

# Chapter 1

## Does Frequency Really Matter?

Dawn Archer

### Words, words, words

A hypothesis popular amongst computer hackers – the infinite monkey theorem<sup>1</sup> – holds that, given enough time, a device that produces a random sequence of letters *ad infinitum* will, ultimately, create not only a coherent text, but also one of great quality (for example, Shakespeare’s *Hamlet*). The hypothesis has become more widely known thanks to David Ives’ satirical play, *Words, Words, Words* (*Dramatists Play Service*, NY). In the play, three monkeys – Kafka, Milton and Swift – are given the task of writing something akin to *Hamlet*, under the watchful eye of the experiment’s designer, Dr Rosenbaum. But, as Kafka reveals when she reads aloud what she has typed thus far, the experiment is beset with seemingly insurmountable difficulties:

“K k k k k, k k k! **K k k!** K ... k ... k.” I don’t know! I feel like I’m repeating myself!<sup>2</sup>

In my view, Kafka’s concern about whether the simple repetition of letters can produce a meaningful text is well placed. But I would contend that the frequency with which particular words are used in a text can tell us something meaningful about that text and also about its author(s) – especially when we compare word choice/usage against the word choice/usage of other texts (and their authors). This can be explained, albeit in a simplistic way, by inverting the underlying assumption of the infinite monkey theorem: we learn something about texts by focussing on the frequency with which authors use words precisely because their choice of words is seldom random.<sup>3</sup>

As support for my position, I offer to the reader this edited collection, which brings together a number of researchers involved in the promotion of ICT methods such as frequency and keyword analysis. Indeed, the chapters within *What’s in a Word-list? Investigating Word Frequency and Keyword Extraction* were originally

---

1 The infinite monkey theorem was first introduced by Émile Borel at the beginning of the twentieth century, and was later popularized by Sir Arthur Eddington.

2 D. Ives, *Words, Words, Words*, Dramatists Play Service, New York.

3 Of course, the extent to which this process is a *completely* cognitive one is a matter of debate.

presented at the Expert Seminar in Linguistics (Lancaster 2005). This event was hosted by the AHRC ICT Methods Network as a means of demonstrating to the Arts and Humanities disciplines the broad applicability of corpus linguistic techniques and, more specifically, frequency and keyword analysis.

### Explaining frequency and keyword analysis

Frequency and keyword analysis involves the construction of word lists, using automatic computational techniques, which can then be analyzed in a number of ways, depending on one's interest(s). For example, a researcher might focus on the most frequent lexical items of a number of generated word frequency lists to determine whether all the texts are written by the same author. Alternatively, they might wish to determine whether the most frequent words of a given text (captured by its word frequency list) are suggestive of potentially meaningful *patterns* that they might have missed had they read the text manually.<sup>4</sup> They might then go on to view the most frequent words in their word frequency list *in context* (using a concordancer) as a means of determining their collocates and colligates (i.e. the content and function words with which the most frequent words keep regular company). For example, the word 'ago' occurs 19,326 times in the *British National Corpus* (BNC)<sup>5</sup> and, according to Hoey,<sup>6</sup> 'is primed for collocation with *year, weeks and days*'. We can easily confirm this by entering the search string '\* ago' into Mark Davies's relational database of the BNC. In fact, we find that nouns relating to periods of time account for the 20 most frequent collocates of 'ago' (see Davies, this volume, for a detailed discussion of the relational database employed here, and Scott and Tribble,<sup>7</sup> for a more extensive discussion of the collocates of 'ago').

The researcher(s) who are interested in keyword analysis may also be interested in collocation and/or colligation, but they will compare, initially, the word frequency list of their chosen text (let's call it text A) with the word frequency list of another *normative* or *reference*<sup>8</sup> text (let's call it text B) as a means of identifying both words that are frequent and also words that are infrequent in text A, *statistically speaking*, when compared to text B. This has the advantage of removing words

---

4 M. Scott and C. Tribble, *Textual Patterns: Keyword and Corpus Analysis in Language Education* (Amsterdam: Benjamins, 2006), p. 5.

5 Produced in the 1990s, the BNC is a 100 million-word corpus of modern British English containing registers that are representative of the spoken and written medium.

6 M. Hoey, *Lexical Priming: A New Theory of Words and Language* (Routledge, 2005), p. 177.

7 Scott and Tribble, *Textual Patterns*, p. 43.

8 *Normative* corpus and *reference* corpus are often used interchangeably by corpus linguists.

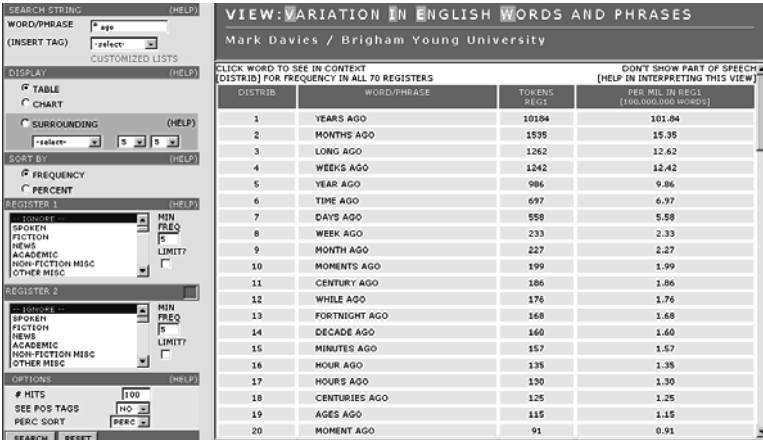


Figure 1.1 Results for \* ago in the BNC, using VIEW

that are common to both texts, and so allows the researcher to focus on those words that make text A *distinctive from* text B (and *vice versa*).

In the case of the majority of English texts, this will mean that function words ('the', 'and', 'if', etc.) do not occur in a generated keywords list, because function words tend to be frequent in the English language as a whole (and, as a result, are commonly found in English texts). That said, function words can occur in a keyword list if their usage is strikingly different from the norm established by the reference text. Indeed, when Culpeper<sup>9</sup> undertook a keywords analysis of six characters from Shakespeare's *Romeo and Juliet*, using the play *minus* the words of the character under analysis as his reference text, he found that Juliet's most frequent keyword was actually the function word 'if'. On inspecting the concordance lines for 'if' and additional keyword terms, in particular, 'yet', 'would' and 'be', Culpeper concluded that, when viewed as a set, they served to indicate Juliet's elevated pensiveness, anxiety and indecision, relative to the other characters in the play.

#### *Text mining techniques as indicators of potential relevance*

As the example of Juliet (above) reveals, a set of automatically generated keywords will not necessarily match a set of human-generated keywords at first glance. In some instances, automatically generated keywords may also be found to be

9 J. Culpeper, 'Computers, Language and Characterisation: An Analysis of Six Characters in *Romeo and Juliet*', in U. Melander-Marttala, C. Ostman and M. Kytö (eds), *Conversation in Life and in Literature: Papers from the ASLA Symposium* (Uppsala: Association Suédoise de Linguistique Appliquée, 2002), pp. 11–30.

*insignificant* by the researcher in the final instance (*see*, for example, Archer et al., this volume), in spite of being classified as statistically significant by text analysis software. This is not as problematic as it might seem. The reason? The main utility of keywords and similar text-mining procedures is that they identify (linguistic) items which are:

1. *likely* to be of interest in terms of the text's *aboutness*<sup>10</sup> and structuring (that is, its genre-related and content-related characteristics); and,
2. *likely* to repay further study – by, for example, using a concordancer to investigate collocation, colligation, etc. (adapted from Scott, this volume).

Put simply, the contributors to this edited collection are not seeking (or wanting) to suggest that the procedures they utilize can replace human researchers. On the contrary, they offer them as a *way in* to texts – or, to use corpus linguistic terminology, a way of *mining* texts – which is time-saving and, when used sensitively, informative.

### **Aims, organization and content of the edited collection**

The aims of *What's in a Word-list?* are similar to those of the 2005 Expert Seminar, mentioned above:

- to demonstrate the benefits to be gained by engaging in corpus linguistic techniques such as frequency and keyword analysis; and,
- to demonstrate the very broad applicability of these techniques both within and outside the academic world.

These aims are especially relevant today when one considers the rate at which electronic texts are becoming available, and the recent innovations in analytic techniques which allow such data to be mined in illuminating (and relatively trouble free) ways. The contributors also identify a number of issues that are crucial, in their view, if corpus linguistic techniques are to be applied successfully within and beyond the field of linguistics. They include determining:

- what counts as a *word*
- what we mean by *frequency*
- why frequency matters so much
- the *consistency* of the various keyword extraction techniques
- which of the (key)words captured by keyword/word frequency lists are the most relevant (and which are not)

---

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

- whether the (de)selection of keywords introduces some level of bias
- what counts as a reference corpus and why we need one
- whether a reference corpus can be *bad* and still show us something
- what we gain (in real terms) by applying frequency and keyword techniques to texts.

*Word frequency: use or misuse?*

John Kirk begins the edited collection by (re)assessing the concept of the *word* (as *token*, *type* and *lemmatized type*), the range of words (in terms of their functions and meanings) and thus our understanding of *word frequency* (as a property of data). He then goes on to refer to a range of corpora – the *Corpus of Dramatic Texts in Scots*, the *Northern Ireland Transcribed Corpus of Speech*, and the Irish component of the *International Corpus of English* – to argue that, although word frequency appears to promise precision and objectivity, it can sometimes produce imprecision and relativity. He thus proposes that, rather than regarding word frequency as an end in itself (and something that requires no explanation), we should promote it as:

- something that needs interpretation through contextualization
- a methodology, which lends itself to approximation and replicability.

Kirk also advocates that there are some advantages to be gained by paying attention to words of low frequency as well as words of high frequency. In his concluding comments, he touches on the contribution made to linguistic theory by word frequency studies, and, in particular, the usefulness of authorship studies in the detection of plagiarism.

*Word frequency, statistical stylistics and authorship attribution*

David Hoover continues the discussion of high versus low frequency words, and authorship attribution, focussing specifically on some of the innovations in analytic techniques and in the ways in which word frequencies are selected for analysis. He begins with an explanation of how, historically, those working within authorship attribution and statistical stylistics have tended to base their findings on fewer than the 100 most frequent words of a corpus. These words – almost exclusively function words – are attractive because they are so frequent that they account for most of the running words of a text, and because such words have been assumed to be especially resistant to intentional manipulation by an author.<sup>11</sup> Hoover then goes on to document the most recent work on style variation which, by concentrating on word frequency in given sections of texts rather than in the

---

<sup>11</sup> This means that their frequencies should reveal authorial habits which remain relatively constant across a variety of texts.

entire corpus, is proving more effective in capturing stylistic shifts. A second recent trend identified by Hoover is that of increasing the number of words analysed to as many as 6,000 most frequent words – a point at which almost all the words of the text are included, and almost all of these are content words.

The final sections of his chapter are devoted to the authorship attribution community's renewed interest in Delta, a method for identifying differences between texts that is based on comparing how individual texts within a corpus differ from the mean for that entire corpus (following the innovative work of John Burrows). Drawing on a two million word corpus of contemporary American poetry and a much larger corpus of 46 Victorian novels, Hoover also argues that refinements in the selection of words for analysis and in alternative formulas for calculating Delta may allow for further improvements in accuracy, and result, in turn, in the establishment of a theoretical explanation of how and why word frequency analysis is able to capture authorship and style.

### *Word frequency in context*

In Chapter 4, Mark Davies introduces some alternatives to techniques based on word searching. In particular, he focuses on the use he has made of architectures based on relational databases and n-gram<sup>12</sup> frequencies when developing corpora (including the 100-million-word *Corpus del Español*,<sup>13</sup> a BNC-based 100-million-word corpus modelled on the same architecture (*Variation in English Words and Phrases*, VIEW),<sup>14</sup> and a 40-million-word *Corpus of Historical English*.<sup>15</sup> Davies' main proposal is that such architectures can dramatically improve performance in the searching of corpora. For example, the following capture three of the many simple word frequency queries that take no more than one to two seconds on a 100-million-word corpus:

- overall frequency of a given word, set of words, phrase, or substring in the corpus;
- 'slot-based' queries, e.g. the most common nouns one 'slot' after 'mysterious', or z-score rank words immediately preceding 'chair'; and,
- wide-range collocates, e.g. the most common nouns within a ten-word window (left or right) of 'string' or 'broken'.

Davies also highlights the importance of developing an architecture that can account for variation through the creation of n-gram frequency tables for each register within a given corpus. The advantage of such an approach is that

---

12 An n-gram is a (usually consecutive) sequence of items from a corpus. The items in question can be characters (letters and numbers) or more usually words.

13 <<http://www.corpusdelespanol.org/>>.

14 <<http://corpus.byu.edu/bnc/>>.

15 <<http://view.byu.edu/che/>>.

each n-gram will have an associated frequency according to historical period and register, and this information will be directly accessible as part of a given query.

Like Kirk (chapter 2, this volume), Davies is zealous about word frequency being something that needs interpretation through contextualization. Indeed, he advocates that word frequency, 'be analyzed not just as the overall frequency of a given word or lemma in a certain corpus, but, rather, as the frequency of words in a wide range of related contexts' (p. 66). Unlike Kirk, however, he does not seem to be readily concerned about the inclusion of low frequency words in any given query. This is because of a potential 'size issue' which means that n-gram tables can 'become quite unmanageable' when dealing with excessively large corpora (p. 57). Consequently, Davies advocates that, for such corpora, we include just those n-grams that occur three times or more. This is not a problem if one is interested in only the highly-frequent n-grams, of course, but it could make a detailed comparison of sub-corpora potentially problematic.

### *Issues for historical and regional corpora – first catch your word*

In Chapter 5, Christian Kay focuses primarily on variable spelling within historical texts, and the difficulties that this occasions when seeking to 'catch a word' in corpora, especially corpora such as the *Historical Thesaurus of English* (HTE), and a semantic index to the *Oxford English Dictionary*,<sup>16</sup> which is supplemented by Old English materials (published separately in Roberts et al.'s *A Thesaurus of Old English*<sup>17</sup>) and, as such, captures English vocabulary from the earliest written records to the present.<sup>18</sup>

Kay goes on to point out that spelling variation can also create problems when searching corpora relating to (modern-day) non-standard varieties such as the *Scottish Corpus of Texts and Speech* and the *Dictionary of the Scots Language*. Indeed, even the specialized dictionaries that lemmatize common variants (for example, the *Dictionary of the Scots Language*) are by no means comprehensive. She also demonstrates how homonymy and polysemy can create additional problems for those working with (historical and dialectal) corpora – and this is something that lemmatization may not be able to solve. Kay concludes by suggesting ways of addressing some of these problems using the resources described above, including the development of a rule-based system which predicts possible variants and maps them to the relevant headwords (See also chapter 9). In addition, Kay touches on

---

16 *Oxford English Dictionary* (Oxford: Oxford University Press, 1884–, and subsequent edns); *OED Online*, ed. J.A. Simpson. (Oxford: Oxford University Press, 2000–).

17 J. Roberts, C. Kay and L. Grundy, *A Thesaurus of Old English* (Amsterdam: Rodopi, 2000 [1995]).

18 Word senses within the thesauri are organized in a hierarchy of categories and subcategories, with up to 14 levels of delicacy. The material is held in a database that can be searched on the Internet, and is likely to be of use in a range of humanities disciplines.

the relationship between *e*-texts (of which there are many) and structured corpora (of which there are few).

*In search of a bad reference corpus*

Mike Scott's contribution to this edited collection tackles the issue of reference corpora. More specifically, he is interested in determining how *bad* a reference corpus can be before it becomes unusable (in the sense that it generates keywords that do not help to clarify the *aboutness* of a target text). As previous chapters have revealed, this issue is particularly pertinent, as *good* reference corpora are not available for all genres / periods / languages.

Using the keywords facility of his own text analysis program, WordSmith Tools,<sup>19</sup> Scott's starting point is the formula proposed by Berber Sardinha, which suggests that the larger the reference corpus, the more keywords will be detected.<sup>20</sup> Berber Sardinha also suggests that, as a reference corpus that is similar to the target text (i.e. the text being analysed) will filter out genre features common to both, an optimum reference corpus is one that contains several different genres. Drawing on a series of reference texts of varying lengths (32 in total: 22 BNC texts and 10 Shakespeare plays), Scott explores the different keyword results that are generated by WordSmith Tools for two target texts: an extract from a book profiling business leaders and a doctor/patient interaction. Scott pays particular attention to their 'popularity' and 'precision' scores as a means of answering three research questions:

1. To what extent does the size of the reference text impact on the quality of the keywords and, if so, is there a point at which the size of the reference text renders the (quality of the) keywords unacceptable?
2. What sort of keyword results obtain if a reference text is used which has little or no relation to the target text (beyond them both being written in the same language)?
3. What sort of keyword results obtain if *genre* is included as a variable?

Popularity relates to the presence of each keyword in the majority of the reference texts (for example, 20 out of the 22 BNC texts). This is based on the rationale that keywords which are identified using most of the reference texts are more likely to be useful than those identified in only a minority of the reference texts. Precision is

---

19 M. Scott, *WordSmith Tools, Version 4* (Oxford: Oxford University Press, 2004). WordSmith Tools is probably the most popular text analysis program in corpus linguistics. For more information, see <<http://www.lexically.net/wordsmith/>>.

20 The critical size of a reference corpus is said to be about two, three and five times the size of the node text: A.P. Berber Sardinha, *Linguística de Corpus* (Barueri, São Paulo, Brazil: Editora Manole, 2004), pp. 101–103).

computed following Oakes,<sup>21</sup> and involves dividing the total number of keywords for each reference text by the number of popular keywords (as determined by the popularity test).

Whilst Scott admits that *usefulness* is a relative phenomenon, which is likely to vary according to research goals (and research goals cannot be predicted with certainty), he contends that it is still worth undertaking such a study, not least because it will help to determine the dimensions that appear to effect the *meaningfulness* (or not) of generated keywords. These include size in tokens (i.e. frequency), similarity of text-type, similarity of historical period, similarity of subject-matter, etc. More importantly, perhaps, this and later studies will provide a useful means of determining the robustness of the keywords procedure and thus, in turn, its potential usefulness in (non-)linguistic fields. And the indications from this preliminary study look promising; indeed, Scott suggests that even relatively restricted reference corpora can give good results in keyword extraction. That said, Scott notes that a small reference corpus containing a mixture of texts is likely to perform better than a larger corpus with more homogeneous texts.

#### *Keywords and moral panics – Mary Whitehouse and media censorship*

Tony McEnery also utilizes the keywords facility of WordSmith Tools – in conjunction with his own lexically driven model of moral panic theory<sup>22</sup> – as a means of determining the extent of *moral panic* in the books penned by Mary Whitehouse during the period 1967–77. In brief, words that are found to be *key* (i.e. statistically frequent) in the writings of Whitehouse (relative to a reference corpus<sup>23</sup>) are classified according to McEnery’s moral panic categories.<sup>24</sup> These categories are heavily influenced by the moral panic theory of the sociologist, Stanley Cohen. Indeed, they capture the discourse roles thought to typify moral panic discourse, including ‘object of offence’, ‘scapegoat’, ‘moral entrepreneur’, ‘corrective action’, ‘consequence’, ‘desired outcome’ and ‘rhetoric’. Some of the categories are also sub-classified according to pertinent semantic fields: for example, the *scapegoat* category contains the semantic fields of ‘people’, ‘research’, ‘broadcast programmes’, ‘media’, ‘media organisations and officers’ and ‘groups’. These semantic fields have been generated using a ‘bottom-up’ approach: that is to say, they have been constructed by McEnery, rather than being

---

21 M. Oakes, *Statistics for Corpus Linguistics* (Edinburgh: Edinburgh University Press, 1998), p. 176.

22 A.M. McEnery, *Swearing in English: Bad Language, Purity and Power from 1586 to the Present* (Routledge, 2005).

23 McEnery opted to use the Lancaster-Oslo-Bergen (LOB) corpus as his reference corpus. The LOB captures 15 text categories, containing 500 printed texts of British English (approximately 2,000 words each) all of which were produced in 1961.

24 Ibid. see also A.M. McEnery, ‘The Moral Panic about Bad Language in England, 1691–1745’, *Journal of Historical Pragmatics*, 7/1 (2006): 89–113.

automatically identified by a text analysis tool (*see* Baker, Chapter 8, and Archer et al., Chapter 9, for a useful comparison with the 'bottom-up' approach). Using this procedure, McEnergy is able not only to capture keywords that help establish the *aboutness* of the moral panic, but also to determine those words (like 'violence') that are actually *key* keywords, i.e. are *key* in a number of related texts (as well as moral panic categories) in the corpus.<sup>25</sup>

McEnergy's chapter is a good example of the benefits to be gained from combining a keywords methodology with other theories (linguistic and non-linguistic) – not least because it demonstrates the usefulness of corpus linguistic techniques beyond linguistics. In addition, McEnergy is one of several authors in this edited collection (*see*, for example, Baker, Chapter 8, and Archer et al., Chapter 9) who seek to combine a quantitative approach to text analysis with a qualitative approach. Indeed, McEnergy specifically focuses on the issue of bad language and, in particular, how bad language was represented by Whitehouse's organization VALA (Viewers and Listeners' Association), through an investigation of the collocations and colligations of several of the more prominent key keywords in Whitehouse's books. Moreover, he discusses those findings not only in respect of their semantic importance, but also in respect of their ideological significance. He argues, for example, that the key keywords within the corrective action category, 'parents' and 'responsible', serve to generate in and out-groups, the former being regarded as serious, reasonable and selfless, and the latter, as the antithesis of these qualities. In addition, a closer inspection of the key keywords in context (using a concordancer) suggests that the in/out-group distinction is heavily related to a dichotomy between (religious) conservatism and liberalism, which, in turn, is comparable to the opposition to bad language voiced by seventeenth-century religious organizations.

### *Keywords, fox hunting and the House of Commons*

Paul Baker is the third author in this edited collection to utilize the keywords facility in WordSmith Tools – in this case, to examine a small corpus of debates on fox hunting (totalling 130,000 words). The debates took place in the (British) House of Commons in 2002 and 2003, prior to a ban being implemented in 2005. For the purposes of this study, Baker split the corpus into two sub-corpora (depending on whether speakers argue for or against fox hunting to be banned) so that they could be compared with each other, rather than with a more general reference text.

The bulk of Baker's chapter is dedicated to a discussion of the different discourses (or ways of looking at the world) that speakers access in order to persuade others of their point of view, which Baker identifies using concordance

---

25 McEnergy suggests that the key keywords approach is especially useful when one is working with large volumes of data (and the volume is such that the number of keywords generated is overwhelming). Key keywords are also useful if the transience of particular keywords may be an issue.

analyses of pertinent keywords. For example, he notes how the pro-hunt speakers overused ‘people’, relative to the anti-hunt speakers. Moreover, they tended to use the term to identify those:

- who would be adversely affected by the ban if it was implemented (because of losing their jobs and/or their communities and/or facing the possibility of a prison sentence, if they opted to ignore the ban), and
- who do not hunt, but were not upset or concerned by those who do.

In addition, the pro-hunt speakers also utilized the keywords ‘fellow’, ‘citizens’, ‘Britain’ and ‘freedom’, the first two occurred together as a noun phrase, e.g. ‘fellow citizens’, and when used as such were preceded in all cases by a first person possessive pronoun (‘my’ or ‘our’). Baker argues that the pro-hunt speakers could thus be seen to use an hegemonic rhetorical strategy to intimate that it was they (and not their opponents) who were able to speak for and with the people of Britain.

Baker also explores additional ways of using keyness to find salient language differences in texts, including the identification of key semantic categories (also referred to as key domains). A tool that enables such analysis to be undertaken automatically is the UCREL Semantic Analysis System (henceforth USAS, also referred to as the UCREL Semantic Annotation System). Developed at Lancaster University, USAS consists of a part-of-speech tagger, which utilizes CLAWS (the Constituent Likelihood Automatic Word-tagging System), and a semantic tagger that, at its conception, was loosely based on McArthur’s *Longman Lexicon of Contemporary English*,<sup>26</sup> but has since been revised in the light of practical application.<sup>27</sup> Currently, the semantic tagset consists of 21 macro categories that expand into 232 semantic fields (see Appendix 2). Once again, Baker focuses on just a few of the most salient key semantic categories. For example, he points out how the semantic category ‘S1.2.6 sensible’ is overused by the pro-hunt speakers in the parliamentary debates (relative to the anti-hunt speakers): words like ‘sensible’, ‘reasonable’, ‘common sense’ and ‘rational’ are used when discussing the reasons for keeping hunting, and ‘ridiculous’, ‘illogical’ and ‘absurd’, when describing the proposed ban on hunting, which prompts Baker to suggest that this may be another example of their hegemonic rhetorical strategy (i.e. presenting one’s view of the world as ‘right’ or ‘common sense’).

---

26 T. McArthur, *Longman Lexicon of Contemporary English* (Longman, 1981).

27 See, for example, A. Wilson and J. Thomas, ‘Semantic Annotation’, in R. Garside, G. Leech and A. McEnery (eds), *Corpus Annotation: Linguistic Information from Computer Texts* (Longman, 1997), pp. 55–65; P. Rayson, D. Archer, S.L. Piao. and T. McEnery, ‘The UCREL Semantic Analysis System’, proceedings of the workshop on Beyond Named Entity Recognition Semantic Labelling for NLP Tasks in association with the fourth international conference on Language Resources and Evaluation (LREC, 2004), pp. 7–12.

Baker concludes by suggesting that keywords offer a potentially useful way of focussing researcher attention on aspects of a text or corpus, but that care should be taken not to over-focus on difference/presence at the expense of similarity/absence. He also suggests that the best means of gaining the fullest possible picture of the *aboutness* of text(s) is to use multiple reference corpora. For example, one might wish to compare texts of the same type against a (larger) corpus of (more) general language usage as a means of capturing those words that, because they are typical of the text-type or genre, may be too similar to show up as keywords in a same text-type comparison (*see my discussion of Romeo and Juliet under 'Explaining Frequency and Keyword Analysis'*).

*An exploration of key domains in Shakespeare's comedies and tragedies*

Archer, Culpeper and Rayson also utilize USAS, in this case to explore the concept of *love* in three Shakespearean love-tragedies (*Othello, Anthony and Cleopatra* and *Romeo and Juliet*) and three Shakespearean love-comedies (*A Midsummer Night's Dream, The Two Gentlemen of Verona* and *As You Like It*). Their aim is to add a further dimension to approaches that:

- use corpus linguistic methodologies such as keyword analysis to study Shakespeare,<sup>28</sup> by systematically taking account of the semantic relationships between keywords through an investigation of key domains; and
- study Shakespeare from the perspective of cognitive metaphor theory,<sup>29</sup> by providing empirical support for some of the love-related conceptual metaphors put forward by cognitive metaphor theorists.

In brief, their *top-down*<sup>30</sup> approach involves determining how love is presented in the two datasets and then highlighting any resemblances between their findings and the conceptual metaphors identified by cognitive metaphor theorists. They also discuss how the semantic field of love co-occurs with different domains in the two datasets, and assess the implications this has on our understanding of the concept of love.

As the original USAS system is designed to undertake the automatic semantic analysis of present-day English language, they have opted to utilize the historical version of the tagger. Developed by Archer and Rayson, the historical tagger includes supplementary historical dictionaries to reflect changes in meaning over

---

<sup>28</sup> See, for example, Culpeper, 'Computers, Language and Characterisation'.

<sup>29</sup> See, for example, D.C. Freeman, "'Catch[ing] the nearest way": *Macbeth* and Cognitive Metaphor', *Journal of Pragmatics*, 24 (1995): 689–708.

<sup>30</sup> *Top-down* captures the fact that the categories are pre-defined and applied automatically by USAS.

time and a pre-processing step to detect variant (i.e. non-modern) spellings.<sup>31</sup> The inclusion of a variant detector is important when automatically annotating historical texts as it means that variant spellings can be mapped to spellings that the text analysis tool can recognize; this, in turn, means that standard corpus linguistic methods (frequency profiling, concordancing, keyword analysis, etc.) are more effective (see Kay, Chapter 5). The taxonomy of the historical tagger is the same, at present. However, Archer et al. are using studies such as this to evaluate its suitability for the Early Modern English period.<sup>32</sup> Indeed, they comment on the semantic domains that seem to capture the data well in their chapter, whilst also pointing out semantic domains that do not work as well. For example, they explain how the overuse of L3 ‘Plants’ in the love-comedies (relative to the love-tragedies) can be explained in large part by ‘Mustardseed’ (a character’s name) and ‘flower’ (part of the phrase, ‘Cupid’s flower’, i.e. the flower that Oberon used to send Titania to sleep in *A Midsummer Night’s Dream*). In addition, the bulk of the remaining items in the L3 category capture features of the setting (for *As You Like It* and *A Midsummer Night’s Dream* are set in the woods).

Nevertheless, even within the L3 category, there are items which have a strong metaphorical association with ‘love’ or ‘sex’. By way of illustration, in *As You Like It*, Silvius uses an agricultural metaphor (‘crop’, ‘glean’, ‘harvest’, ‘reaps’) to confirm that he is prepared to have Phoebe as a wife in spite of her less-than-virginal state. According to Oncins-Martínez,<sup>33</sup> the ‘sex is agriculture’ metaphor and its sub-mappings (‘a woman’s body is agricultural land’, ‘copulation is ploughing or sowing’, etc.) were common in the Early Modern English period.

As Archer et al.’s findings demonstrate, then, a keyness analysis does not merely capture *aboutness*; it can also uncover metaphorical usage, as in this case, or character traits, as in the case of Culpeper,<sup>34</sup> discussed above. In addition, their approach can confirm – and also suggest amendments to – existing conceptual metaphors. By way of illustration, they suggest that the container idea within

---

31 D. Archer, T. McEnery, P. Rayson and A. Hardie, ‘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*, UCREL Technical Paper Number 16 (Lancaster: UCREL, 2003), pp. 22–31; Rayson et al. 2005); Rayson, P., Archer, D. and Smith, N., ‘VARD Versus Word: A Comparison of the UCREL Variant Detector and Modern Spell Checkers on English Historical Corpora’, *Proceedings of the Corpus Linguistics Conference Series On-Line E-Journal* 1:1 (2005).

32 Archer and Rayson are also exploring the feasibility of mapping the USAS tagset to, first, the categories utilized by Spevack, in his *A Thesaurus of Shakespeare* and, then, to the Historical Thesaurus of English.

33 J.L. Oncins-Martínez, Notes on the Metaphorical Basis of Sexual Language in Early Modern English, in J.G. Vázquez-González et al. (eds), *The Historical Linguistics-Cognitive Linguistics Interface* (Huelva: University of Huelva Press, 2006).

34 Culpeper, ‘Computers, Language and Characterisation’.

Barcelona Sánchez's<sup>35</sup> 'eyes are containers for superficial love' (which, in itself, is a development of Lakoff and Johnson's<sup>36</sup> 'eyes are containers for the emotions') is not clearly articulated in the (comedy) data, and that the latter would be better captured by the conceptual metaphor, 'eyes are weapons of entrapment'. This particular finding is made possible because of their innovative analysis of key collocates at the domain level, using Scott Piao's *Multilingual Corpus Toolkit*.<sup>37</sup>

Like Baker, Archer et al. believe that key domains can capture words that, because of their low (comparative) frequency, would not be identified as keywords in and of themselves.<sup>38</sup> However, they are acutely aware that the USAS process is an automatic one, and so will mis-tag words on occasion. Archer et al. therefore suggest that researchers thoroughly check the results of such processes, using a manual examination of concordance lines to determine their contextual relevance. By way of illustration, they comment on the occurrence of 'deer', which is assigned to the category L2 'Living creatures' by USAS. Archer et al. found that deer (like many items assigned to L2) was used metaphorically, and can be captured by the conceptual metaphor 'love is a living being' and the related metaphor 'the object of love is an animal'. When the concordance lines of these items were checked, they discovered that, although correctly assigned, the bulk of them had strong negative associations, semantically speaking. This finding contrasts with the items that Barcelona Sánchez<sup>39</sup> discusses in respect of *Romeo and Juliet*. Indeed, even the 'deer' example is problematic: it is linked to cuckoldry in many of Shakespeare's plays (e.g. *Love's Labours Lost*, *The Merry Wives of Windsor*) and may indicate that it, too, had negative undertones for both Shakespeare and his audience.<sup>40</sup>

---

35 A. Barcelona Sánchez, 'Metaphorical Models of Romantic Love in *Romeo and Juliet*', *Journal of Pragmatics*, 24 (1995): 667–88, 679.

36 G. Lakoff, and M. Johnson, *Metaphors We Live By* (Chicago and New York: University of Chicago Press, 1980).

37 S.L. Piao, A. Wilson and T. McEnery, 'A Multilingual Corpus Toolkit', paper given at AAACL–2002, Indianapolis, Indiana, USA, 2002.

38 Given many authors/(public) speakers seek to avoid unnecessary repetition by using alternatives to a given word, I would suggest that key domain analysis provides us with a useful means of capturing low frequency words that (although not key in and of themselves) do become 'key' when viewed alongside terms with similar meaning (see Rayson 2003, 100–113, for a more detailed exploration of the advantages of the key domains approach).

39 Barcelona Sánchez, 'Metaphorical Models of Romantic Love in *Romeo and Juliet*', p. 683.

40 Culpeper's investigation is another useful reminder of the importance of checking – as a means of contextualizing – any generated keywords (or key domains). For Culpeper found that some of the nurse's keywords in *Romeo and Juliet* ('god', 'warrant', 'faith', 'marry', 'ah') did not relate to her character at all – or to *aboutness* for that matter. Rather, they were surge features (or outbursts of emotion), which occurred at points in the play when the nurse was reacting to traumatic events (involving Juliet, in particular). Culpeper, 'Computers, Language and Characterisation'.

Their final sentence is devoted to a call for quantitative analysis to be combined with qualitative analysis. For, like Baker (Chapter 8), they recognize that it is the researcher who must determine their cut-off points in respect of (contextual) salience in the final instance. Indeed, how the researcher chooses to interpret the data is probably the most important aspect of corpus-based research.

### *Promoting the wider use of word frequency and keyword extraction*

In this final Chapter, I report on several AHRC ICT Methods Network promotional events (some of which were inspired by the Expert Seminar in Linguistics) that have helped to bring frequency and keyword extraction techniques to a wider community of users. I also address ways in which we might promote word frequency and keyword extraction techniques to an even wider community than we have at present (commercial and academic). In particular, I stress the need for (ongoing) dialogue, so that:

- the keyword extraction community can discover what it is that other research communities are interested in finding out, and then determine how their tools might help them to do so; and
- ‘other’ research communities keep the keyword extraction community informed of (the successes and failures of) research that makes use of text mining techniques, which will allow the latter, in turn, to improve (the functionality of) their text analysis tools further.

### *How to use this book*

I have deliberately incorporated detailed summaries of the contributors’ chapters in this introductory chapter so that readers can ‘pick and choose’ those chapters that seem most relevant to their interests. That said, I would encourage readers with the time and inclination to read the edited collection as a whole, so that they gain a better sense of the different issues that must be considered if we are to utilize word frequency and keyword extraction techniques successfully. The most important message of this edited collection, however, is that the researcher who engages in word frequency/keyword analysis has at their disposal a relatively objective means of uncovering lexical salience/(frequency) patterns that invite – and frequently repay – further qualitative investigation.