gender is constructed between them. Erotic narratives often detail idealistic, surreal events that are unrepresentative of most people's experience. Differences in the vocabulary of these texts are therefore not reflective of "real-life" differences in how people really think, talk, and act but are more indicative of how people *believe* they should behave in erotic situations. Erotic narratives could therefore function as instructional discourses in the same way that advertisements instruct heterosexual women to desire taller boyfriends (Goffman 1976; Eckert 2002, 109). It was therefore the intention to explore how identity is constructed differently in each of the erotic genres and the discourses that the authors draw on, in order to create recognizably (or not) gendered characters.

There were roughly equal numbers of texts under examination (354 gay texts vs. 342 lesbian texts), with the gay narratives being slightly longer on average than the lesbian texts (the mean text length being 2,898 words vs. 2,775 words, respectively). An examination of the standardized type token ratio, average word length,

and average sentence length also showed the two sets of data to be remarkably similar.[2] The "cutoff" point for determining whether a word was a keyword was whether the difference in frequency between the two files was significant at a level less than $p = .000001$ using the log-likelihood statistical test.[3] Even at this extremely high level, a total of 1,055 keywords were found, 504 that occurred significantly more often in the gay texts and 551 that occurred significantly more in the lesbian texts. In this article I am not so much focusing on the discourses that were elicited but more on the method of analysis that was used to find them.

## Difference

The first observation that should be made when comparing corpora to elicit keywords is that the comparison will not reveal words that would normally be keywords when compared to other genres (e.g., nonerotic narratives, if these words are keywords in both sets of files). So, for example, it is likely that a word such as *sex* would be key in most types of erotic texts when compared to a corpus of general English, but this will not be revealed in this analysis. Therefore, a keyword analysis will focus only on lexical differences, not lexical similarities. Such a feature of WordSmith is not necessarily problematic, depending on the researcher's focus, but it may result in the researcher making claims about differences while neglecting similarities to the point that differences are overemphasized. For example, if the word *large* appears in a keyword list, we may theorize that this reveals an important difference—that one genre or set of texts is concerned with size much more than the other. However, other words, such as *big*, *huge*, *enormous*, *small*, *tiny*, and so on, may occur with equal numbers in both sets of texts, suggesting that the overall pattern is that size per se is not particularly important, but for some reason use of the word *large* is. Care must therefore be taken when generalizing beyond the lexical level.

Therefore, one way of analyzing similarities between texts is to carry out comparisons on more than two sets of data. For example, the gay and lesbian narratives were compared with the Frown (Freiberg-Brown) corpus of general American English, taken from the same time period. This gave two further lists of keywords, which could then be compared against each other. Table 1 shows how keywords associated with verbs that showed communication (e.g., *said*, *replied*) and facial reactions (*blush*, *smile*) were then categorized. The gay keywords were key when compared to both the lesbian texts *and* the Frown corpus. The lesbian keywords were key when compared to the gay texts and Frown, whereas in the final row, words that were not key when the gay and lesbian texts were compared with each other, but were key when each were compared with Frown, are shown.

Table 1 therefore shows us differences as well as similarities between the key communicative verbs in the gay and lesbian narratives. We could then make further

**TABLE 1**
Keywords Related to Communication and Facial Reactions

| Genre | Keywords |
|---|---|
| Gay | grinned, groaned, grunted |
| Lesbian | asks, asked, blush, giggle, giggled, giggling, replied, smile, smiled |
| Both | answered, begged, begging, blushed, cried, decided, gasp, gasping, grin, heard, laughed, laughing, moan, moaned, moaning, moans, panting, reply, responded, said, scream, screamed, sigh, smiling, tease, teased, teasing, tell, told, whispered |

investigations based on the fact that in the gay texts people appear to grunt, groan, and grin, whereas in the lesbian texts they giggle, blush, and smile. In addition, communicative verbs that signify a range of reactive states (*moan*, *tease*, *beg*) appear to be key in both sets of texts (when compared to the Frown corpus), and it may be interesting to examine why different forms of the same lemma are not consistently key across each text type; for example, *grinned* is key in the gay texts when compared against the lesbian texts, but *grin* is key in both the gay and lesbian texts when compared against the Frown corpus.

## Frequency

A second problem with a keywords analysis is particularly salient when working with groups of multiple texts. There were about 350 individual texts in each of the gay and lesbian corpora that were used. Therefore, potentially, a word may be key but will only occur in a very small number of texts. For example, the word *wuz* is a gay keyword, being used as a nonstandard spelling of *was* (occurring thirty-two times in the gay texts and never in the lesbian texts). However, all of the cases of *wuz* are restricted to one narrative, which suggests that this word is key because of a single author's use of a word in a specific case, rather than being something that indicates a general difference in language use.

One way to counter this problem is to consider what Scott calls "key keywords." A key keywords list reveals how many texts a keyword appears in as key. For example, in the lesbian texts the word *herself* is a keyword. It occurs 1,168 times across 216 texts, although when each lesbian text is analyzed separately against the gay texts as a whole, it only occurs as a keyword in 91 of them. Table 2 shows the top 20 key keywords for the lesbian and gay male texts.

However, one problem with key keywords is that the strongest words tend to reveal the most obvious differences; in this case, they reveal keywords that we could have probably made a good educated guess at in advance. So the lesbian texts contain more female pronouns and more words relating to female parts of the body and clothing. There are some interesting points of interest here: for example, the use of

**TABLE 2**
Top 20 Key Keywords in the Gay and Lesbian Texts

| Lesbian Key Keywords | | | Gay Key Keywords | | |
|---|---|---|---|---|---|
| Word | Key in Number of Texts | Overall Frequency | Word | Key in Number of Texts | Overall Frequency |
| her | 327 | 33,708 | his | 334 | 24,516 |
| she | 320 | 23,224 | he | 328 | 20,647 |
| breasts | 136 | 1,339 | him | 249 | 7,574 |
| pussy | 123 | 2,021 | cock | 203 | 4,914 |
| clit | 109 | 747 | dick | 109 | 1,834 |
| herself | 91 | 1,168 | I | 103 | 36,975 |
| woman | 61 | 1,433 | balls | 87 | 1,059 |
| panties | 54 | 649 | my | 83 | 20,826 |
| bra | 48 | 468 | ass | 51 | 2,151 |
| cunt | 44 | 531 | himself | 43 | 657 |
| skirt | 42 | 283 | man | 37 | 1,262 |
| the | 41 | 39,861 | load | 35 | 490 |
| breast | 34 | 431 | guy | 33 | 777 |
| my | 32 | 16,398 | we | 31 | 5,158 |
| girl | 32 | 789 | cum | 31 | 1,404 |
| dildo | 31 | 328 | shaft | 28 | 521 |
| she'd | 30 | 497 | he'd | 26 | 437 |
| you | 29 | 8,448 | hole | 24 | 717 |
| lesbian | 26 | 320 | me | 21 | 10,190 |
| girls | 26 | 501 | penis | 21 | 396 |

nonstandard sexual terms rather than formal terms and the use of more first-person pronouns in the gay male texts, but on the whole, the key keywords list confirms expectations, rather than revealing hidden patterns. By the time that the twentieth words in the list are reached, they are only key in 20 or so texts out of a possible 350, so the sense of looking at any more key keywords than this is debatable.

Therefore, it would be useful to find a way that combines the strengths of key keywords with those of keywords but is neither too general or exaggerates the importance of a word based on the eccentricities of individual files. Two suggestions are proposed. First, when analyzing individual keywords, it is possible to ascertain how many files they occur in and to present or take into account this information in addition to the frequency count. For example, the word *wife* occurs 223-81 (223 times in 81 texts). However, one problem with this strategy is in establishing cutoff points. One could specify, for example, that a keyword has to occur at least *x* times and/or in *y* or more of the individual texts in a corpus, relative to its frequency, in order for it to be viewed as a representative keyword. This relates to a more general concern about keyword analyses in that there is no popular consensus about cutoff points. So researchers who derive a list of keywords may be unsure about how

many words they should examine or how small to specify the *p* value. Scott (1999) says that "with keywords where the notion of risk is less important than that of selectivity, you may wish to set a comparatively low p value threshold such as 0.000001 [1 in one million] so as to obtain fewer keywords." As different researchers will work with different types of corpora and different research questions, reaching a consensus over cutoff points is unlikely and possibly undesirable in any case. For the sake of this analysis, my weakest keyword, by setting *p* at .000001, was *bloated*, which occurred eighteen times in the gay texts and zero times in the lesbian texts. I also discarded keywords that only appeared in fewer than ten narratives, which suggests that they are not particularly representative of that genre. While *bloated* was infrequent, it did occur in thirteen separate gay texts, demonstrating that at least it had a relatively even distribution. These cutoff points were derived from testing a number of different formulas and then settling on one that was felt to be a good compromise between giving enough words to analyze, but not so many that the representativeness of a keyword across a range of individual files became negligible.[4]

A second solution, based not on placing restrictions on frequencies but on a more inclusive and subjective analysis, could be to carry out a close analysis of concordances and collocations of individual keywords and then group them together according to the purposes that they serve in contributing to particular discourses. For example, the gay keywords *sweat*, *smelly*, *beer*, *football*, *duty*, *army*, and *military* all contributed toward a discourse of hypermasculinity within the gay narratives. Some of these keywords have semantic links—for example, *army* and *military*—but it is only by looking at their overall functions in the texts that stronger links can be made between them (e.g., there is no immediately obvious link between the words *smelly* and *military*). Only through a concordance-based analysis of these words was it made clear that *smelly* was consistently used in a way to construct hypermasculine identities in the gay texts.

In addition, examining both the gay and lesbian sets of keywords together is a useful strategy. For example, where *beer* was a gay keyword, *wine* was a lesbian keyword. Both words served the same purpose in their respective texts—the consumption of alcoholic drinks was important in the early parts of the narratives in that this enabled characters to lose inhibitions. However, these drinks also helped to construct gender identities, with beer-drinking gay male characters displaying a traditionally working-class masculine identity, while wine was associated with a more sophisticated lesbian identity.

Another aspect of a keyword analysis is that relatively low frequency words can be revealed as being key. As mentioned previously, *bloated* occurred eighteen times in the gay texts and never in the lesbian texts and was therefore (just) flagged as key. Depending on what the researcher is looking for, low-frequency keywords may be welcome or not. Changing the *p* value to a lower number would result in

**TABLE 3**
Cumulative Keyness of Words Indexing Size

| Word | Frequency in Gay Texts | Frequency in Lesbian Texts | Keyness |
|---|---|---|---|
| bloated | 18 | 0 | 25.1 |
| fat | 165 | 39 | 85.0 |
| thick | 474 | 132 | 208.5 |
| huge | 293 | 99 | 102.4 |
| massive | 99 | 27 | 44.5 |
| bulging | 48 | 7 | 34.7 |
| total | 1,097 | 304 | 481.27 |

*bloated* not appearing as a keyword. In addition, specifying a higher cutoff point for the minimum frequency with which a word must occur before it can be key would remove *bloated* from the list of keywords. However, low-frequency keywords may be useful in that they can often be combined into similar categories of meaning or function. For example, as well as *bloated*, the words *fat*, *thick*, *huge*, *massive*, and *bulging* are also key in the gay texts, all serving very similar uses.

When the frequencies of these words are added together, their cumulative keyness increases (see Table 3). Note that cumulative keyness does not mean simply taking a combined total of the keyness scores of all of the words but requires that keyness is recalculated using a log-likelihood or chi-square calculation based on the frequency data and the relative size of each corpus—in this case it is 481.27, rather than 500.2 (which is what would have been achieved by simply adding up the numbers in the final column of Table 3). WordSmith does not offer a way of calculating an individual log-likelihood score, although Web sites exist where numbers can be entered into a form and the calculation is carried out automatically.[5]

Therefore, a key category was found by conducting separate analyses of individual keywords in order to note their general functions and then combining words together in ways that made sense. However, one problem with combining words into conceptual groups is that it is a subjective process. Some groups may suggest themselves more clearly to the researcher than others, and it may be difficult to know how to specify a cutoff point. Carrying out concordance-based analyses of individual keywords should ensure that the researcher first has an understanding of what such words are used to achieve in a text, before erroneously combining words that may appear similar at face value. Like many other forms of linguistic analysis, researchers are required to develop skills of interpretation, which suggests that corpus-based research is not a merely quantitative form of analysis.

Another, more mechanical way of considering keywords as related groups is to use a form of semantic annotation, discussed in more detail below. This method should also help to identify important concepts consisting of low-frequency words.

## Sense

A related issue, which is potentially more far-reaching, is concerned with the fact that keywords only focus on lexical differences, rather than semantic, grammatical, or functional differences. For example, consider the word *sessions*. This word appears as a keyword in the gay narratives, occurring thirty-five times (as opposed to only three in the lesbian texts). It is also reasonably well distributed, appearing in twenty-five gay texts between one and three times in each. A concordance analysis reveals that *sessions* has a relatively specific meaning in the gay texts—in twenty-two out of its thirty-five cases it is used to refer to instances of sex between men, usually with little emotional commitment. In a smaller number of cases, it refers to therapy sessions, conference sessions, or study sessions. We could therefore conclude that in the gay narratives, when people use the word *sessions*, they are referring to sex rather than anything else. Arguably this does appear to be the case—even if all of the cases of *sessions* in the lesbian texts referred to sex, the number of cases where this happens in the gay texts would still be higher.

However, none of the three lesbian uses of *sessions* refer to sex; instead, they refer to psychiatrist sessions or exercise sessions in the gym. Although it is difficult to draw conclusions from such small frequencies, in the lesbian texts it seems we are less likely to find conceptualizations of sex as a session. Therefore, a closer analysis has shown that not only is a word key because it occurs more frequently, but it is key because it can occur more frequently *within a restricted set of meanings*.

Now consider a related word, *session*. This occurs fifty-five times in the gay texts and twenty-four times in the lesbian texts—meaning that although it still occurs more frequently in the gay texts, the difference is not statistically significant enough to make *session* a keyword. However, when we consider the senses of the word *session* in the two corpora, we find a similar pattern to that of *sessions*, which was key. Forty-nine cases of *session* in the gay texts refer to sex, whereas only twelve have sexual meanings in the lesbian texts. As a result, if only the sexual meanings of *session* are taken into account, it *would* count as a keyword. But because the original keywords calculation took into account the nonsexual meanings, *session* was not a keyword. It is therefore unlikely that the word *session* would have been brought to our attention at all, had it not been for the fact that its plural form occurred as a keyword. A simple keyword list therefore may obscure the fact that, in a text, certain *senses* of words can be key, but not others. As a result, a keyword analysis on plain unannotated text will only uncover keywords that indicate differences in lexical frequency, meaning that the researcher may overattend to such differences and subsequently overlook the more subtle cases based on word meaning such as *session*.

One strategy that some researchers have used is to lemmatize the data before calculating keywords. For example, Utka (2004) in his analysis of keywords in George

Orwell's *1984*, takes the one hundred most frequent lemmatized noun forms in the text and calculates keywords based on frequencies of lemmas, rather than individual word forms. Carrying out such a strategy on the gay and lesbian narratives would have enabled a more inclusive form of analysis as it most likely would have resulted in the lemma Session being key rather than just the word *sessions*. However, a lemma-based analysis may not always be a useful strategy as particular word forms can contain specific collocations or senses that would be lost when combining word forms together. For example, Stubbs (2001, 27-28) points out that *seek*, *seeks*, *seeking*, and *sought* have different sets of collocates, only some of which are shared.

Another option would be to carry out a form of annotation of the corpus before calculating keywords. For example, if we used a grammatical markup scheme such as the C5 tagset, which was used for annotating the British National Corpus (Garside and Smith 1997), we could distinguish between uses of words based on different grammatical functions—for example, gaze_NN1 (noun) vs. gaze_VVB (verb). We could also annotate a corpus according to semantic classifications, such as USAS (UCREL Semantic Analysis System) (Wilson and Thomas 1997). This semantic tagset was originally loosely based on McArthur's (1981) *Longman Lexicon of Contemporary English*. It has a multitier structure with twenty-one major discourse fields, subdivided, and with the possibility of further fine-grained subdivision in certain cases. In some cases, tags can be assigned a number of plus or minus codes to show where meaning resides on a binary or linear distinction. For example, the code T3 refers to "Time: Old, new and young; age," so the word *kids* is assigned T3– placing it at one end of a linear scale, whereas a word like *pensioner* would receive T3+.

However, tagging is time-consuming when carried out by hand, and it is likely to be error-prone when done automatically. In addition, such schemes may not be able to show all of the subtleties of word meaning, which often will not be made apparent until the word is analyzed via a concordance. According to the USAS scheme, *session* would always be tagged as "T1.3: Time: period," whether it occurs as a sexual session or a session with a psychiatrist as the word still has roughly the same surface meaning. Therefore, semantic taggers that make more fine-grained distinctions would be useful. However, more subtle semantic categories may also make it more difficult to group words automatically based on the same sense or function. For example, looking back at an earlier example, USAS categorized the words *bloated*, *fat*, *thick*, *huge*, and *massive* with the same general tag (N3 measurement), although beyond this, these words received more fine-grained distinctions: for example, *huge*, *bloated*, and *massive* were categorized as measurement of size, whereas *fat* was categorized as measurement of volume.

Another approach to keyness involves moving beyond the lexical level. Scott (1999) defines a word as "a sequence of valid characters with a word separator at

**TABLE 4**
Key Three-Word Lexical Bundles

| Lexical Bundle | Frequency in Gay Texts | Frequency in Lesbian Texts | Keyness |
|---|---|---|---|
| Key in the gay texts | | | |
| out of my | 258 | 113 | 59.1 |
| a couple of | 293 | 143 | 53.6 |
| up and down | 423 | 242 | 51.0 |
| back of my | 137 | 57 | 34.5 |
| I got a | 51 | 13 | 24.3 |
| Key in the lesbian texts | | | |
| asked in a | 0 | 19 | 26.2 |
| cum for me | 3 | 34 | 30.3 |
| the folds of | 0 | 25 | 34.5 |
| oh my god | 11 | 62 | 39.0 |
| glass of wine | 0 | 34 | 46.9 |

each end." However, there is no reason why keywords need to consist of single words. A further method of comparison can also be achieved by building word lists of two-, three-, and four-word "clusters" (Scott 1999), what Biber et al. (1999, 990) refer to as "lexical bundles" rather than single words. In this case, we would be looking at key clusters rather than keywords. Table 4 shows a list of key three-word clusters, which could be useful in helping researchers understand how individual keywords are used in context. Such a method is also useful in showing up key differences that are overlooked at the single word level: for example, while *session* may not occur as a keyword because of similar frequencies in both texts due to the range of possible meanings of the word, a more specific term like *fuck session*, which is more frequent in the gay texts than in the lesbian texts, might be revealed as key.

To give an illustrative example of the use of key clusters, the word *good* is key in the gay texts, occurring 1,479 times as opposed to 1,165 times in the lesbian texts. However, this does not tell us a great deal about how *good* is used within these texts. Does it occur in the same way, or does it have different types of uses? An examination of key two-word clusters reveals that in the gay texts *real good* (32 vs. 3 occurrences) and *good looking* (55 vs. 15 occurrences) are key. This finding gives us two of the most significant ways that the use of *good* differs between the two text types. In particular, the use of *real good* is interesting because it is a nonstandard use of *really good*, which could be an indication that the gay texts use more nonstandard language than the lesbian texts (an examination of other cases would be needed to test this). We could have probably derived the same information from carrying out a concordance or collocational analysis on the word *good*, although here we must rely on the word *good* to be key in the first place in order to bring it to our attention,

and there is no guarantee that this would have been the case. Using a variety of techniques when eliciting keywords, combined with a thorough analysis of how the keyword occurs in *all* the data (not just the texts where it appears key but also the comparison corpus) is therefore likely to result in more interesting and detailed research findings.

## Conclusion

A keyword list is a useful tool for directing researchers to significant lexical differences between texts. However, care should be taken to ensure that too much attention is not given to lexical differences while ignoring differences in word usage and/or similarities between texts. Carrying out comparisons among three or more sets of data, grouping infrequent keywords according to similar meaning or function, showing awareness of keyword dispersion across multiple files by using key keywords, carrying out analyses on key clusters and on grammatically or semantically annotated data, and conducting supplementary concordance and collocational analyses will enable researchers to obtain a more accurate picture of how keywords function in texts. Although a keyword analysis is a relatively objective means of uncovering lexical salience between texts, it should not be forgotten that the researcher must specify his or her cutoff points in order to determine levels of salience: such a procedure requires more analysis to establish how cutoff points can influence research outcomes.

When used sensitively, keywords can reveal a great deal about frequencies in texts, which is unlikely to be matched by researcher intuition. However, as with all statistical methods, how the researcher chooses to interpret the data is ultimately the most important aspect of corpus-based research.

## Notes

1. These data were "cleaned" by removing headers that gave extraneous information such as e-mail addresses, date of publication, and disclaimers. As a result, the final word counts were 985,331 words for the gay texts and 991,189 words for the lesbian texts.

2. Lesbian texts: standardized type token ratio: 40.08, average word length: 4.18, average sentence length: 17.72. Gay texts: standardized type token ratio: 40.01, average word length: 4.02, average sentence length: 17.73.

3. In corpus analysis, as Oakes (1998) points out, a significant amount of the data are skewed and do not therefore follow a normal distribution. However, log-normal distributions can help to overcome skewed data.

4. In WordSmith Version 4, the user will be able to specify a minimum consistency range (Mike Scott, personal communication, March 2004).

5. See, for example, http://ucrel.lancs.ac.uk/llwizard.html.

## References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Burr, Vivien. 1995. *An Introduction to Social Constructionism*. London: Routledge.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19 (1): 61-74.

Eckert, Penelope. 2002. Demystifying Sexuality and Desire. In *Language and Sexuality: Contesting Meaning in Theory and Practice*, edited by Kathryn Campbell-Kibler, Robert J. Podesva, Sarah Roberts, and Andrew Wong, 99-110. Stanford, CA: Center for the Study of Language and Information.

Fairclough, Norman. 2000. *New Labour, New Language?* London: Routledge.

Firth, J. R. 1957. *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Flowerdew, John. 1997. The Discourse of Colonial Withdrawal: A Case Study in the Creation of Mythic Discourse. *Discourse and Society* 8 (4): 453-77.

Garside, Roger, and Nicholas Smith. 1997. A Hybrid Grammatical Tagger: CLAWS4. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, edited by Roger Garside, Geoffrey Leech, and Anthony McEnery, 102-12. London: Longman.

Goffman, Erving. 1976. Gender Advertisements. *Studies in the Anthropology of Visual Communication* 3 (2): 69-154.

Johnson, Sally, Jonathan Culpeper, and Stephanie Suhr. 2003. From "Politically Correct Councillors" to "Blairite Nonsense": Discourses of Political Correctness in Three British Newspapers. *Discourse and Society* 14.1:28-47.

Krishnamurthy, Ramesh. 1996. Ethnic, Racial and Tribal: The Language of Racism? In *Texts and Practices: Readings in Critical Discourse Analysis*, edited by Carmen Rosa Caldas-Coulthard and Malcolm Coulthard, 129-49. London: Routledge.

McArthur, Tom. 1981. *Longman Lexicon of Contemporary English*. London: Longman.

Oakes, Michael P. 1998. *Statistics for Corpus Linguistics*. Edinburgh, Scotland: Edinburgh University Press.

Parker, Ian. 1992. *Discourse Dynamics: Critical Analysis for Social and Individual Psychology*. London: Routledge.

Parker, Ian, and Erica Burman. 1993. Against Discursive Imperialism, Empiricism and Constructionism: Thirty Two Problems with Discourse Analysis. In *Discourse Analytical Research*, edited by Erica Burman and Ian Parker, 155-72. London: Routledge.

Piper, Alison. 2000. Some People Have Credit Cards and Others Have Giro Cheques: "Individuals" and "People" as Lifelong Learners in Late Modernity. *Discourse and Society* 11 (4): 515-42.

Scott, Mike. 1999. *WordSmith Tools Help Manual*. Version 3.0. Oxford, UK: Mike Scott and Oxford University Press.

Stubbs, Michael. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. London: Blackwell.

Teubert, Wolfgang. 2000. A Province of a Federal Superstate, Ruled by an Unelected Bureaucracy: Keywords of the Eurosceptic Discourse in Britian. In *Attitudes towards Europe: Language in the Unification Process*, edited by Andreas Musolff, Colin Good, Petra Points, and Ruth Wittlinger, 45-86. Aldershot, UK: Ashgate.

Tribble, Christopher. 2000. Genres, Keywords, Teaching: Towards a Pedagogic Account of the Language of Project Proposals. In *Rethinking Language Pedagogy from a Corpus Perspective*, edited by Lou Burnard and Tony McEnery, 75-90. Frankfurt: Peter Lang.

Utka, Andrius. 2004. Analysis of George Orwell's Novel 1984 by Statistical Methods of Corpus Linguistics. *Sankirta: A Yearly Internet Journal of Lithuanian Corpus Linguistics*. Retrieved from http://donelaitis.vdu.lt/publikacijos/adrtmain.htm.

Williams, Raymond. 1983. *Keywords*. London: Fontana.

Wilson, Andrew, and Jenny Thomas. 1997. Semantic Annotation. In *Corpus Annotation: Linguistic Information from Computer Texts*, edited by Roger Garside, Geoffrey Leech, and Anthony McEnery, 55-65. London: Longman.

*Paul Baker is a lecturer in linguistics and modern English language at Lancaster University. His previous research includes papers in the* Journal of Sociolinguistics *and* Literary and Linguistic Computing*, as well as the monograph* Polari: The Lost Language of Gay Men *(Routledge).*